# Task Oriented Privacy-preserving (TOP) Technologies Using Automatic Feature Selection

Yasser Jafer

This thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
in partial fulfillment of the requirements
for the Doctorate in Philosophy degree in
Electrical and Computer Engineering

Ottawa-Carleton Institute for Electrical and Computer Engineering
Faculty of Electrical Engineering and Computer Science
University of Ottawa

# Abstract

A large amount of digital information collected and stored in datasets creates vast opportunities for knowledge discovery and data mining. These datasets, however, may contain sensitive information about individuals and, therefore, it is imperative to ensure that their privacy is protected.

Most research in the area of privacy preserving data publishing does not make any assumptions about an intended analysis task applied on the dataset. In many domains such as healthcare, finance, etc; however, it is possible to identify the analysis task beforehand. Incorporating such knowledge of the ultimate analysis task may improve the quality of the anonymized data while protecting the privacy of individuals. Furthermore, the existing research which consider the ultimate analysis task (e.g., classification) is not suitable for high-dimensional data.

We show that automatic feature selection (which is a well-known dimensionality reduction technique) can be utilized in order to consider both aspects of privacy and utility simultaneously. In doing so, we show that feature selection can enhance existing privacy preserving techniques addressing *k*-anonymity and differential privacy and protect privacy while reducing the amount of modifications applied to the dataset; hence, in most of the cases achieving higher utility.

We consider incorporating the concept of privacy-by-design within the feature selection process. We propose techniques that turn filter-based and wrapper-based feature selection into privacy-aware processes. To this end, we build a layer of privacy on top of regular feature selection process and obtain a privacy preserving feature selection that is not only guided by accuracy but also the amount of protected private information.

In addition to considering privacy after feature selection we introduce a framework for a privacy-aware feature selection evaluation measure. That is, we incorporate privacy

during feature selection and obtain a list of candidate privacy-aware attribute subsets that consider (and satisfy) both efficacy and privacy requirements simultaneously.

Finally, we propose a multi-dimensional, privacy-aware evaluation function which incorporates efficacy, privacy, and dimensionality weights and enables the data holder to obtain a best attribute subset according to its preferences.

# Acknowledgements

I would like to express my sincere gratitude to my advisors Dr. Stan Matwin and Dr. Marina Sokolova. It has been an honour to be their student, to work with them, and learn from them. Dr. Matwin, thank you for your support, insightful remarks, valuable advice and suggestions which always enhanced my research and guided me throughout my work. Dr. Sokolova, thank you for your support, critical comments and suggestions which added considerable value to my work. Your mentorship was always needed and appreciated throughout those years.

I would like to express my deep gratitude to my parents for their continued love and support and in their encouragement in pursuing my graduate studies.

Endless thanks to my wife Zahraa for being patient, caring, encouraging, and supportive throughout those years.

*To my dear Zahraa, Mohammad, Nour*

# List of Tables

# List of Figures

# Acronyms

| | |
|---|---|
| PPDP | Privacy Preserving Data Publishing |
| PPDM | Privacy Preserving Data Mining |
| KDD | Knowledge Discovery in Databases |
| PbD | Privacy-by-Design |
| TOP | Task Oriented Privacy-preserving |
| TOP_KDD | Task Oriented Privacy-preserving Knowledge Discovery in Databases |
| QI | Quasi-Identifier |
| DH | Data Holder |
| DR | Data Recipient |
| PPFS | Privacy Preserving Feature Selection |
| PW | Privacy-aware Wrapper |
| PF-IFR | Privacy-aware Filters- Individual Feature Ranking |
| PETs | Privacy Enhancing Technologies |
| SDC | Statistical Disclosure Control |
| PBI | Privacy Breach Increase |
| EC | Equivalence Class |
| IG | Information Gain |

# Contents

# Chapter 1

# Introduction

A large amount of digital information collected and stored in datasets creates tremendous opportunities for knowledge discovery and data mining. These datasets, however, may contain sensitive information about individuals and therefore, necessary measures should be put in place to ensure that the privacy of individuals is protected.

In general, data privacy and Privacy Enhancing Technologies (PETs) are spread across several fields. The main goal is to study the protection of information by preventing disclosure of sensitive information of individuals. Statistical Disclosure Control (SDC), which is rooted in the statistics community, is one of the first disciplines that dealt with this problem. SDC aims at developing techniques to enable publishing data from statistical agencies while protecting the privacy of their users (Navarro-Arribas et al. 2014)

In addition to the statistics community, data privacy has become important research area in computer science as well. We find privacy enhancing technologies ranging from cryptography to privacy preserving data mining and private information protocols. It is difficult to classify the privacy enhancing technologies because of the overlap between most of disciplines.

A well accepted categorization of dataset privacy was proposed in (Domingo-Ferrer 2007). In their research, techniques are being classified based on whose privacy they are trying to protect. Three dimensions have been identified; namely, *respondent* privacy, *owner* privacy, and *user* privacy. In the case of *respondent* privacy, respondent is the subject of the data. For example, in the area of SDC, given a census dataset, the respondents are the subjects included in the census. The respondent is a passive subject that cannot act within the system in order to protect its own data. *Respondent* privacy's goal is to avoid re-identification of, say, patients or other individuals whom dataset records refer to. In the case of *owner* privacy, owner is the administrator holder of the data. The owner is normally liable for disclosure of sensitive data. Therefore, the aim of owner privacy is to ensure that the owner does not give away his dataset. Finally, user in the *user* privacy dimension is considered the active counterpart of respondent. In this case, the user is responsible for protecting his/her own private data and is able to interact with the system for this purpose.

*Respondent* privacy is perhaps the most common case found in data privacy scenarios. Most common SDC and Privacy Preserving Data Mining (PPDM) techniques usually fit in this category. Their common goal is to protect a dataset (i.e., to protect the privacy of the respondents) and yet being able to apply statistical analysis and data mining techniques (Navarro-Arribas et al.). In our work, we mainly consider *respondent* privacy.

Two closely related research areas in protecting the privacy of individuals in the datasets include Privacy Preserving Data Publishing (PPDP) and Privacy Preserving Data Mining (PPDM). PPDP is mainly concerned with developing tools and methods that enable publishing data in an insecure environment. PPDP aims at publishing modified data such that while individuals' privacy is preserved, the published data remains practically useful (Mohammed et al. 2011). PPDP mainly focuses on application-free protection of data when the data is required for research purpose or business transactions (Lin et al. 2011b). PPDM, on the other hand, aims at extending traditional data mining techniques in order to work effectively with the modified data. That is, in PPDM, the data mining algorithms are analyzed for their potential side-effects on data privacy. The main objective of PPDM is to develop algorithms that modify the original data such that the private data is protected even after the mining process (Verykios et al. 2004b). PPDM focuses on privacy issues when data miners want real data in order to perform data mining (Lin et al. 2011b). Two

main reasons for PPDM's increasing importance in recent years include improved ability to store data that contains personal information about users, and increasing sophistication of data mining techniques to leverage such information (Aggarwal and Yu 2008).

Most research in the areas of PPDP do not make any assumptions about an intended analysis task applied on the dataset. In many domains such as healthcare, finance, etc, however, it is possible to identify the analysis task beforehand. In general, considering a particular publishing scenarios is a fruitful direction and leads to identifying the best algorithm based on the scenario at hand (Ayala-Rivera et al. 2014). Incorporating the knowledge of a given ultimate analysis task in PPDP may improve the quality of the anonymized data while protecting the privacy of individuals. Consider the following example borrowed from (Sun et al. 2009): Two distinct data recipients use the same data for different purposes. In the first scenario, a research center from a university requests the census data from the Census Bureau in order to conduct a demographic analysis in a local area. In the second scenario, a PhD student from the faculty of business requires the same census data in order to investigate or predict the potential business opportunities in a local area. In each case, the analysis task is different. As a result, the published data may be good for one purpose and not for another or may not be good for either purpose. Such 'one-size-fits-all' methodology usually does not result in the most optimal solution from a utility point of view. When data undergoes general-purpose anonymization (without considering the intended analysis task) we may end up with over-protection of data, i.e., over-anonymization which eventually results in adverse impact on its utility.

In our work, one main case is to take advantage of feature selection as a dimensionality reduction tool. It is, therefore, important to make assumptions about the ultimate usage of the data and in order to tailor a given data release to the intended purpose. This purpose-based or application-oriented privacy has been addressed previously and our work fits into this dimension.

To this end, the impact of the ultimate analysis "task" in PPDP and PPDM may be studied in a bigger dimension, i.e., in the process of Knowledge Discovery in Databases and the notion of privacy-by-design. We first briefly discuss these two concepts.

## 1.1. Knowledge Discovery in Databases (KDD)

Knowledge Discovery in Databases (KDD) refers to the process of transforming raw data into useful information. A more formal definition of the KDD process is given as "the non trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data" (Fayyad et al. 1996b).

The KDD process is summarized in Figure 1.1. Such process is both interactive and iterative with many decisions that are made by the user.

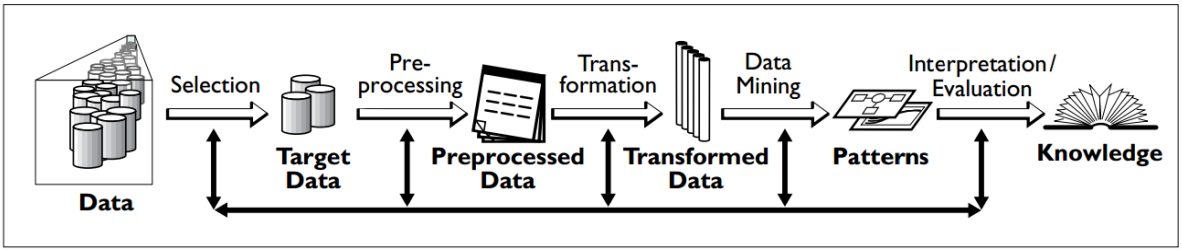



Figure 1.1: The KDD process steps (Fayyad et al. 1996a).

The steps of the KDD process are outlined in (Fayyad et al. 1996b) as follows:

- Step 1: Understanding the application domain and identifying the goal of the KDD process from the customer's point of view.
- Step 2: Creating a target dataset (and to select a subset of data samples or variables) that will be used for discovery.
- Step 3: Performing data cleaning and data preprocessing such as noise removal, handling missing data, etc.
- Step 4: Data reduction which consists of identifying relevant features according to a given analysis goal. This step involves dimensionality reduction or transformation methods.
- Step 5: Matching given data mining methods (e.g., classification, regression, clustering, etc) to the goals of the KDD process.
- Step 6: Choosing the data mining method(s).
- Step 7: Performing the actual data mining and searching for patterns of interest such as classification rule, classification trees, clustering, regression, and so on.
- Step 8: Interpreting the mined patterns and possibly returning to any of the steps 1 through 7.

- Step 9: Acting on the discovered knowledge.

In this process, data mining is considered one step, although it is, by far, the most important step and has received most attention in the literature (Fayyad et al. 1996b).

## 1.2. The Importance of Protecting Privacy

"Privacy is not simply a precious and often irreplaceable human resource; respect to privacy is the acknowledgment of respect for human dignity and the individuality of man" (https://www.priv.gc.ca).

Privacy has been associated with Maslow's hierarchy of human needs (Clarke 2006). Abraham Harold Maslow was an American psychologist. His proposed idea (Maslow 1943) was that people gets motivated to achieve certain needs. After fulfilling a given need person seeks to fulfill the next one. These needs start with "Physiological" at the bottom of the hierarchy, followed by "Safety", "Love/belonging", "Esteem", and finally "Self-actualization" at the top of the hierarchy.

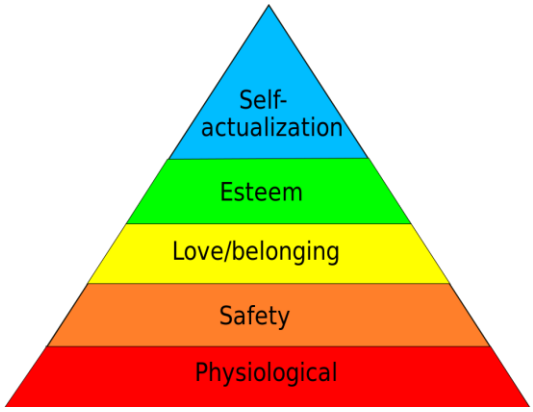



Figure 1.2: Illustration of Maslow's hierarchy of human needs (Wikipedia)

In (Clarke 2006) an interesting implicit association between these needs and different types of privacy was made. "Physiological" and "Safety" needs are associated with bodily privacy and privacy of the person. "Love/belonging" and "status and self-esteem" are associated with privacy of personal communication and personal behavior, and "Self-actualization" is associated with privacy of personal data.

Jeopardizing individual privacy due to improper share of information amoung different parties has been addressed in the past. Two well publicized cases include Massachusetts voter list and Netflix prize data. They are briefly discussed below:

The Massachusetts voter list is an early privacy incident which was the main motivation for proposing *K*-anonymity privacy model outlined in (Sweeney 2002b). The Group Insurance Commission (GIC) in the US published anonymized dataset which was assumed to be anonymous. This dataset included the medical visits of patients i.e., all state employees that were managed under the insurance plan. In the published dataset, the explicit identifiers were removed. However, the data included zip code, date of birth, and gender. These were sufficient to identify a large portion of population. Sweeney (Sweeney 2002b) showed that it is possible to correlate this data with the publicly available Voter Registration List of Cambridge Massachusetts which she purchased for $20. She was able to identify the former governor of the state of Massachusetts in the dataset. The problem nowadays is that, even without accessing the public voter registration list, such privacy breaches may occur due to the availability of individuals' gender, zipcode, and birthdate on social media such as Facebook (Li et al. 2011).

In the case of Netflixe prize data, Netflix released the anonymized movie rating of 500,000 subscribers for a contest. The purpose was to improve the correctness of its recommendation system. Netflix had to cancel the contest because the work in (Narayanan and Shmatikov 2008) showed that when the movie ratings are combined with the information available in the public Internet Movie Database (IMDb), it is possible to re-identify users in the released Netflix dataset with high probability.

A research done in (Tsai et al. 2011) showed that a more prominent display of privacy policies in online retailers will make consumers to incorporate privacy considerations when making online purchasing decisions. In fact, this study showed that customers tend to purchase from online retailers that better protect their privacy. Furthermore, the research shows that when privacy information is made more accessible, some customers are willing to pay more in order to purchase from sites that protect their privacy more. As a result, the study suggests that business could leverage privacy protection as their selling point.

In health care, a given study showed that Canadian health providers resist to share data with public health because of concerns due to patient privacy and performance evaluation. In another example, students were able to re-identify individuals in the Chicago homicide database by linking it with the social security death index (El Emam et al. 2011a) (El Emam et al. 2011b).

## 1.3. Privacy-by-Design (PbD)

The concept of Privacy-by-Design (PbD) was initiated by Ann Cavoukian, the Information and Privacy Commissioner of Ontario, Canada, in the '90s. The objectives of PbD are accomplished by practicing the seven foundational principles (Cavoukian 2009). These principles are listed as follows: *First*, this approach has proactive rather than preventive nature. In other words, PbD does not wait for privacy risk to occur. It tends to anticipate and prevent privacy leakages even before they occur. *Second*, PbD ensures that personal data are automatically protected in any given business practice or IT system. That is, even if an individual does not do anything, his/her privacy is preserved by default. *Third*, privacy is embedded into the architecture and design of IT systems and business practices and is considered an essential component of the core functionality being delivered. *Fourth*, PbD offers to accommodate all legitimate objectives and interests in a positive-sum manner. It avoids the false claims such as privacy vs. security and demonstrates that it is possible to achieve both simultaneously. *Fifth*, it offers end-to-end security and full lifecycle protection. *Sixth*, it assures all stakeholders that the business practice or technology operates according to the stated objectives and promises. *Finally*, PbD requires that the interests of the individuals are kept uppermost and keep the process user-centric.

KDD and PbD can be combined using the ultimate analysis task. In a bigger picture we may consider a Task Oriented Privacy-preserving KDD (TOP_KDD) process in which the KDD application is informed about privacy. This requires considering privacy in every step of the KDD process. In this work we focus on the data reduction step during pre-processing stage. Though not the focus of this work, such incorporation of privacy in every step of the KDD process could be potentially integrated into the existing KDD standards (e.g. CRISP (CRISP-DM 2001)). Integrating privacy into the CRISP-DM has been studied in the past in (Hu 2011).

To this end, we consider automatic feature selection, an effective dimensionality reduction technique and incorporate privacy into its functionality. In doing so, we effectively address two principles of PbD; namely, "*privacy as the default setting*" and "*privacy embedded into design*".

Making assumptions about the (sensitive) personal data, the background knowledge of the attacker(s), and the final analytical questions to be answered are considered fundamental assumptions for designing privacy-preserving framework (Monreale 2011). Here, background knowledge of the attacker refers to any additional information that could be obtained from external sources which would help the adversary (i.e. attacker) in breaching the privacy of a given individual in the dataset. Further assumptions may include the amount of privacy that ought to be protected for different individuals in the dataset. When we want to apply data mining, while some methods are appropriate for categorical data, some other methods are more suitable for numerical data, and so on. Some other techniques would handle both types of data. This is directly related to the final analysis task. If the goal is to build a classifier without specifying the type of classifier, then the nature of data plays an important role in selecting the appropriate classification algorithm. If, on the other hand, the goal is to build a specific classifier such as C4.5, then such knowledge may lead the data pre-processing and transforms the data to fit the goal of building a C4.5. classifier. Full privacy means not revealing anything, and full utility means not hiding anything. The goal is to find a trade-off between privacy and utility. By this trade-off, in our work, we refer to controlling the amount of privacy that could be gained without degrading our classification accuracy. Our work is based on feature selection and one main property of feature selection is that it would reduce the dimensionality of data without negatively impacting the accuracy of the resulting dataset or in many cases while improving its accuracy. We believe that such trade-off is best realized when the purpose of analysis is taken into consideration. The knowledge of adversary is yet another important element to consider. Recall that (with the exception of differential privacy) each privacy model is built to address specific attack models. Therefore, it is important to consider such knowledge by the adversary in a privacy preserving design/framework in order to obtain "reasonable" level of privacy (Monreale 2011), call it "realistic" privacy. Furthermore, not

all data attributes are potentially privacy breaching, and therefore considering all attributes to be quasi-identifier attributes is unrealistic.

## 1.4. Problem Statement

The TOP framework and some of its applications are shown in Figure 1.3. Our proposed TOP framework is different than workload-aware anonymization (e.g., (Iyengar 2002, Wang et al. 2004, Fung et al. 2005, LeFevre et al. 2006a)) in which anonymization techniques are modified in order to enhance the utility of the datasets for classification purposes. We consider TOP framework to be a general framework with different applications. One such application is workload-aware anonymization. Another application is in the area of cryptography e.g., digital credential technique proposed in (Brands 2000). We consider KDD process with the ultimate goal of turning it into a privacy-preserving process by applying privacy by design. In order to realize privacy-by-design in the context of KDD, every step of the process should become privacy-aware. In our work, we focus on step four, i.e., the dimensionality reduction step and in particular on feature selection which is a well-known dimensionality reduction technique.

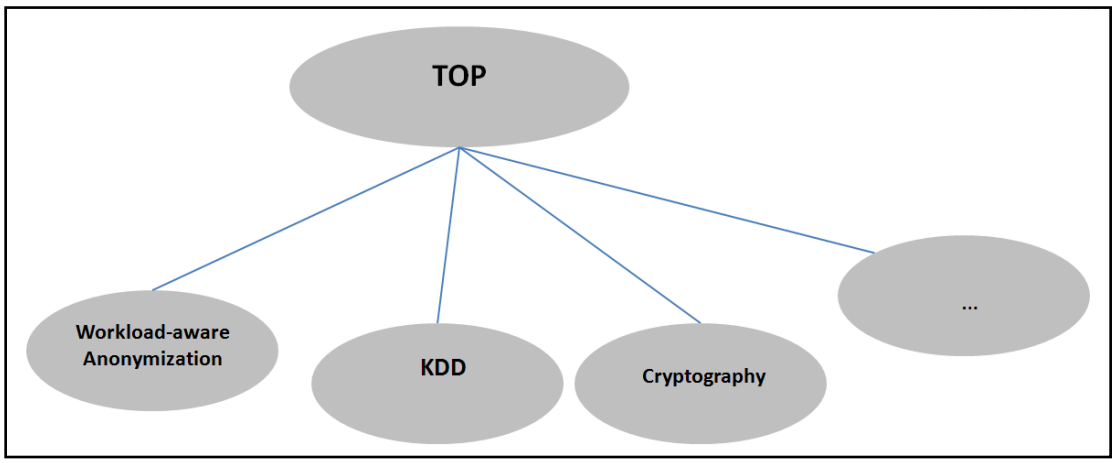



Figure 1.3: TOP Framework.

To achieve this objective, we propose different techniques to turn automatic feature selection into a privacy-aware process. This is done in a step-by-step approach via the three dimensions which constitute the contributions of our work. We start with

investigating the impact of automatic feature selection on privacy without any adjustment made to the feature selection process. We then turn automatic feature selection into a privacy-aware process. Finally, we consider privacy as another factor (in addition to accuracy) when evaluating selected attribute subsets.

Throughout our work, we consider the classification accuracy as our utility function. In all experiments, our baseline accuracy (associated with the baseline dataset) is the reference point. We either aim to improve the classification accuracy or at least ensure that under any circumstances our proposed techniques will not degrade the classification accuracy significantly lower than the baseline accuracy. We use statistical significance *t*-test to make our comparisons. In our work, privacy gain is either identified as reducing the risk of re-identification of an individual in the dataset, or reducing the ability of an attacker to infer sensitive attributes of individuals in the dataset. We evaluate our classifiers using 10-fold cross validation.

The experiments are conducted using real world datasets taken from the UCI repository (http://archive.ics.uci.edu/ml/). The summary of the datasets is shown in Appendix C in Table C.1. We considered nine of the UCI repository datasets mainly belonging to medical and financial fields. One advantage of the UCI datasets is that they are available for researchers to use and apply their techniques. It is possible to reproduce the results.

## 1.5. Contributions and Thesis Organization

Turning automatic feature selection into a privacy-aware process is a step forward in transferring the KDD process into a privacy-aware one. We assume a case in which the Data Holder (DH) wants to utilize automatic feature selection and to publish a task-oriented dataset that is tailored for a specific analysis task. This is referred to as Task Oriented Privacy-preserving (TOP) data publishing.

We assume a scenario that consists of a DH who holds the original data on the one hand and, a Data Recipient (DR) who wants the data in order to apply certain data mining task on the other hand. In our work, we assume that DR wants to classify the dataset. DH uses automatic privacy-aware feature selection and publishes a customized dataset which takes the intended analysis task of DR into consideration. Since the dataset is going to be

published in an unsecure environment, in practice, it will also be available to the attacker. However, since the dataset is tailored for a given analysis task, if the attacker uses the same dataset to do say clustering analysis instead, the results would be misleading and of less or no value. As examples, DH and DR could be a hospital and a research center respectively.

In our work we specifically consider individual's privacy. That is, the privacy concerns in this work pertain to individuals not a group of people. Considering the privacy of a group of people is addressed in the field of discrimination-aware data mining or fair data mining (Pedreshi et al. 2008) and is out of the scope of our research.

Our contributions in this work are three-fold:

(1) In the first dimension, we show that automatic feature selection (which is an effective dimensionality reduction technique) could be well utilized in order to improve privacy. In doing so, we show that feature selection can enhance existing privacy models such as *K*-anonymity and differential privacy. To the best of our knowledge, such role of feature selection as a privacy preserving tool has not been addressed in the past. We show that with feature selection, while satisfying privacy, high utility of the resulting dataset for the given task is maintained.

(2) In the second dimension, we turn the two main categories of feature selection, namely filters and wrappers, into privacy-aware processes. As such, we build a layer of privacy on top of automatic feature selection and turn it into a privacy-aware process. To this end, we introduce two systems (i.e., *PF-IFR* and *PW*) corresponding to filters and wrappers respectively. We show that using correlation among (predictor) attributes on the one hand and correlation between (predictor) attributes and the class attribute on the other hand, we can further refine the feature selection process taking into account privacy considerations.

(3) In the third dimension, we incorporate privacy into the very evaluation process of automatic feature selection. We consider privacy "during" feature selection and as such introduce a measure of privacy based on increase/decrease of attacker's ability to

infer the value of sensitive attribute of individuals. Consequently, a list of privacy-aware candidate subsets satisfying both objectives of privacy and efficacy (e.g. accuracy) is obtained. This enables the data holder to trade-off privacy and efficacy based on its preferences. In the same dimension, we introduce a (multi-dimensional privacy-aware evaluation function $E(S)$) which allows the data holder to select and release a single best attribute subset that incorporates efficacy, privacy, and dimensionality objectives according to the data holder's preferences. Such evaluation function enables the data holder to trade-off privacy, accuracy, and dimensionality of data in a controlled manner.

This thesis is organized as follows. Part I includes three chapters, namely, Chapters 2, 3, and 4 provide background information in the areas of privacy and feature selection. Part II includes four chapters. Chapter 5 summarizes the TOP framework and discusses the first dimension discussed above. The content of this chapter is mainly based on the material that appeared in:

Y. Jafer. (2014). Task Oriented Privacy (TOP) Technologies. Advances in Artificial Intelligence. 2014.M. Sokolova and P. van Beek, Springer International Publishing. **8436:** 375-380.

Y. Jafer, S. Matwin and M. Sokolova. (2014). Task Oriented Privacy Preserving Data Publishing Using Feature Selection. Advances in Artificial Intelligence. 2014. M. Sokolova and P. van Beek, Springer International Publishing. **8436:** 143-154.

Y. Jafer, S. Matwin and M. Sokolova. (2014). Using Feature Selection to Improve the Utility of Differentially Private Data Publishing. 2014. Procedia Computer Science **37**(0): 511-516.

Chapters 6 and 7 discuss privacy-aware filters and privacy-aware wrappers respectively and thus present the second dimension discussed above. The content of these two chapters is mainly based on the material that appeared in:

Y. Jafer, S. Matwin and M. Sokolova. (2014). Privacy-aware Filter-based Feature Selection. Proceedings of First IEEE International Workshop on Big

Data Security and Privacy (BDSP 2014) in IEEE International Conference on Big Data, Washington DC, USA, 1-5.

Y. Jafer, S. Matwin and M. Sokolova. Privacy-aware Wrappers. (2015). Advances in Artificial Intelligence. D. Barbosa and E. Milios, Springer International Publishing. **9091:** 130-138.

Chapter 8 presents the third dimension. It discusses the proposed framework for a privacy-aware feature selection evaluation measure. Furthermore, it discusses the proposed multi-dimensional privacy-aware evaluation function. The content of this chapter are mainly due to the material that was appeared in and submitted to:

Y. Jafer, S. Matwin and M. Sokolova. (2015). A Framework for A Privacy-aware Feature Selection Measure, Proceedings of Privacy, Security, and Trust (PST 2015), Izmir, Turkey, 62-69.

Y. Jafer, S. Matwin and M. Sokolova. A Multi-Dimensional Privacy-Aware Evaluation Function In Automatic Feature Selection, (submitted to) Transactions on Data Privacy (TDP) journal, Submission No. 20150915.

Finally, Chapter 9 concludes the thesis and present some potential future work.

# Part I

# Background and Relevant Work

# Outline Part I

In Part I we discuss some of the techniques and concepts that are most relevant to our work and constitute the basis upon which we build our methods. This part consists of three chapters. The main objective of this part is to provide background information and to set up the stage for introducing the proposed techniques in Part II.

Chapter 2 provides information about data collection and data publishing, attack models, anonymization operations, privacy vs. utility trade-off, and information metrics amongst others. In Chapter 3 we specifically discuss the relation between anonymization and classification and consider related works in anonymization where application purpose is taken into account. In addition to privacy, feature selection is the other pillar of the proposed methodology. We provide a detailed chapter about feature selection and its role in dimensionality reduction. Chapter 4 introduces feature selection techniques, the main categories, their advantages and disadvantages, etc.

In each of these chapters we make frequent connection with our proposed methodology and identify the scope of our work accordingly.

# Chapter 2

# Privacy-Preserving Data Publishing

## 2.1. Data Collection and Data Publishing

The focus in Privacy Preserving Data Publishing (PPDP) is on publishing person-specific data for the benefit of research. A typical scenario for data collection and publishing is demonstrated in (Fung et al. 2010a). In this scenario, in the *data collection* phase, the data publisher collects data from *record owners* (e.g. patients), and in the *data publishing* phase, the data publisher releases the collected data to the *data recipient*.

Two models for data publishers have been outlined in the literature (Gehrke 2005); namely, the *untrusted* model and the *trusted* model. In the first model, the data publisher is not trusted and may try to infer sensitive information from record owners. In the second model, the data publisher is trusted, and the record owners are willing to provide the data publisher with their personal information. An example of using the trusted model is when a patient gives his/her personal information to hospitals, clinics, family doctors, or when a customer reveals his/her personal information and financial details to a bank in order to obtain a loan, mortgage, etc. Our work assumes trusted publisher model at the data collection level. The question is whether this trust is transitive to the data recipient or not?

Furthermore, the recipient of the published data could be either known (private data recipient) or unknown (public data recipient). The work presented in (Bhumiratana and Bishop 2009) refers to the first case as *data sharing* and the second case is referred to as *data publishing*. In the case of data publishing, when an entity publishes the data it loses all control over how this data would be used. When the data recipient is known, it is likely that we know what this published data is intended to be used for. Knowing vs. not knowing of the data recipient needs to be discussed further in the light of trust. The general assumption in PPDP is that since we do not know how (and by whom) the released dataset will be used and since it will be available to both legitimate users and attackers, this trust is not transitive to the data recipient. However, when the recipient of the released data set is known in advance, appropriate assumptions about it being trustworthy or not could be made. When the recipient is unknown, it is more difficult to predict the ultimate usage of this data and the type of analysis that will be applied on it. The solution usually involves applying general purpose anonymization to the dataset prior to its publication.

In our work we assume that the data recipient is known and is trusted. However, this scope of trust is limited to the intended analysis task. In other words, the data recipient, though being a legitimate and recognized entity (e.g. a research center), should only see or know what is necessarily required in order to achieve the analysis task at hand. Any additional information which has no relation with the task should not be given/released to the data recipient.

### 2.1.1. Definition of Explicit Identifiers and Quasi-identifiers

The attributes in the original dataset are categorized into (Fung et al. 2010a):

*D(Explicit_Identifier(s), Quasi_Identifiers, Sensitive_Attribute(s), Non-Sensitive_Attribute(s))*

In this categorization, *Explicit_Identifier(s)* refers to a set of attributes that could identify individuals explicitly. One example is SSN in the US. An extended list of explicit identifiers could be found in HIPAA (Health Insurance Portability and Accountability Act) which is a legislation in the US that provides provisions for data privacy and security in order to protect medical data (HIPPA). *Quasi_Identifiers (QI)* is a set of attributes (e.g. demographic attributes such as zip-code) that could potentially pin point an

individual in the dataset. These quasi-identifier attributes can be obtained by the attacker (adversary) from personal knowledge or from publicly available datasets, *Sensitive_Attribute(s)* correspond to a person-specific information such as salary, disease, and so on which may be considered confidential. Finally, *Non-Sensitive_Attribute(s)* contain the attributes that do not fall into any of the above categories. These attributes are also called "natural" attributes; i.e., neither sensitive nor quasi-identifying (Brickell and Shmatikov 2008). Note that, although a single attribute in the QI set is not able to uniquely identify an individual, a combination of such attribute (i.e. the QI set) is. The very definition of quasi-identifers refers to a set of attributes that when their values are taken together could potentially identify an individual. More specifically, in the record linkage attacks the attacker already knows that a given individual's record exists in the published dataset *T*. S/he also observes the QI values from external source and if the two records match, individual's SA value could be obtained with high probability (Ke et al. 2009).

Another work categorized the attributes into three categories; namely, explicit identifiers, quasi-identifiers, and sensitive attributes (Li et al. 2007). A fundamental assumption in previous research such as *K*-anonymity, *l*-diversity, etc is that dataset's attributes can be divided vertically into *SA* and QI attributes.

In general, the original dataset with the above attribute structure (i.e. having explicit identifiers, quasi-identifiers, and sensitive attribute(s)) does not satisfy privacy requirements and therefore, it has to be modified before release. Such modification is performed via anonymization operations. Anonymization (Cox 1980) refers to hiding the identity and/or the sensitive information of individuals with the assumption that the sensitive data should be retained for analysis purposes. In the context of the above dataset format, anonymization requires that the explicit identifiers are removed from the dataset. This is, however, not sufficient. In (Sweeney 2002a), Sweeny showed that, even if all explicit identifiers are removed, an individual's record in a published medical dataset could be linked to his/her name in a public voter list. Such a case of privacy breach was verified for William Weld, the former governor of the state of Massachusetts. To prevent linking attack, the data publisher releases an anonymous dataset which would have the following format (Fung et al. 2010a):

*T(QI', Sensitive_Attribute(s), Non-Sensitive_Attribute(s))*

Note that, in this format, the *Explicit_Identifier(s)* no longer exists and is(are) removed, and the *QI* set is transformed into a less specific form (i.e. *QI'*) by applying different anonymization operations.

These operations include generalization, suppression, anatomization, permutation, and perturbation. Before discussing the anonymization operations in more detail it is important to discuss attack models and privacy models.

Privacy scenarios for privacy-preserving data sharing include interactive and not interactive. Each of these two privacy scenarios have advantages and disadvantages associated with them which are listed in Table 2.1.

In our work, we consider a non-interactive privacy scenario.

**Table 2.1: Advantages and disadvantages of nonn-interactive vs. query-based interactive privacy scenarios.**

| Privacy Scenario | Advantages | Disadvantages |
|---|---|---|
| Non-Interactive | • Offers constant availability of data<br>• No infrastructure cost is required<br>• Good for hypothesis generation and testing | • It is necessary for the data holder to specify privacy and utility requirements before sharing the data<br>• The data holder has not control over the released data since the dataset may be susceptible to attacks which have not been discovered when the dataset is released |
| Query-based Interactive | • The data holders can audit the use of their data and apply access control policies. This means that it is possible to identify attackers and hold them accountable.<br>• The protection mechanism can be updated and the data holder can provide state-of-the-art protection for the sensitive data in the dataset.<br>• Since the types of posted queries are known beforehand, it helps the data holder to decide on appropriate level of privacy when the queries are answered. | • It is difficult to support complex queries.<br>• There is usually a restriction on the number of queries that can be answered.<br>• There are analysis tasks such as visualization which require individual records not aggregate results or models and it is difficult to support such task in the interactive scenario. |

## 2.2. Attack Models and Privacy Models

The work in (Sankar et al. 2013) divides existing work in data privacy into two main categories; namely, *heuristic* and *theoretical* approaches.

Research in (Matwin and Szapiro 2010) and (Acquisti 2010) have studied privacy from an economic point of view. The work in (Matwin and Szapiro 2010) argues that approaching data privacy from an economic perspective helps us to better understand privacy as a subject of an exchange (i.e. a good) in the economic sense. This will also lead to realize multiple dimensions of data privacy including the *value of the data* to its owner and the recipient, the *cost (or effort) of attacking the data*, and the *cost of protecting the data.*

To achieve a proper privacy protection measure, three criteria have been defined in (Zhang and Zhao 2007). A privacy protection measure should reflect the fact that adversaries may have different level of interests in different data values. It should considers data providers differences in privacy concerns (for example, some people consider age to be private while others are willing to disclose it publicly), and it should satisfy the minimum necessary rule.

Due to either mutual interests, or by regulations which require that certain data should be published, there is a demand to exchange and publish data among different parties. Medical research depends on sharing data; the National Institutes of Health stated that "[w]e believe that data sharing is essential for expedited translation of research results into knowledge, products, and procedures to improve human health" (Health 2003). Sharing data is crucial for nation's security. The US Department of Homeland Security stated that "it is critical that each DHS component gives the highest priority to the sharing of potential terrorism, homeland security, law enforcement and related information" (May 2009). In the US, the National Cancer Institute initiated the Shared Pathology Informatics Network (SPIN) for researchers throughout the United States to share pathology-based data sets that are annotated with clinical information in order to discover and validate new diagnostic tests and therapies with the goal of ultimately improving patient care (Xiong and Rangachari 2008).

Dalenius (Dalenius 1977) provided a strict definition of dataset's privacy. In that definition, privacy is assumed to be preserved if accessing the published dataset does not give the attacker any information about a targeted individual that could not be obtained without accessing the dataset. Dwork (Dwork 2006) challenged this definition and showed

the impossibility of Dalenius' privacy protection due to existing background knowledge by the attacker which could be obtained even without participation of a given individual in the dataset. Such auxiliary information, i.e., information that can be obtained by the attacker even without accessing the statistical dataset was illustrated via the following example provided in (Dwork 2006). In this example, the assumption is made that the exact height of a given person is considered sensitive information. Now, let's assume that there is a statistical dataset which shows the average heights of women from different nationalities. Considering an attacker who has access to the statistical dataset and the auxiliary information "Terry Gross is two inches shorter than the average Lithuanian women". In this scenario, the attacker would identify Terry Gross' height, however, if the attacker knew only the auxiliary information, without accessing the average heights, s/he would learn very little.

Privacy is also measured as the information gain of a given observer. That is the difference between the prior belief of an observer about sensitive attribute value of an individual (i.e. before seeing the released table) and the posterior belief of the observer (after seeing the released table) (Li et al. 2007). The work in (Li et al. 2007) separates the information gain into two parts: that about specific individual and that about the whole population.

In PPDP the notion of privacy is attached to a sanitization technique based on a given attack model. The attack model may depend on the attacker's ability to uniquely identify an individual in the released dataset (i.e. *identity disclosure* (Sweeney 2002a)). Examples of privacy models that address identity disclosure includes the very well-known *K*-anonymity (Sweeney 2002b) model. The attack model could also depend on the ability of the attacker in determining/predicting the sensitive information of a given individuals (i.e. *attribute disclosure*(Li et al. 2007)). Examples of models addressing attribute disclosure include *l*-diversity (Machanavajjhala et al. 2006) and *t*-closeness (Li et al. 2007), etc. The models addressing these attacks compare privacy disclosure in terms of prior and posterior probability before and after releasing the dataset. In all of these approaches notions of Quasi-Identifiers (QI) and equivalence class play an important role. Recall from Section 2.1.1 that QI attributes refer to a set of attributes that when are "combined", could be linked to external datasets and potentially identify individuals. An equivalence class

consists of the number of tuples that share the same quasi-identifier values. In these models the assumption is made that the attacker knows the QI attributes/values of the individuals in the dataset. Choosing/identifying the quasi-identifiers is a challenging task and remains to be an open question in PPDP.

In the second category, the goal is to achieve the *uninformative principle* (Machanavajjhala et al. 2006). This principle states that a published dataset should only provide the attacker with small degree of additional information beyond his/her background knowledge. If this added information is large, the attack is called a *probabilistic attack*. In other words, this attack refers to the case where there exists a large variation between the prior and posterior belief of the attacker after accessing the dataset. Many privacy models belonging to this category do not classify attributes into QI and sensitive attribute(s) (Fung et al. 2010a). One rigorous notion of privacy in this category is the well-known $\varepsilon$-differential privacy proposed by Dwork (Dwork 2006).

Another categorization of the attacks models was provided in (Kifer 2009). The four categories of attack listed in that work are linkage, exploitation of properties of the sanitization, use of background knowledge, and reasoning how an attacker's prior belief changes into posterior belief.

Any privacy research usually makes assumptions about the attack model and its corresponding privacy model. According to (Clifton and Tassa 2013), it is important that the selection of an appropriate privacy model should be made according to the use scenario.

In the first dimension of our proposed TOP framework in which feature selection is utilized as a privacy enhancing technique, we consider *K*-anonymity from the first category (which addresses the record linkage attack) and Differential Privacy from the second category (which addresses table linkage attacks and probabilistic attacks).

### 2.2.1. *K*-anonymity

To illustrate how privacy leakage can result from publishing data, consider the following example provided in (Jian et al. 2009). A hospital releases patient records to enable researchers to study the characteristics of different diseases. The released information does not contain the names, IDs, and other individually identifiable attributes of the patients. There are, however, some attributes (i.e. QI attributes) that, combined with external datasets, enable an attacker to access confidential information and hence result in identity disclosure. In our example, a hospital releases the data in Table 2.2 after removing the names of the patients. If an attacker obtains access to the (publicly available) voter's registration list shown in Table 2.3, he could easily discover the identity of the patients by joining the two tables on {Age, Sex, Zipcode} which are considered the QI attributes.

Table 2.2: Patients' Microdata (Jian et al. 2009).

| ID | Attribute | | | |
|---|---|---|---|---|
| | Age | Sex | Zip code | Disease |
| 1 | 28 | M | 83661 | Headache |
| 2 | 25 | M | 83634 | Headache |
| 3 | 30 | M | 83967 | Cough |
| 4 | 38 | F | 83949 | Toothache |

Table 2.3: Voter Registration List (Jian et al. 2009).

| ID | Attribute | | | |
|---|---|---|---|---|
| | Name | Age | Sex | Zip code |
| 1 | Mark | 28 | M | 83661 |
| 2 | Joe | 25 | M | 83634 |
| 3 | Tom | 30 | M | 83967 |
| 4 | Judith | 38 | F | 83949 |

Table 2.4: The 2-anonymous table corresponding to Table 2.1 (Jian et al. 2009).

| ID | Attribute | | | |
|---|---|---|---|---|
| | Age | Sex | Zip code | Disease |
| 1 | 2* | M | 836** | Headache |
| 2 | 2* | M | 836** | Headache |
| 3 | 3* | * | 839** | Cough |
| 4 | 3* | * | 839** | Toothache |

Samarati and Sweeney (Samarati and Sweeney 1998a, Samarati and Sweeney 1998b) proposed the notion of *K*-anonymity in order to prevent record linkage. A table is considered *K*-anonymous if the QI values of each tuple are indistinguishable from "at least" *K*-1 other tuples. In other words, if a tuple in the table has some QI value, there are "at least" *K*-1 other records which have the same QI value. The records that share the same QI values constitute an Equivalence Class (EC). *K*-anonymity is mainly used to address record linkage attacks. Table 2.4 shows a 2-anonymous table corresponding to Table 2.2. Even though the attacker has access to the voter registration list, he could only infer that Tom may be the person in the last two records of the table. Therefore, the exact record is identifiable only with a probability of 50%.

There are a number of limitations that have been associated with *K*-anonymity (Friedman et al. 2008): First, it is not trivial for the database holder to identify the attributes that are (or are not) available in the external tables. We discuss this shortcoming further in Section 2.2.4. Second, while *K*-anonymity assumes a certain model of attack, there is practically no reason why the attacker should not try other methods. Third, the implicit assumption that within an equivalence class group, the tuples will have different private values. The deficiency of this assumption has lead to the introduction of *l*-diversity (Machanavajjhala et al. 2006) privacy model discussed in Section 2.2.2 and subsequently t-closeness (Li et al. 2007) privacy model discussed in Section 2.2.3. However, despite these limitations, *K*-anonymity is the most commonly accepted privacy model and provides the theoretical basis for privacy related legislation (Office for Civil Rights 2002). The work in (Friedman et al. 2008) identifies several advantages of *K*-anonymity including the following: (1) Rather than defining the process itself a *K*-anonymity model only defines the privacy of the output of a process, (2) *K*-anonymity is a simple, and well-understood model, and (3) It is easy to validate if the outcome is actually *K*-anonymous and therefore, non-expert data owners could be easily assured that they are using the model in a proper way, (4) The assumptions about separation of quasi-identifiers, variability of private data, and mode of attack has been so far correct in real-life scenarios. For these reasons, in our work, we consider *K*-anonymity as one of our main privacy models. There have been a number of algorithms to enforce K-anonymity.

#### 2.2.1.1. Extending K-anonymity to achieve personalized privacy preferences

According to (Wang and Zhang 2012), the essence of privacy requires that the record owners should be able to select the amount of their desired privacy preferences. In conventional anonymization techniques (e.g. *K*-anonymity), the same amount of privacy is applied on all the tuples in the raw data (microdata). The issue with such an approach (although commonly used and widely accepted) is that it does not take into account the personal preferences of individuals with respect to their sensitive attributes in the dataset. One direct implication of such "one-size-fits-all" measure could cause unnecessary utility loss without privacy gain. For instance, given a dataset of patients with different diseases, it is typically the case that a flu patient would require weaker protection compared with a patient with HIV (Xiao and Tao 2006). In another example, given an application which tracks the data for brokerage customers, it is likely that the privacy requirements by institutional investors is different than retail investors and even within a given class of customers it is possible that customer with high net-worth may require higher level of privacy protection compared with others (Aggarwal and Yu 2005). The work in (Xiao and Tao 2006) argues that when "universal" privacy standard is provided, some individuals will be over-protected while some others will be under-protected. It follows that, by controlling/setting the amount of privacy per individual, the personalized privacy approach could result in less utility loss compared with the conventional methods of exerting equal amount of privacy control over all the tuples in the dataset (Xiao and Tao 2006). Therefore, rather than imposing identical amount of privacy protection when anonymizing the dataset, the amount of generalization could be adjusted according to the personal needs/preferences of the individuals whose data are in the dataset (e.g. patients, bank customers, etc) (Xiao and Tao 2006). Personalized privacy has been also studied under the notion of stochastic privacy (Singla et al. 2014). Stochastic privacy guarantees to the users the upper bound on the probability in which their personal information will be used.

One basic technique of integrating such personal preferences into privacy preservation is to extend *K*-anonymity and *l*-diversity in order to accommodate personal preferences (Xiao and Tao 2006). In the case of *K*-anonymity, it is possible to allow every individual to

select a tailored value of *K* to determine the smallest size of QI-group desired for his/her tuples. For a dataset with a cardinality of *m*, there are *m* personalized *K*-values, each associated with one tuple (Xiao and Tao 2006). We know that based on the contents of the datasets, participating parties, and analysis tasks, the measure of privacy and utility could change. In addition to the utility requirements identified by the data recipient and the privacy requirements identified and enforced by the data holder, our framework can be easily extended to add extra degree of flexibility by incorporating the notion of personalized privacy in which it is possible to define how much protection each individual requires when the ultimate usage of the data is known in advance.

The existing research does not address personalized privacy under predetermined ultimate usage consideration. We believe that personalized privacy is naturally integrated (both conceptually and practically) into the TOP framework, and any attempt for personalized privacy preservation without considering the ultimate analysis task may result in a sub-optimal solution. By identifying the recipient and the analysis task, individuals will be more comfortable in making decisions with respect to their privacy. With such knowledge, some individuals may be willing to share their private information for the benefit of society and to have positive contribution in the advancement of research. In other cases, some individuals may be willing to reveal their private information when they are offered rewards and financial incentives in exchange.

Since the dataset is going to be published, in practice, it could also be available to the attacker. However, since the dataset is tailored for a given analysis task, while guaranteeing privacy according to the privacy model, if the attacker uses the same dataset to do other analysis (e.g. to perform clustering instead of classification), the results would be misleading and of less or no value. In other words, we want to reveal to the data recipient information "just enough" for the given purpose.

### 2.2.2. *l*-diversity

Although *K*-anonymity protects against identity disclosure, it does not provide enough protection against attribute disclosure by *background knowledge attack* and *homogenous attack*. Assuming that an attacker knows that Mark is a 28 years old male living in Zip code

83687 and Mark's record is in the table. From Table 2.4, the attacker knows that Mark has headache. This is called a homogenous attack. Now assume that the attacker knows Tom's age and zip code. The attacker concludes that Tom belongs to the second equivalence class. The attacker also knows that there Tom is low risk of toothache. Such background knowledge enables the attacker to conclude that most likely has Couth. *l*-diversity was proposed (Machanavajjhala et al. 2006) as an extension of K-anonymity in order to protect against attribute disclosure. It requires that there are at least *l* well-represented values for the sensitive attribute within each equivalence class. As such, a table is considered *l*-diverse if every equivalence class of the table has *l*-diversity.

### 2.2.3. *t*-closeness

A number of shortcomings were associated with *l*-diversity (Li et al. 2007). First, it may be difficult and unnecessary to achieve. Second, it is insufficient to prevent attribute disclosure. Two attacks against *l*-diversity were presented, namely *skewness attack* and *similarity attack*. In order to address these issues, *t*-closeness was proposed by Li et al (Li et al. 2007). *t*-closeness requires that the distribution of the sensitive values in each equivalence class is close to the distribution of the attribute in the overall table. An equivalence class satisfies *t*-closeness if the distance between the distribution of a sensitive attribute in this class and the distribution of the attribute in the whole table is not more than a threshold *t*. When all equivalence classes have t-closeness, the table is said to have *t*-closeness (Li et al. 2007). *t*-closeness does not limit the observer's information gain about the population as a whole; rather, it limits the amount of additional information that the observer could learn about a specific individual.

### 2.2.4. Challenges of Choosing Quasi-Identifier (QI) Attributes

Classifying attributes into three disjoint sets; namely, sensitive, non-sensitive, and quasi-identifiers is one of the main challenges faced by the data holder is to (Fung et al. 2010b). The main reason for identifying a given attribute *X* as a QI attribute is if the attacker can potentially obtain that attribute from other external sources.

Furthermore, misclassifying *X* into the sensitive or the non-sensitive attribute sets may compromise another sensitive attribute *Y*. For example, it is possible for the attacker to obtain X from other resources and use it to perform record linkage or attribute linkage on *Y*. In another case, misclassifying a sensitive attribute *Y* into QI may compromise the sensitive attribute *Y* belonging to some individuals. For example, it is possible for the attacker to use the attributes in QI-*Y* in order to do attribute linkage on *Y* (Fung et al. 2010b). When *Y*, on the other hand, is incorrectly included in the QI set, unnecessary information loss is incurred because of the curse of dimensionality (Aggarwal 2005).

The work in (Motwani and Xu 2008) proposed a method in order to determine the minimal set of quasi-identifiers in a data table that is capable of almost distinctly identifying a record and capable of separating two data records. However, according to (Fung et al. 2010b), identifying the minimal set of QI does not result in the most appropriate privacy protection setting. The reason is that, the proposed method does not consider the attributes that an attacker could potentially get access to. In general, the choice of identification of QI remains an open issue.

In our approach (specifically the first and second dimension), we only retain the QI attributes which would be necessary for our analysis purpose. Without having a full QI set, it becomes much more difficult for the attacker to single out an individual via identity disclosure recalling that we need QI attributes "combined".

In a real datasets, the set of quasi-identifier attributes is considered domain-specific. This includes the set of attributes that the attacker will gain access to from an external dataset such as the demographic information (Brickell and Shmatikov 2008). One method of choosing QI attributes is to consider all non-sensitive attributes in the dataset. Such larger set of quasi-identifiers, however, gives the adversary more prior information and requires heavier amount of generalization and suppression (Brickell and Shmatikov 2008). As such, large quasi-identifiers underestimate utility of the dataset and increase the risk of privacy breach. The criterion followed in (Brickell and Shmatikov 2008) in choosing QI is to keep it small in order to make adversary's task as difficult as possible. One issue is that if the sensitive attribute is also the researcher's target attribute while all attributes are quasi-identifiers, both the attacker and the researcher will try to use QI to predict the sensitive attributes. We address the case where C=SA and C≠SA in Chapter 8.

### 2.2.5. Differential Privacy

In order to respond to the need for a firm foundation for privacy preserving data publishing, differential privacy was proposed by Dwork (Dwork 2006). In general, differential privacy ensures that the existence of any individual does not, substantially, influence the outcome of any analysis on the dataset. Differential privacy requires that the answer to any query to be probabilistically indistinguishable with or without any record in the dataset. Therefore, it becomes difficult for the attacker to make inference attacks on any of the records in the dataset. Learning whether a given individual is in the dataset or not is also called *membership disclosure* (see (Dwork 2006, Nergiz et al. 2007)). Membership disclosure is the focus of differential privacy.

The ε-differential privacy is described as follows (Dwork 2006): *a randomized function K gives ε-differential privacy if for all datasets D and D' differing by at most one element, all S ⊂ Range (K),*

$$Pr[K(D) \in S] <= exp(\varepsilon) \times Pr[K(D') \in S] \quad (2.1)$$

The ratio of $Pr[K(D) \in S]/ Pr[K(D') \in S]$ is interpreted as the "knowledge gain ratio from one data set over the other" (Dwork 2008). It follows that, differential privacy requires that the knowledge gain be bounded by $exp(\varepsilon)$. From the definition of the differential privacy, if a given participant's record is removed from the dataset, the limited knowledge gain implies that no output becomes more or less likely in any significant way (Dwork 2008).

In formula (2.1), parameter $\varepsilon$ is called the privacy budget ($\varepsilon > 0$). The privacy budget is public and is specified by the data holder, and with a lower value of ε, stronger privacy is guaranteed. The value of $\varepsilon$ should be less than 1, and typically is chosen to be 0.01, 0.1, ln 2, or ln 3 (Dwork 2008).

Currently, there are two main models for guaranteeing differential privacy, including *interactive* and *non-interactive* techniques. In the *interactive* approach the data recipient (i.e. the data miner) posts aggregate queries through a private mechanism and the data holder answers those queries in response (Dwork et al. 2006). In the case of the *non-*

*interactive* approach, the data owner anonymizes the raw data and publishes an anonymized version of the dataset to the data recipient. One of the advantages associated with the *non-interactive* approach compared to the *interactive* approach is that the data recipient gets a full access to the anonymized data and, thus is given flexibility to perform the required data analysis on the complete dataset (Mohammed et al. 2011). In the *interactive* approach, on the other hand, noise is being added to each query response. The cumulative "cost" of all queries is calculated. A privacy "budget" is assigned to the dataset as a whole, and if the cumulative cost reaches the budget, a policy decision is made by the data holder to increase the amount of distortion and the noise level to a degree that the answers become useless(Microsoft 2012). As such, given the privacy budget, only a fixed number of (useful) queries could be made in the *interactive* approach.

### 2.2.6. Relationship between Differential Privacy and K-anonymity

*K*-anonymity belongs to the syntactic anonymity approaches known to be susceptible to various attacks such as record linkage, attribute linkage, etc (Section 2.2). There are also common limitations associated with the syntactic approaches such as information loss, ad hoc assumption on auxiliary information, and sub-optimality (Nguyen et al. 2013). Differential privacy is a rigorous notion of privacy; however, a study of its utility is still in its infancy (Clifton and Tassa 2013). One main limitation associated with differential privacy is the fact that it is a perturbative method and usually noise needs to be added in order to satisfy this privacy model. In privacy preserving data mining, while protecting the privacy is the main goal, the other goal is to keep the data useful for data mining purposes. A fruitful research direction is to combine the benefits associated with syntactic anonymity approaches and differential privacy. A recent attempt (Soria-Comas et al. 2013) combines *K*-anonymity and differential privacy to improve the utility of the data using the very notion of indistinguishability offered by *K*-anonymity. The authors in (Soria-Comas et al. 2013) use microaggregation to achieve *K*-anonymity, however, their technique is limited to datasets with numerical features only. (Li et al. 2011) show that "safe" *K*-anonymization followed by random sampling could achieve differential privacy. The authors of this work show that applying random sampling to the data followed by a generalization step that is

independent from the dataset (i.e. applying a fixed generalization step) and then suppression of all tuples which occur less than *K* times results in a dataset that satisfies ($\varepsilon$ , $\delta$)-differential privacy. One issue with this technique is that using a data independent generalization may lead to poor utility (Dankar and El Emam 2012). Furthermore, according to (Zang and Bolot 2011) adding a random sampling step negatively impacts the utility as well. (Mohammed et al. 2011) presented a novel technique for generalizing the contingency table then adding noise to the counts. Their technique handles both categorical and numerical attributes. According to this technique, the generalization step increases the cell counts and therefore, counts become much larger than the added noise.

Incorporating the ultimate analysis task and feature selection into anonymization is a step forward to combine the benefits of syntactic anonymization and differential privacy. This objective is discussed in details in Chapter 5 where we present the results of our work in (Jafer et al. 2014b) and (Jafer et al. 2014c). In (Jafer et al. 2014b) we show that incorporating feature selection into the anonymization process results in enhanced *K*-anonymized datasets and eventually leads to building classifiers with higher accuracy. In (Jafer et al. 2014c) we further showed that, it is possible to use feature selection and K-anonymity in order to improve the utility of differentially private data publishing.

## 2.3. Anonymization Operations

Depending on their effect on the original data, the anonymization methods are divided into two categories (Willenborg and Waal 2001), namely, *perturbative* and *non-perturbative*. In the *perturbative* category, the data is distorted before being published. Perturbation is achieved via adding noise, swapping values, aggregating values, and generating synthetic data according to the statistical properties of raw data (Fung et al. 2010b). As a result of this perturbation, unique combinations of attributes in the original dataset are replaced with new unique combinations in the perturbed dataset so that the statistical confidentiality of the data is preserved. The key point in the *perturbative* methods is that, there should be no significant differences between the statistics obtained from the perturbed dataset compared with the statistics computed from the original dataset (Domingo-Ferrer 2008). One issue with these techniques is their failure to preserve data truthfulness (e.g. changing patient's age from 60 to 6) which can severely impact the utility

of the published data in domains such as healthcare (Gkoulalas-Divanis and Loukides 2013).

In the *non-perturbative* category, the data is modified such that the data truthfulness in retained while its detail is reduced. This category includes anonymization operations such as generalization and suppression, and, anatomization and permutation (Fung et al. 2010b). Detailed description of generalization and suppression and additive noise operations is given in the following sections. These anonymization operations are used in our selected privacy models, i.e. *K*-anonymity and differential privacy.

### 2.3.1. Generalization and Suppression

Generalization and suppression result in a dataset that is semantically consistent with the raw data. In the case of generalization, we generalize the attribute values of the quasi-identifiers in order to reduce the granularity of their representation. In other words, we replace some values with a parent value in its corresponding taxonomy tree. For example, age = 27 will be replaced by [20-30]. As for suppression, we replace some values with a special value so that the replaced values do not get disclosed. For example, Zip Code = 83661 will be replaced by 83***. Different algorithms that implement *K*-anonymization use these operations in order to satisfy the *K*-anonymity privacy model.

### 2.3.2. Additive Noise

Additive noise belongs to the category of random perturbation. An example is to add Laplace noise to achieve differential privacy. The main idea in random perturbation is to use synthetic data values to replace the original data values such that the statistical information computed from the original data is similar to the statistical information computed from the perturbed data. Since the perturbed data does not correspond to real world data it is not possible for the adversary to perform linkage attacks, or to confidently infer sensitive information from the published data (Fung et al. 2010b). Additive noise is considered an important and frequently used privacy preserving technique in statistical disclosure control (Adam and Worthmann 1989, Brand 2002).

## 2.4. Privacy vs. Utility Trade-off and Utility Measures

In addition to preserving the privacy when publishing a dataset, it is equivalently important to ensure that the released data is useful for further analysis (Fung et al. 2007). Protecting the privacy, alone, without considering the usefulness of the modified dataset for future analysis may make the dataset useless. At one extreme end of the spectrum, if we suppress all of the records in the dataset, the privacy of the individuals is 100% guaranteed but the utility of the dataset is completely destroyed. At another end, if we release the raw data, as is, without any modification the utility of the dataset is fully preserved but privacy of individuals is endangered. Therefore, it is crucial to find a balance between the amount of privacy and utility to make the dataset useful and yet privacy preserving. We believe such balance can be fully realized when the ultimate task guides the modification/anonymization process by means of 'intelligent anonymization'.

Let us consider the following example borrowed from (Xu et al. 2006) which consists of patients' records that are used for disease analysis. To anonymize this dataset, it is possible to generalize the *zip_code* attribute from a five-digit full zip code to three digit prefix (e.g. from 64534 to 645**). It is also possible to generalize the *age* attribute to *age_group* (such as [20- 25] instead of '24'). Consider the following two analysis examples:

Task 1: An analyst wants to study the relation between disease and age. Solution: In such case, the age information is critical for disease analysis more than the exact location of the patients. Therefore, we need to preserve more details of the age attribute while use coarser generalization for the *zip_code* attribute

Task 2: In another study, an analyst wants to indentify disease spread geographically. Solution: This may requires very precise information about the city and street names and a less detailed information about the age.

Given the same published dataset; from a researcher point of view, we are more interested in the utility of the dataset. As such, we may hide or modify attributes that have no or small values for the intended task and that may lead to high privacy risks. In the above example, if age and zip_code are considred QI attribute, while satisfying *K*-anonymity, guided by the analysis task we can adjust the level of generalization to retain more information.

The research in (Bhumiratana and Bishop 2009) proposed a model to balance privacy and utility of data. In that model, both the data provider and the data user are allowed to negotiate these two requirements until a satisfactory balance is reached, or rather, to determine that such a balance is not reachable. This is an interactive model in which the data publisher has control over the data after data publishing. Other research such as (Iyengar 2002, Bayardo and Agrawal 2005, LeFevre et al. 2006b) introduce utility metrics to guide the transformation process; however, they do not take into account the importance of attributes. (Brickell and Shmatikov 2008) proposed a technique where they directly compare the privacy loss with the utility gain caused by anonymization. In their technique, the privacy loss is measured as the improvement of adversary's ability in guessing individual's sensitive attributes values. The utility gain, on the other hand, is measured as the improvement of researcher's ability in building accurate classification model. This work shows that in order to achieve even modest privacy, the data mining utility will be completely destructed (Brickell and Shmatikov 2008). The same work argues that the assumption that, privacy gain equals utility loss is not reasonable since privacy and utility have very different characteristics, and therefore in any privacy preserving data publishing/mining effort, studying privacy and utility should be conducted in parallel. Therefore it is crucially important to measure both privacy and utility using the same methodology because otherwise, maximizing utility will inevitability result in violation of privacy (Brickell and Shmatikov 2008).

With respect to the relation between privacy and utility, (Li and Li 2009) provide three reasons to why these two objectives cannot be directly compared. These reasons are summarized as follows:

First, Privacy is an individual concept which should be measured separately for every individual. However, utility is an aggregate concept that must be measured in an accumulative way for all useful knowledge. This implies that for privacy, the worst-case privacy loss should be measured. However, for utility we need to measure the aggregated utility. Second, specific knowledge (which is about a small group of individuals) has a higher impact on privacy whereas aggregate information (which is about a large group of individuals) has a higher impact on utility. Third, any information which deviates from a given prior belief, being correct or false, may result in privacy loss but only correct

information contributes to utility. It follows that, privacy should be measured against the trivially-anonymized data (the case where all quasi-identifiers along identifiers are removed from the dataset) whereas utility should be measured using the original data as the baseline.

Two main characteristics of the data utility are that it is *relative* and *specific* (Hua and Pei 2008). It is *relative* because rather than considering the absolute utility of the data, we measure how much utility is preserved in the anonymized dataset compared to the original dataset. It is *specific* since different applications need different information in the dataset which leads to conclusion that, in designing utility measures, the context of certain applications should be taken into consideration. Therefore, the use of the published data is considered important in determining the appropriate utility measure. These measures include *query answering accuracy*, *classification accuracy*, and *distribution similarity* (Hua and Pei 2008). In the case of *query answering accuracy*, the data quality depends on how far away each attribute value is from the original data after the dataset is modified and published. For instance, when generalization is used to anonymize the dataset, specific values of attributes are replaced by more general values. One important technique that uses the query answering accuracy as a utility measure is differential privacy (Dwork 2006) (Section 2.2.5). In the case of *classification accuracy* the released data is usually used to train a classifier and to build a classification model. The data quality is therefore measured based on how well the class structure is retained. Traditionally, data mining has been employed to measure the usefulness of the sanitized dataset (Sramka 2010). This is done via measuring and comparing the prediction accuracy over the original and the sanitized datasets. In our work, we consider classification accuracy and similarity in clustering to be our utility measures (i.e. utility functions).

Any reasonable trade-off between utility and privacy requires knowing the maximum utility achievable for a given level of privacy and vice versa (Sankar et al. 2013). The work in (Sankar et al. 2013) argues that this could be achieved by borrowing an important concept from information theory called *rate distortion theory*. That is, utility can be quantified through distortion and the rate distortion should be augmented with privacy constraints quantified via equivocation which is related to entropy.

Determining the amount of desired/required utility is essentially task dependent, and it is realistic to make a judgment about the obtained utility according to the predefined ultimate analysis task. The work in (Sramka et al. 2010) argues that similar to the utility, privacy requirement is also scenario dependent; for example, in one case a selected group of individuals need higher protection for specific attributes, whereas in another case the same privacy level is required for all of the users. This is directly related to the personalized privacy concept which was discussed in Section 2.2.1.1.

Constraint-based anonymization of transactions was studied in (Loukides et al. 2011). This was the first approach which anonymizes transactional data under application-specific privacy and utility requirements. These requirements are modeled as constraints which are satisfied using an algorithm that anonymizes transactions using a flexible anonymization scheme. This approach follows a greedy fashion and performs both generalization and suppression. The selection of generalized item is guided by the utility constraints which are specified by the data publisher. They correspond to the most generalized items that could be used to replace a set of items. As such, they limit the generalizations to items that are acceptable for the given application. When it is not possible to generalize an item with respect to a specific utility constraint, this algorithm suppress that item to ensure privacy.

Depending on whether the ultimate usage of the anonymized dataset is (not) taken into consideration during anonymization, two major categories of approaches have been introduced in the literature. These categories include general purpose and specific purpose anonymization and are discussed in the next section.

## 2.5. Information Metrics

In order to measure the data usefulness, two broad categories of information metrics exist, namely, *data metric* and *search metric*. *Data metric* measures the data quality in the anonymous table and compares it to the data quality in the original table. *Search metric* guides an anonymization algorithm - step by step - to find an anonymous table that has maximum information, and minimum distortion. The identified anonymous table is eventually evaluated using data metric. Information metric is also categorized based on its

information purpose into *general purpose metrics*, *special purpose metrics*, and *trade-off metrics* (Fung et al. 2010a).

### 2.5.1. General Purpose metrics

The goal stated for privacy preserving data publishing is to produced sanitized datasets which have a "good" utility with respect to various workloads (Brickell and Shmatikov 2008). The point is that the unknown workload is an essential premise in privacy preserving data publishing. The workload-independent measures have been proposed for this purpose. One approach is to minimize the amount of suppression and generalization (Ciriani et al. 2008). Such minimization is measured via relative distance, absolute difference, maximum distribution, or minimum suppression. Other work considered other syntactic metrics such as average size of quasi-identifiers equivalence classes, the number of generalization steps, the sum of squares of class sizes (Machanavajjhala et al. 2006), as well as preservation of marginals as it was shown in (Kifer and Gehrke 2006). The observation made by (Brickell and Shmatikov 2008) is that although workload-independent metrics quantify the amount of "damage" caused by santitization, they do not measure the amount of utility that remains. Such unclear correlation between information loss metric and future use of data was pointed out in (LeFevre et al. 2006b). The problem is that the definition of utility is ill-defined on some unknown use of the data in the future (Cormode et al. 2013). This is why it has been recognized that the utility of sanitized datasets should be measured empirically with respect to specified workloads such as classification algorithms (Iyengar 2002) (Wang et al. 2005, LeFevre et al. 2006a).

The *General purpose metrics* are used when the data publisher does not know how the published data will be used by the recipient. Reasonable information metric in such scenarios is to measure "similarity" between the original data and the anonymous data which is related to the *principle of minimal distortion* (Sweeney 1998, Samarati 2001, Sweeney 2002b). One method is to minimize the amount of generalization and suppression that is applied to the quasi-identifier attributes while achieving a given privacy level (Ciriani et al. 2007). Other metrics are the number of steps required to perform generalization, the average size of quasi-identifier equivalence classes, preservation of marginals (Kifer and Gehrke 2006), and the sum of squares of class sizes (Machanavajjhala

et al. 2006). Workload-independent metrics while quantifying the "damage" that is caused by sanitization, they do not measure how much utility remains in the dataset (Brickell and Shmatikov 2008). The special purpose metrics are discussed next.

### 2.5.2. Special Purpose metrics

The *Special purpose metrics* are used when the purpose of the data is known and can be taken into account during anonymization in order to retain information. This is directly related to our work. For example, if the data is published for a classification task, the values whose distinctions are essential for discriminating the class labels in the target attribute should not be generalized (Fung et al. 2010a).

Without a workload context, stating whether a dataset is "useful" or "not useful" becomes irrelevant (Brickell and Shmatikov 2008). The stated goal of PPDP is to produce santitized datasets that have "good" utility for different kinds of workloads. However, one often-asked question is that, if the ultimate goal of data usage is known beforehand why not, instead, publishing the data mining results (i.e. a classifier)? According to (Nergiz and Clifton 2007), publishing data mining results is a commitment at the algorithmic level. This is an impractical solution for the non-expert data publisher. Furthermore, there are many ways to mine the data even if the purpose is given and, typically, it is not clear which one is the best unless the data is received and different ways are tried (Fung et al. 2010a). In our TOP framework, we address this particular concern. We insist that the data recipient is still in charge of the data mining and model building process and will generate the data mining model. The difference is that, rather than accessing the data directly (which is prohibited due to privacy restrictions), it views a subset of data that best fits its analysis goal and no further details are released by the data holder.

Let us consider the following example borrowed from (Fung et al. 2010a): we have a classification problem in which the goal is to classify future cases into some predetermined classes drawn from the same underlying population as the sample data in the published data. The sample data contains both the useful classification information which could improve the classification model and the useless noise which can degrade the classification model. The useful classification information (which clearly should be retained) holds for the sample data and the future data whereas the useless noise holds only for the sample

data. For instance, a patient's birth year would most likely to be part of the information that is required to classify lung cancer if the disease occurs more frequently among elderly people; however, the exact birth data is likely considered to be noise. In such cases, by generalizing the birth date to birth year we actually eliminate the noise and achieve better classification accuracy.

The first work on privacy preserving data publishing for classification was presented by (Iyengar 2002). That work proposed a classification metric, a.k.a. *CM*, to measure the classification error on the available sample data. This method charges a penalty for each record that is suppressed or generalized to a group where the record's class is not the majority class. It follows that, a record which does not have a majority class in the group will be classified as a majority class. This is clearly incorrect and contradicts the record's original class. When the data mining task (e.g. classification, regression, etc) is known in advance, this knowledge could be used in order to improve the anonymization process and eventually improve the quality of the results. Such knowledge also impacts the utility metric used (Section 2.5). For instance, we may use accuracy for classification applications and intra-cluster similarity and inter-cluster dissimilarity for clustering applications.

Such incorporation of the data mining task into the anonymization process has an important implication. Recall from Section 2.5.1 that, general purpose anonymization aims at minimizing the distortion of the dataset while preserving the privacy according to a given privacy model. The importance of incorporating the data mining task in the anonymization process when the data mining task is known in advance (e.g. classification) was outlined by (Fung et al. 2007). Their work offered a *K*-anonymization solution for classification with the goal of finding a *K*-anonymized dataset which is not necessarily optimal in terms of minimizing data distortion on the sample data but, provides a solution with the aim of preserving the classification structure. According to this work (Fung et al. 2007), it is more important for the classifier to use the structure which will be repeated in the future data rather than noise which would occur only in the sample data. Usually, the sample data consists of overly specific "noise" which are harmful for classification. Therefore, in order to construct a classifier, noise should be generalized into patterns which are shared by more records in the same class. The data also consists of redundant structure.

The work in (Fung et al. 2007) argues that the cost metric for anonymization should be measured via the classification error on the future data. It is not correct to replace this cost metric by the classification error on the masked table due to the fact that a perfect classifier for the masked table may be inaccurate for the unseen data.

Utility in (Cormode et al. 2013) is defined as the empirical utility and described as the (relative) error of COUNT(*) queries with range conditions of the attribute. The work in (Brickell and Shmatikov 2008) argues that this does not contradict the "unknown workload" principle of sanitization necessarily. Rather, it shows that even when syntactic damage minimization requirement is satisfied, such sanitization may destroy the utility of a dataset for certain data mining tasks.

The findings of (Fung et al. 2007) are concluded as follows: this work considered the problem of satisfying the anonymity of individuals while releasing person-specific data for classification analysis. The authors argued that the existing optimal *K*-anonymization based on a closed form of cost metric does not address the classification requirement. Two observations were considered in this approach: information specific to individuals tend to be overfitting and therefore has little utility value to classification. It is possible that a masking operation eliminates some useful classification structures and alternative structures in the data emerge to help. As a result, not all data items are useful for classification and less useful data items may provide the room for anonymizing the data without compromising the utility. Our results shown in Chapter 5 verify this observation.

## 2.6. Multiple-View Data Publishing

Most of the algorithms in the area of PPDP/M consider, only, a single release of a given dataset. An extension to this model which has gained less attention considers multiple-view data publishing. It is observed that when multiple-views of the same data set are being released, although each of the individual releases is anonymized, it is still possible for an adversary to break the anonymity when comparing different anonymized views (Fung et al. 2010b). The assumption of multiple releases from the same dataset has been addressed in (Nergiz et al. 2009, Stokes and Torra 2012).

Few projects considered potential privacy breach resulting from linking two or more views. (Yao et al. 2005) proposed a technique for detecting violation of *K*-anonymity on a set of views. According to their work, the views are obtained by performing projecting and selection query. Another work by (Kifer and Gehrke 2006) considered releasing additional marginals in order to increase the utility of published data where marginals are duplicate preserving projection views.

Multiple-view data publishing is directly related and applicable to our TOP data publishing. If the analysis task is the main factor that guides the anonymization process, it would be legitimate to assume that the same dataset may be anonymized and released for different purposes. In order to consider multiple-view publishing in the context of TOP data publishing the data holder must ensure that if multiple views of the same dataset are requested by the data recipient(s), the union of all published anonymized views is still *K*-anonymous. This is achieved by introducing an anonymization budget, partially consumed, when an additional view of the dataset is being requested. If the total allowed budge is consumed the default case should rule and instead of task oriented anonymization, a general purpose anonymization is enforced. We consider multiple-view data publishing as a possible future work and will not discuss it further in our work.

## 2.7. Summary

This chapter provided background information about privacy-preserving data release including data collection and data publishing, possible attacks and their corresponding privacy models and anonymization operations. It also listed and analyzed some of the works considering privacy-utility trade-off measures, information metrics such as general purpose and special purpose metrics. A special attention was given to special purpose metrics due to their direct relation with TOP. The aforementioned topics were discussed and their potential relation with our current work was discussed. In the next chapter, we focus of background information and related work in the area of privacy preserving data mining with a focus on classification which is the core data mining task that is used throughout our work.

# Chapter 3

# Privacy Preserving Data Mining

Data mining and knowledge discovery in datasets aims at automatically extracting previously unknown patterns in a large amount of data (Han et al. 2012). In the area of privacy preserving data mining, the data mining algorithms are analyzed for their impact on data privacy. The goal of privacy preserving data mining is to develop algorithms to modify the original dataset so that the privacy of confidential information remains preserved and as such, no confidential information could be revealed as a result of applying data mining tasks. Data mining tasks include classification, clustering, association rule mining, etc.

## 3.1. Relations between *K*-anonymity and Classification

Let us consider the case where a given *K*-anonymous table will be used for data mining. Such knowledge about the purpose for which the anonymized data will be used has important implication. As it was outlined in (Ciriani et al. 2008), when the data is intended for data mining, the aim of *K*-anonymity should not be the minimization of information loss, rather, optimizing a measure that is suitable for data mining purpose.

Different methods have been proposed in the past that aim at protecting individuals' privacy (through achieving *K*-anonymity and/or *l*-diversity) while preserving the data utility for certain data mining tasks such as classification. These works mainly focused on optimizing anonymization for classification applications. All of these works consider special purpose metric.

(LeFevre et al. 2006a) proposed a suite of greedy algorithms in order to address the *K*-anonymization problem for a number of analysis tasks such as classification and regression analysis for single/multiple categorical and single/multiple numerical target attribute(s) respectively. This technique provides a very flexible framework for anonymizing different attributes. The technique employs multidimensional generalization such as Mondrian (LeFevre et al. 2006b) that helps reducing information loss with respect to classification and regression analysis. However, there are two issues associated with this technique (Fung et al. 2010b). Firstly, multidimensional generalization leads to higher computational cost due to data exploration problem. Second, the fact that Mondrian algorithms enforce traditional *K*-anonymity and *l*-diversity results in high information loss due to the curse of dimensionality (Aggarwal 2005). The work in (LeFevre et al. 2006a) argues that it is best to judge quality with regard to the workload for which the data will be eventually used.

In a number of projects (Iyengar 2002, Wang et al. 2005, LeFevre et al. 2006a) it has been shown that the utility of sanitized datasets must be measured empirically. Such observation does not contradict the "unknown workload" premise of sanitization.

Research such as top-down specialization (Fung et al. 2005), bottom-up generalization (Wang et al. 2004), and genetic algorithm (Iyengar 2002) used the target classification model in order to evaluate the recoding. These works evaluated the *K*-anonymization results according to a given data mining task. In the top-down specialization (Fung et al. 2005), data is anonymized according to the class conditional entropy measure. For example, top-down specialization is an anonymization technique designed for building accurate decision trees based on anonymized datasets. The work in (Iyengar 2002) is the only work which evaluated the impact of anonymity on classification with single dimensional generalization.

### 3.1.1. Anonymization Based on the Application Purpose

The work in (Byun et al. 2005) proposed a comprehensive approach for privacy preserving access control based on the notion of *purpose*. Hippocratic databases (Agrawal et al. 2002, Agrawal et al. 2003) used *purpose* in order to enforce fine-grained disclosure policies at data level.

There are works that considered the notion of application purpose during the anonymization process. (Xiong and Rangachari 2008) presented an application-oriented approach for data anonymization which considers the relative attribute importance for the target applications. Their work identifies three types of target applications, namely, *query applications supporting ad-hoc queries*, *applications with a specific mining task* such as classification or clustering, and *exploratory applications without a specific mining task*. The work in (Xiong and Rangachari 2008) characterizes the attributes with respect to the applications into *selection attributes, feature attributes*, and *target attributes*. *Selection attributes* are those used to identify a subpopulation, *feature attributes* are used to perform analysis such as clustering or classifying, and *target attributes* are the class attributes corresponding to attributes that classification or predication are trying to predict. As such, *target attributes* are only applicable to supervised learning tasks such as classification. The authors proposed a prioritized anonymization scheme in which the attributes are prioritized based on how critical and important they are for a given application. According to (Xiong and Rangachari 2008), these priority weights could be either explicitly specified by the users or implicitly learned by the system based on a set of sample queries and analysis. In the latter case, the more frequently an attribute is queried, the more crucial it may be to the application which follows that it should undergo less generalization. One main shortcoming associated with (Xiong and Rangachari 2008) is the fact that such attribute priority assignment by the user is impractical in many scenarios and "it is not always possible for users to specify that attribute priorities beforehand" (Xiong and Rangachari 2008). A recent work by (Pattuk et al. 2015) explores the role of selecting features to preserve privacy in a dynamic client-service provider environment. The idea is that during model deployment, only a subset of the available features are requested by the service provider from the client. The goal is to produce results with maximal confidence

and yet minimize the ability of violating a client's privacy. The authors proposed an iterative approach in which the server requests information about of one feature at a time until the client-specified privacy budget is consumed. The feature selected in each step depends on the previously selected features and their values.

In a relevant work, (Sun et al. 2009) used the notion of purpose in terms of applications queried by different data requesters. They provide an example in which the same dataset may be queried by a PhD student for the disease investigation purpose and by a research center for population analysis. Thus the disease attribute which is necessary for disease investigation may not be necessary for the population analysis. Therefore, the purpose of the request determines the priority of each attribute. The work in (Sun et al. 2009) aimed at developing a "much finer" data anonymization strategy that takes both the reliability of the data requester (i.e. trust), and the application purpose into consideration. The authors highlight the importance of the purpose of the request in determining the priority of each of the attributes. By specifying priorities, the data requesters are able to determine the acceptable degree of generalization and information loss.

In order to address the shortcoming associated with (Xiong and Rangachari 2008), the work in (Sun et al. 2009) devises a method to automatically derive the attribute priorities by adopting the concept of entropy to measure the independency among attributes. However, one main disadvantage associated with this technique is that it requires the most useful and the least useful attribute to be determined by the user. Only then, it is possible to derive the attribute priority of the set of attributes based on the measure of independency between them.

Our work is different from the above proposals. We neither make assumptions about the most relevant and the least relevant attribute with respect to the target class nor do we set the priority of the attributes. We employ feature selection in order to automatically obtain/retain the list of the most important attributes that are relevant to the (classification) task at hand taking privacy into account and to the best of our knowledge such solution was not proposed in the previous works. Such technique is crucially important when we are dealing with high-dimensional data.

## 3.2. Anonymization in Data Mining

In general, two approaches have been proposed to preserve *K*-anonymity in data mining (Ciriani et al. 2008): First approach is to anonymize the data and then perform data mining, and second approach is to perform data mining on the original data and anonymize the data results later.

### 3.2.1. Privacy of the Models Built Using Anonymized Data

One main advantage of the Anonymize-then-Mine approach is the decoupling of data protection and data mining (Ciriani et al. 2008). Such decoupling, according to (Ciriani et al. 2008), has two benefits: First, it guarantees a "safe" data mining practice since it is performed on the anonymized dataset and by definition the data mining results can not violate e.g. the k-anonymity of the original table. Second, it provides a degree of flexibility by allowing the data mining to be performed by other parties (other than the data holder). This relaxes the condition when the data holder is not a data mining expert.

However, there are two challenges associated with this approach. First, could such anonymized data be effectively used to build accurate models? Second, could the resulting model (which was built using anonymized data), itself, breach the privacy?

The work in (Inan et al. 2009) argues that, in some data mining tasks, it is possible to find effective answers to the first question. One such data mining task is to build a decision tree. To do so, it is possible to expand the domain of quasi-identifier attributes when adding generalized value. Several papers have addressed such a case. For example, top-down specialization (Fung et al. 2005) is an anonymization techniques for building accurate decision trees based on anonymized datasets. In that approach, the data is anonymized according to the class conditional entropy measure. Other papers in (Martin 1995, Zhang and Honavar 2003, Zhang et al. 2006) use anonymized data to perform classification.

Despite the above, in some other data mining tasks, mining anonymized data becomes more challenging (Inan et al. 2009). An example of such difficulty is the case where the data mining method requires distance computation between attributes. Inan et al. (Inan et al.

2009) proposed a new approach to build a classifier using anonymized data to address this requirement.

To answer the second question, few studies have investigated potential privacy breaches in the case of association rule mining. The work in (Aggarwal et al. 2006) argues that suppressing the sensitive values that are selected by individual data contributors is insufficient since the attacker could use association rules that are learnt from the data in order to estimate the suppressed values. In their proposed solution, they introduced a heuristic algorithm to hide a minimal set of entries to prevent the privacy violations by such attacks. (Verykios et al. 2004a) proposed algorithms to hide sensitive associate rules in a given transactional dataset. Their general idea was to hide one rule at a time by decreasing its confidence or its support which is achieved by removing items from the transactions. As such, rules that satisfy a specified minimum support and minimum confidence will be removed.

The above studies consider association rule mining in particular and much research is yet needed to investigate the potential output privacy breach even when the data mining model is built using anonymized data.

(Friedman et al. 2008) identified two problems with building data mining models from anonymized data. First, there is a performance cost of the anonymization process which may be very high, particularly in large and sparse datasets. As a result, the cost of the anonymization may even exceed the cost of mining. Second, anonymization may inadvertently result in deleting features which are crucial for data mining and leaving out attributes that are useless. Thus, "it is more reasonable to perform data mining first and anonymization later". The same problem was highlighted in (Ciriani et al. 2008) that due to anonymization, the mining is performed on "less specialized and complete" data which could eventually impact the significance and the usefulness of the data mining results.

Most research in the area assumes that the classifiers will be trained and tested over anonymized data. We follow this assumption in our work. In our framework, by integrating the application purpose and identifying ultimate analysis goal we are able to obtain anonymized datasets that implicitly address the above two problems. We show that by employing feature selection we substantially reduce the dimensionality of high dimensional datasets and as a result, lower the cost of anonymization. Furthermore, rather

than deleting features which turn out to be important for a given data mining task, with feature selection, we instead remove redundant and irrelevant features and keep the ones that are relevant to the task at hand.

### 3.2.2. Privacy of the Models Built Using Original Data

In the Mine-then-Anonymize approach, the data mining models are built based on the original datasets and then, anonymization process is applied on the data mining results. The work in (Atzori et al. 2008) suggested that if the goal is to release data mining results (e.g. frequent patterns), then it should be sufficient to anonymize the patterns rather than the data itself. According to their study, such approach results in a much better information utility compared with performing data mining on anonymized data.

In the approach of building the models using the original data, *K*-anonymity needs to be satisfied either after data mining is completed (a *two-step* method) or within the data mining process itself (a *one-step* method) (Ciriani et al. 2008). In either case, the mined results should not enable inference of the existence of sets of quasi-identifier values which have less than *K*-occurrences in the original dataset. The *one-step* method requires modification of the mining process. That is, it requires modifying the data mining algorithms in order to enforce *K*-anonymity directly. However, it usually results in performance advantages (Ciriani et al. 2008). The work in (Friedman et al. 2008) identifies two major benefits when the data mining techniques are used as the basis for k-anonymization. First, the anonymization algorithms are optimized in order to preserve specific data patterns based on the data mining technique. Second, when the anonymization algorithms are employed based on data mining techniques, instead of having the same generalization for all tuples, different generalizations for several groups of tuples may be applied. This results in retaining more useful information. It follows that, one of the main advantages of this approach is higher quality of the data mining results and increased efficiency. This approach has been studied with respect to association rules (Atzori et al. 2008, Friedman et al. 2008), decision tree (Fung et al. 2007, Friedman et al. 2008), and clustering (Friedman et al. 2008).

One main disadvantage of this approach is that data mining can be performed by the data holder only (Ciriani et al. 2008). Such requirement negatively impacts the applicability of the method since it implicitly requires the data holder be a data mining expert.

Table 3.1 lists the advantages and disadvantages associated with each of the approaches. By integrating the concept of application purpose and allowing both approaches in our framework, we benefit from advantages of each of the two approaches while eliminating their shortcomings.

Table 3.1: Comparison of 'anonymize-then-mine' and 'mine-then-anonymize' approaches.

| Approach | Advantage | Disadvantage |
|---|---|---|
| Anonymize-then-Mine | • Flexible: Allows parties other than the data owner to perform data mining | • Anonymization cost specially in sparse and high dimensional datasets<br>• Information loss due to anonymization which may negatively impacts the quality of data mining results |
| Mine-then-Anonymize | • Better models are obtained since data mining algorithms are applied on original data<br>• Possible to have different generalizations for several groups of tuples | • Not flexible: Data mining can be performed by the data owner only |

We aim at presenting a framework that is constructed around the concept of utilizing the intended task in order to guide the data publishing and data sharing process. In our proposal, one method is that we anonymize the dataset based on the task and the desired amount of privacy and utility. In other words, unlike the strategy followed in differential privacy (Dwork 2006) in which the utility is achieved based on privacy goals, we first define the utility goal (since the analysis task is already defined) and then explore how privacy can be achieved given that the utility goals will be satisfied. In this strategy, we may hold the utility as a constant and the privacy as an adjustable variable based on intended task. This is further explored as we discuss our work in the coming chapters.

## 3.3. Summary

In this chapter, we considered the relation between *K*-anonymity and classification. We showed and discussed the workload-aware anonymization works in literature. We also discussed related work that consider anonymization based on application purpose. Finally, we considered the two major approaches to privacy of models. This includes models that are built using anonymized data vs. models that are build using original data where privacy is then incorporated into the model itself. In the next chapter we will focus on feature selection considering two major techniques of filters and wrappers.

# Chapter 4

# Feature Selection

*Data quality* in literature has been defined in terms of completeness, consistency, accuracy, timeliness, interpretability, and believability of the data (Han et al. 2012). These qualities are accessed depending on the intended use of the data.

Several preprocessing methods have been introduced in the literature in order to improve the quality of data for analysis purposes. These methods include data cleaning, data integration, data reduction, and data transformation (Han et al. 2012). Data cleaning is applied to correct inconsistencies and to eliminate noise in data. Data integration deals with merging data from different sources into a coherent data store (e.g., data warehouse). Data reduction involves reducing the size of data by applying aggregation, clustering, or removing redundant features. Finally, data transformation aims at converting the data into appropriate forms in order to use them for mining purposes. The data transformation techniques include normalization, concept hierarchy generation, and data discretization.

For example, normalization is applied in order to scale the data to fall in a smaller range (e.g. between 0.0 and 1.0).

We particularly focus on data reduction. One main reason is that when dealing with privacy, we mainly aim at reducing the amount (and extent) of the released data. Therefore, it seems natural to study and investigate privacy in the context of and in relation with data reduction. The three main data reduction strategies include *dimensionality reduction*, *data compression*, and *numerosity reduction*. These strategies along their definition and examples are summarized in Table 4.1.

Table 4.1: Data reduction strategies with examples.

| **Data reduction strategy** | **Definition** | **Techinques** |
|---|---|---|
| Dimensionality reduction | Reducing the number of random variables or attributes. | • Wavelet transforms<br>• Principal Components Analysis (PCA)<br>• Attribute subset selection |
| Numerosity reduction | Replacing the original data volume by smaller forms of data representation. Two types: *Parametric*, *Nonparametric* methods.<br><br>*Parametric*: a model is used in order to estimate the data. As such, only the data parameters are stored.<br><br>*Nonparametric*: used to store reduced representations of the data. | *Parametric*<br><br>• Regression models<br>• Log-linear models<br><br>*Nonparametric*<br><br>• Histograms<br>• Clustering<br>• Sampling<br>• Data cube aggregation |
| Data Compression | Obtaining a "compressed" representation of data. Two types: *Lossless*, *Lossy*<br><br>*Lossless*: The original data can be constructed from the compressed data without losing information.<br><br>*Lossy*: Only an approximation of the original data can be constructed. | • Algorithms for string compression<br>• Dimensionality reduction and numerosity reduction techniques are forms of data compression. |

We focus on attribute subset selection in particular. In machine learning, attribute subset selection is also known as feature subset selection and we do use them interchangeably. Features are the "workhorse of machine learning" (Flach 2012). A feature is "an individual measurable property of the process being observed" (Chandrashekar and Sahin 2014), and a good set of features is essential for any successful machine learning.

In the past years the domain of features has been expanded from tens to thousands of features or variables. There have been several methods to address the issue of removing redundant or irrelevant features from a dataset. Some of the benefits of feature selection include improved prediction performance, better understanding of data and reduction of computation time. One main idea in feature selection is that the independent variables bring no additional information about the classes and turn out to be noise for the predictor. As such, total features in the dataset could be minimized to only few features containing maximum discrimination information about the classes. In other words, the goal in feature selection is to measure the relevance of each feature with the output class/labels.

Feature selection is considered an effective dimensionality reduction technique. The goal of feature selection is to obtain a minimum set of attributes so that the probability distribution of the data classes in the reduced feature space is as close as possible to the original distribution of the data classes in the original feature space (Han et al. 2012).

Dimensionality reduction refers to reducing the number of variables. It can be achieved by elimination, extracting, and engineering of features. By reducing the number of features, feature selection aims at enhancing understandability of data, reducing computational cost, reducing the negative impact of the curse of dimensionality, and finally improving the predictive performance (Chandrashekar and Sahin 2014). As the dimensionality of data increases, data analysis such as classification becomes substantially harder. In some cases, data becomes so sparse leading to curse of dimensionality (Powell 2007). It is possible that in the case of classification the available training data may be very small. Therefore, there will very few data objects in order to create a reliable model which assigns a class to all possible objects. As a result, large number of features may lead to lower classification accuracy. Furthermore, a data with high dimensionality is considered a serious problem in many classification techniques due to the associated memory usage and computational cost (Janecek et al. 2008). Reducing the attribute space also results in better understanding of

the model and makes it easier to utilize different visualization techniques (Tan et al. 2006). Reducing the attribute space is an important factor in both supervised and unsupervised classification and regression problems.

Out of the data reduction strategies in Table 4.1, attribute subset selection is the only one that reduces the number of redundant/irrelevant attributes without changing the structure of features via data compression or attribute construction. Furthermore, what distinguishes feature selection from other dimensionality reduction techniques such as PCA or wavelet transforms, is that in feature selection a subset of the original attribute subset is extracted while in other dimensionality reduction techniques, in general, a linear combinations of the original attribute subset is obtained (Janecek et al. 2008). Moreover, in feature selection no information about the importance of a single feature is lost (Janecek et al. 2008) whereas in other dimensionality reduction techniques, the information about how much the original attributes contribute is usually lost and it is not possible to interpret the linear combinations of the original features.

Feature selection was originally meant to reduce dimensionality in order to control efficiency. In our work, we introduce privacy as another factor when performing feature selection. In other words, we consider a case where dimensionality is reduced while efficiency and privacy are controlled.

There are two main categories of automatic feature selection techniques, namely, *filters* and *wrappers* which are discussed next. Another category is called embedded which performs feature selection as an integral part of a given machine learning technique and is not further discussed in this work.

## 4.1. Relevancy and Redundancy

*Relevancy* is the core concept behind feature selection. A feature is considered relevant if it provides information about the class; either individually or together with other features. A relevant feature cannot be independent of the class labels (though can be independent of other features) (Kohavi and John 1997). Two degrees of relevancy have been identified in (John et al. 1994), namely, strong and weak relevance. In other words, features are considered relevant if they are strongly or weakly relevant with the class and otherwise irrelevant (John et al. 1994). Strong and weak relevancy has been formally defined in

earlier work in (John et al. 1994) and we borrow the formal definition from it. The input is a set of $n$ training instances where each instance $X$ is an element of the set $F_1 \times F_2 \times \ldots \times F_m$, and $F_i$ is the domain of the $i$th feature. Training instances are tuples $<\boldsymbol{T}, C>$ where $C$ is the class. The value of feature $X_i$ is denoted by $x_i$. $S_i$ is considered to be the set of all features except $X_i$. That is, $S_i = \{X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_m\}$and $s_i$ refers to the value of all features in $S_i$.

**Definition 1: (Strong Relevance)** (John et al. 1994)

*$X_i$ is relevant if and only if there exists some $x_i$, $s_i$, and c given $p(X_i = x_i, S_i = s_i) > 0$ so that $p(C = c \mid X_i = x_i, S_i = s_i) \neq p(C = c \mid S_i = s_i)$.*

From Definition 1, $X_i$ is relevant if by eliminating $X_i$, the probability of the class given all features changes. Strongly relevant features provide unique information about the class. In other words, they cannot be replaced by other features.

**Definition 2: (Weak Relevance)** (John et al. 1994)

*A feature $X_i$ is weakly relevant if and only if it is not strongly relevant and there exists a subset of features $S'_i$ of $S_i$ for which there exists some $x_i$, c, and $s'_i$ with $p(X_i = x_i, S'_i = s'_i) > 0$ such that*

*$P(C = c \mid X_i = x_i, S'_i = s'_i) \neq p(C = c \mid S'_i = s'_i)$.*

From Definition 2, weakly relevant features provide information about the class; however, they can be replaced by other features without losing information about the class. In other words, when an attribute has weak relevance with the target class it means it can sometimes contribute to prediction accuracy.

Finally, irrelevant features do not have any information about the class and can be removed. A finer classification of features into weakly relevant but redundant and weakly relevant but non-redundant was proposed by (Yu and Liu 2004).

*Redundancy* is associated with the level of dependency between two or more features. An example of such redundancy measures is Mutual Information (MI) can be used to measure the dependency between a given feature $f_i$ with respect to a feature subset $S \subseteq \neg f_i$. According to (Meyer et al. 2008), such information-theoretic measure of redundancy is nonlinear, symmetric, and nonnegative. Also it does not diminish even if new features are added. One issue however, as (Vergara and Estévez 2014) points out is that with such

measure it is not possible to determine which specific feature in $S$, $f_i$ is redundant with. Two other more elaborated criteria of measuring redundancy include the Markov blanket (Yu and Liu 2004) and total correlation (Watanabe 1960).

## 4.2. Filter Methods

The filter model uses general characteristics of the data in order to evaluate attributes. Filter approach attempts to assess the merits of features from the data without considering the induction algorithm.

The filter approach is shown in Figure 4.1.


Figure 4.1: The filter approach where the features are filtered independently of the induction algorithm (Kohavi and John 1997).

In general, filter methods use variable ranking techniques in order to score the variables. A threshold is also used in order to remove attributes below it. Filter methods are applied before induction (Figure 4.1). The idea is to determine the usefulness of a feature in discriminating different classes. This property is referred to as feature relevance (Kohavi and John 1997). There are several definitions and measurements for the relevance of a variable. In (Law et al. 2004) it is stated that "a feature can be regarded as irrelevant if it is conditionally independent of the class labels". In other words, a relevant feature (though can be independent of other features) cannot be independent of the class labels.

Filter approaches can be divided into two techniques, namely, *univariate* (i.e. individual feature evaluation) and *multivariate* (i.e. subset evaluation). In the *univariate* methods (e.g. InfoGain (Quinlan 1993)) feature dependencies is ignored. However, two main advantages of such methods are that they are fast and scalable. *Multivariate* methods (e.g. CFS (Hall 1999)) consider feature dependencies; however, they are slower and less scalable than the univaritate techniques (Bolón-Canedo et al. 2013).

Individual evaluation is what is referred to as feature ranking which determines a weight for attributes according to their relevance to the target class (Yu and Liu 2004). A

main advantage of feature ranking is that it is fast. A disadvantage of this approach is that the selected subset might not be optimal and the correlation among the attributes within the subset is ignored. It is possible that important features are less informative on their own; however, when they are combined with other features they become important. In feature ranking such features could be discarded (Guyon and Elisseeff 2003). Another disadvantage is that filters do not detect dependencies between features since they depend only on marginal distributions (Flach 2012). An argument is made by (Flach 2012) that filters are good in picking up possible root features for decision tree; however, they are not necessarily good in selecting features further down the tree. Another disadvantage of filters is that since the underlying algorithm is ignored it is more difficult to find suitable learning algorithm (Flach 2012).

Subset evaluation, on the other hand, generates candidate features subsets according to a certain search strategy (Yu and Liu 2004). Each candidate subset is evaluated using a given evaluation measure and then is compared with the previous best subset with respect to the same measure. One main advantage of subset evaluation is that it can handle both feature redundancy and feature relevance. The main issue with subset evaluation is issues associated with searching entire feature space.

## 4.3. Wrapper Methods

Wrapper methods use the induction algorithm's performance as the objective function to evaluate the subset. Wrappers assess subsets of variables according to how useful are they with respect to predicting the target class. In doing so the usefulness of features is detected in the context of other features.

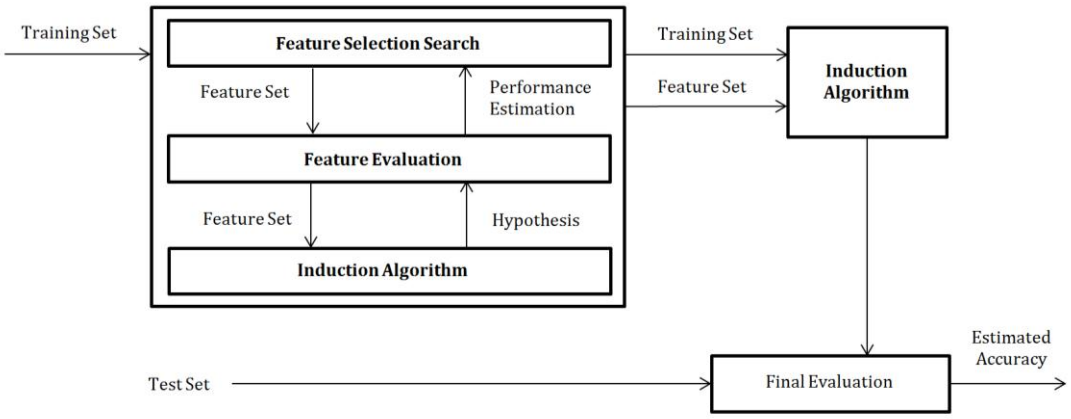



Figure 4.2: The wrapper approach where the induction algorithm is used as "black box" by the subset selection technique (Kohavi and John 1997).

In the wrapper approach, the induction algorithm is used as a black box as shown in Figure 4.2. In this approach, feature selection is 'wrapped' in a search procedure which involves training and then evaluating a model with a candidate set of features (Flach 2012). The idea is described as follows (Kohavi and John 1997): The induction algorithm is applied on the dataset which is partitioned into internal training and hold out sets.

There are two main components associated with feature selection algorithms, namely, feature search and feature subset evaluation (Lin et al. 2011a). Although an exhaustive search leads to the optimal solution, it turns out to be computational impractical since for a set of $N$ features, an exhaustive search would lead to $2^N$ feature combination. It turns out that, the problem of choosing a search method is known to be NP-hard (Amaldi and Kann 1998). Therefore, heuristic methods are commonly used in order to reduce the search space. These methods are usually greedy in which during a search through attribute space, they always what is considered to be the best choice at a time. They aim at making a locally optimal choice hoping that it will lead to a globally optimal solution. (Han et al. 2012).

In general, sequential search techniques use greedy approaches and result in $O(N^2)$ worst case scenario (Lin et al. 2011a). There are different procedures for attribute subset selection including stepwise forward selection, stepwise backward elimination, combination of forward selection and backward elimination (bi-directional), and decision tree induction. These search procedures and their brief descriptions (Han et al. 2012) are listed in Table 4.2.

The stepwise forward selection and backward elimination are commonly used procedures. For example, a well-known search method which searches the space of attribute using greedy search algorithm is BestFirst. Two main advantages of BestFirst include its computational speed and its robustness against overfitting (Guyon and Elisseeff 2003). BestFirst may follow the forward selection or backward elimination procedures in order to obtain the best feature subset. Consequently, such best feature subset is used to build the model. Finally, the resulting classifier is evaluated on an independent test set which was not used during the search.

The main disadvantage of wrapper methods is that it is computationally expensive since for each subset evaluation, the induction algorithm creates a new model. In other words, the classifier is trained for each subset and tested to obtain the classification accuracy. Previous works have shown that wrapper feature selection achieves higher classification accuracy compared with filter feature selection techniques (Hall and Holmes 2003).

Table 4.2: Search procedures and their description.

| Procedure | Description |
|---|---|
| Stepwise forward selection | - Starts with an empty set<br>- the best of the original attributes in added to the list<br>- At each subsequent iteration or step, the best of the remaining original attributes is added to the set. |
| Stepwise backward elimination | - Start with full set of attributes<br>- At each step, removes the worst attribute that remains in the set |
| Combination of forward selection and backward elimination | - Combining stepwise forward selection/backward elimination |
| Decision tree induction | - Constructs a flowchart-like structure<br>- Each internal node is a test on an attribute<br>- Each branch is an outcome of the test<br>- Each external node indicates a class prediction<br>- All attributes that do not appear in the tree are considered irrelvant |

One challenge with regard to overusing the accuracy estimate in feature subset selection is overfitting (Kohavi and John 1997). This issue is described as follows: since there are many feature subsets, it is quite possible that one of them results in a hypothesis which has very good predictive accuracy on the holdout sets. The problem of overfitting in feature subset space has already been addressed in machine learning (Wolpert 1992, Schaffer

1993)and the statistics (Miller 2002) community. However, according to (Kohavi and John 1997), although the theoretical problem remains, overfitting only becomes an issue when the number of instances is small, i.e., when the dataset consists of less than 250 instances. Real world datasets usually contain much more than this amount of instances and therefore, overfitting is not considered a problem in the face of employing wrapper feature selection for the privacy preserving purpose.

In our task-oriented, privacy-preserving data publishing the induction algorithm is pre-determined. Using this very assumption, wrapper feature selection technique was considered a perfect fit for our purpose. There are, however, cases where the final analysis task is known in general (for example, the data recipient knows that the task is classification) however it is not known which particular classification algorithm will be used. In such a scenario filter approaches will be used. We address both approaches in our technique in the context of privacy preserving feature selection.

## 4.4. Summary

In this chapter, we focused on feature selection as a main dimensionality reduction technique. We considered two main approaches to feature selection, i.e., filters and wrappers, and their advantages and disadvantages. We discussed the concept of relevancy and redundancy in the context of feature selection. The chapter is the last chapter of part I. Part I mainly considered background information and related work in three major areas of privacy-preserving data publishing, privacy-preserving data mining, and feature selection.

The information provided in Part I sets up the stage to introduce our proposed techniques and contributions as shown in Part II, next.

# Part II

# Task Oriented Privacy-preserving (TOP) Technologies

# Outline Part II

In this part, we provide an overview of the TOP framework. The framework includes three major dimensions as follows.

**First dimension** is to investigate automatic feature selection as a privacy preserving tool. This dimension explores the role of automatic feature selection in enhancing existing privacy models, namely, *K*-anonymity and differential privacy. The main objective is that utilizing feature selection such that, while satisfying a given privacy model, as much information as possible is preserved for performing the required analysis.

**Second dimension** is to turn two main categories of feature selection, namely filters and wrappers, into privacy-aware processes. As such, we build a layer of privacy on top of automatic feature selection and hence introduce two systems; *PF-IFR* and *PW* corresponding to filters and wrappers respectively. We show that using the correlation among (predictor) attributes on the one hand and correlation between (predictor) attributes and the class attribute on the other hand, we can further refine the attribute selection process. In doing so, we tend to address identity disclosure attack models.

**Third dimension** is to introduce a new privacy measure (*PBI*) for feature selection taking into consideration both efficacy (e.g. accuracy) and privacy simultaneously. we incorporate privacy considerations into the very evaluation measure that is used to evaluate and select feature subsets. We consider privacy during the feature selection process and as such, introduce a two-dimensional measure in automatic feature selection that takes into account both objectives of privacy and efficacy simultaneously and provides the data holder with the flexibility of trading-off one for another. In the same dimension, we introduce a multi-dimensional evaluation function $E(S)$.

Part II consists of a five chapters. In Chapter 5 we discuss the overview of the TOP technologies and show the methodology and experimental results addressing the first dimension discussed above. Chapter 6 and Chapter 7 discuss the privacy-aware filters and wrappers methodology supported by experimental results. These chapters discuss the second dimension. Chapter 8 discusses the concept and methodology of the framework for privacy-aware feature selection evaluation measure and shows the process of generating

candidate privacy-aware subsets that trade-off privacy and efficacy according to the data holder's preferences. It also introduces a multi-dimensional evaluation function for obtaining a single best attribute subset according to the efficacy, privacy, and dimensionality weights of the selected privacy-aware candidate subset.

# Chapter 5

# Enhancing Existing Anonymization Algorithms using Feature Selection

This chapter summarizes our work in (Jafer 2014), (Jafer et al. 2014b), and (Jafer et al. 2014c). In these papers, we show that well-known machine learning techniques such as feature selection can be used to solve privacy-preserving data publishing problems. We present the TOP data publishing model in which feature selection is integrated into the anonymization process. We show that such integration results in an anonymized dataset that, while preserving the privacy, does not have a negative impact on the performance of the resulting models.

This chapter is organized as follows: Section 5.1 provides a general discussion of the Task Oriented Privacy(TOP) concept. Section 5.2 shows how feature selection enhances the existing anonymization algorithms such as *K*-anonymity. In Section 5.3 the impact of feature selection on enhancing differential private data publishing is shown. Finally, Section 5.4 discusses the role of feature selection on the accuracy and performance of the anonymized datasets.

## 5.1. Task Oriented Privacy

One major shortcoming with most of the exiting privacy preserving techniques is that, they do not make any assumption about the ultimate usage of the data. We follow the observation outlined by different studies that the commonly followed strategy of applying general purpose modification of the dataset via 'one-size-fits-all' approach usually leads to over-anonymization of the dataset (Sweeney 2002a). We propose a Task Oriented Privacy (TOP) model and its corresponding software system which incorporates the ultimate usage of the data into the privacy preserving data mining and data publishing process. Our model allows the data recipient to perform privacy preserving data mining including data pre-processing using metadata. It also provides an intelligent privacy preserving data publishing technique guided by feature selection and personalized privacy preferences.

In a typical scenario, the participants of the privacy preserving data mining/publishing process include a data holder and a data recipient. The data holder holds the dataset (which usually consists of identifiers, quasi-identifiers, sensitive information, and non-sensitive information). The data recipient, on the other hand, aims to perform data analysis on the dataset. Before releasing the dataset, the data holder must ensure that the privacy of individuals in the dataset is protected. Two possible solutions exist: The *first solution* includes modifying the dataset and releasing a modified version of it. The *second solution* is to have the data holder to build the model based on the original data and then to release the model to the data recipient. Both solutions, however, have shortcomings associated with them. The problem with the *first solution* is that, data modification usually results in information loss and degradation in the quality of the data for analysis purposes. The problem with the *second solution* is twofold: Firstly, in most cases, the data holder is not a data mining expert. Secondly, since the model is built using the original data, the model itself may violate the privacy.

In privacy preserving data publishing, since the original dataset is modified, eliminating the shortcoming associated with the *first solution* is not possible. It is, however, possible to find methods to protect privacy without negatively impacting the quality of analysis, hence, in this chapter we employ feature selection in order to increase the utility of the resulting dataset without additional privacy side-effects. Different privacy models have been

proposed in the past. These privacy models address number of attack models such as record linkage, attribute linkage, table linkage, and probabilistic attack (a background of privacy models was discussed in Chapter 2).

## 5.2. TOP Data Publishing and K-anonymity

The TOP data publishing incorporates the ultimate usage of the data in order to generate a customized dataset specifically constructed for the intended analysis task. Figure 5.1 illustrates the intersection between all, selected, and QI attributes after applying feature selection to the original dataset. {A} refers to the complete set of attributes after removing the identifiers and the class attribute. From this list, we first identify a subset of attributes selected by the wrapper feature selection algorithm i.e. {B}. We identify {C} which represents the set of QI attributes. We then find {D'} which represents the intersection between {B} and {C}. That is

$$\{D'\} = \{B\} \cap \{C\} \qquad (5.1)$$

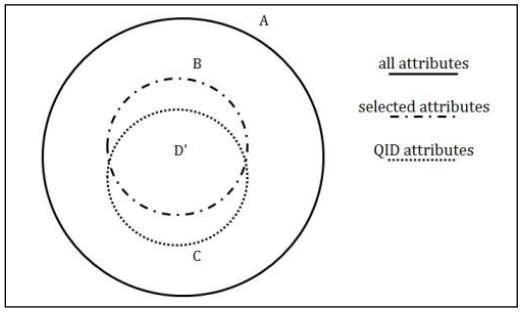



Figure 5.1: The intersection between all, selected, and QI attributes.

Depending on the number of attributes that fall in {D'}, the following outcomes are expected:

- **Case 1**: {D'} = {}. In this case, the features selected by feature selection do not include any the quasi-identifier features.
- **Case 2**: {D'} ≠ {}, B - {D'} ≠ {}, and C - {D'} ≠ {}. That is, {D'} includes some of the quasi-identifier features and excludes some others.

- **Case 3**: {D'} = {C}. In this case, it is possible that all of the quasi-identifier attributes are selected.

Recall from Chapter 2 that a dataset needs to be modified in order to satisfy a given privacy model and therefore, anonymization is applied to the dataset to achieve that. To the best of our knowledge, no previous study has combined feature selection and anonymization to release a task oriented privacy preserving dataset. We investigate such an impact by anonymizying the selected QI attributes, rather than all of the QI attributes and releasing them along other selected attributes according to a specific induction algorithm. Figure 5.2 shows the TOP data publishing model. In this figure, we show how wrapper feature selection and *k*-anonymity (which are seemingly two independent processes) could be combined in order to achieve our objective in turning feature selection into an implicitly privacy preserving process. The first block corresponds to the wrapper feature selection algorithm and is borrowed from (Kohavi and John 1997). It, virtually, represents a black box from which, we first identify a subset of attributes that gives the best performance according to a predetermined induction algorithm *X* (being C4.5, N.B., SVM, etc). The TOP processing block constructs a new dataset with a feature vector equivalent to the selected features. Depending on the ultimate task, the *TOP Processing* block may include selecting a proportion of the dataset. For example, the analysis may include analyzing the records of male patients who are 20-40 years old. In such case only these records are selected.

The resulting dataset is then anonymized using the *Anonymization* block, released, and sent to the data recipient.

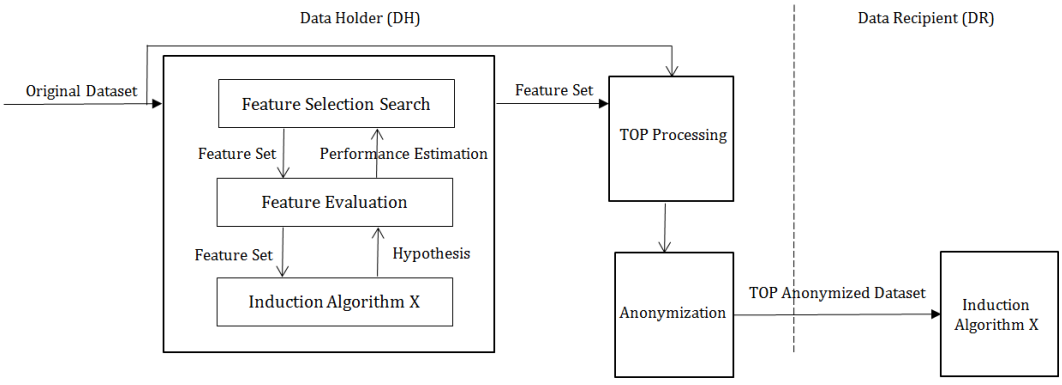



Figure 5.2: Overview of the TOP data publishing model.

We conducted our experiments using some of the UCI datasets identified in Appendix C. Anonymization was done using the Mondrian (LeFevre et al. 2006b) technique. One of the first steps to apply syntactic privacy models such as k-anonymity is to identify the list of quasi-identifiers. In general, demographic attributes are strong candidates to be included in the list. For some datasets, a subset of the QI attributes was selected according to (Keng-Pei and Ming-Syan 2011). This includes {age, sex} in the *heart stat logs* dataset, {preg, mass, age} in the *pima diabetes* dataset, and {purpose, credit amount, personal status, residence since, age, job} in the *german credit* dataset. In the *liver patients* dataset, we identified {age, sex} as quasi-identifiers. In the *diabetes* dataset, we assumed that {age, gender, race, payer code} to constitute our QI attributes. For the remaining datasets, all of the attributes except the target class were selected as the QI set.

We first find the intersection between attribute sets according to the Venn diagrams shown earlier. This includes {B}, {C}, {D'}, and {C} – {D'}.

### 5.2.1. Exclusion of the QI Attributes and Anonymization

We use 10-fold cross validation in order to evaluate our classifiers. As such, in each round of this repeated process, 1/10 portion of the dataset is held as a testing set and the remaining 9/10 portion is used in order to train the classifier. We use two classification algorithms, namely, C4.5 and N.B. (Naïve Bayes) in order to conduct our experiments.

In the case of *heart stat logs* and *liver patients* (which belong to Case 1), when C4.5 is the chosen classifier and in the case of *liver patients,* when N.B. is the selected classifier , no anonymization at all is needed since none of the QI attributes are selected by feature selection. This was the case for the *diabetes* dataset as well. For this dataset, in both cases of C4.5. and N.B. classifiers, none of the QI attributes were picked up by feature selection. This is especially important due to the large size of this dataset. While in the case of C4.5, only {number_inpatient, number_emergency, discharge_disposition_id, rosiglitazone} are selected, in the case of N.B., the set of selected attributes include {discharge_disposition_id, number_emergency, number_inpatient, chlorpropamide, rosiglitazone}. We observe, that with respect to both classifiers the difference in classification accuracy with and without feature selection is not statistically significant. While baseline accuracy (C4.5) is 57.22%,

the accuracy of the dataset projected on the selected attributes is 57.56%. Similarly, while baseline accuracy (N.B.) is 56.28%, the dataset projected on the selected attributes reports a 56.41% accuracy.

For the remaining cases, since feature selection excludes some of the QI attributes, rather than anonymizing all of the attributes in the QI set, only a subset of those attributes needs to be anonymized. This leads to less anonymization and consequently less exposure of potentially harmful attributes. At the same time, less generalization and modification results in preserving more details of the values of relevant attributes w.r.t. to the target class and hence more utility.

Figure 5.3 presents a sample of the *pima diabetes* dataset. The QI attributes associated with this dataset are preg, mass, and age. If the DH publishes this dataset as is, it is possible for an attacker who has access to the external table to link the records in Figure 5.3 (a) and Figure 5.3 (b) and to figure out that a patient named 'Cathy' who is 38 years old, whose BMI is 32, and have been pregnant 5 times in the past, has diabetes. Now, consider the case where Wrapper Feature Selection (WFS) is applied to the dataset using C4.5 as the base classifier. In this case, out of the three QI attributes, only mass and age are retained and preg is excluded (Figure 5.3 (c)).

We may argue that, even without anonymization, such elimination of some of the QI attributes results in more privacy. Assume that the original table has other records x and y with the following QI values: x{preg = 5, mass = 29.8, age = 30, class = 'negative'} and y{preg = 6, mass = 29.8, age = 30, class = 'positive'}. An attacker can link this record with the external table and conclude with 100% accuracy that Samantha has diabetes. However, when preg is excluded, the following QI values are published instead: x{mass = 29.8, age = 30 where class = 'negative'} and y{mass = 29.8, age = 30 where class = 'positive'}. In such case, the attacker can no longer link this record to that of Samantha in the external table confidently. We can argue that, since feature selection excludes some of the QI attributes, such exclusion disrupts the very meaning of the quasi-identifier set. We can make a hypothesis that, since the original QI set is being disrupted, the remaining attributes are no longer considered quasi-identifiers and may be released without any anonymization, although such statement needs to be investigated. To see the impact of feature selection on anonymization consider Figure 5.3 (d): In order to satisfy the 3-anonymity requirement the

dataset needs to undergo coarse generalization. As such, all values of the mass attribute are generalized to [0:68) and all values of the age attribute are generalized to [0:82). By anonymizing the selected QI attribute only (via automatic feature selection) first, fewer attributes need to be anonymized, and second, finer generalization (of the mass attribute) is required. In other words, more details are preserved w.r.t. the target class. This, as the results show, leads to better classification accuracy, thus better utility.

(a) Original pima diabetes records

| preg | plas | pres | skin | insu | mass | pedi | age | class |
|---|---|---|---|---|---|---|---|---|
| 2 | 112 | 66 | 22 | 0 | 25 | 0.307 | 24 | negative |
| 1 | 80 | 55 | 0 | 0 | 19.1 | 0.258 | 21 | negative |
| 4 | 123 | 80 | 15 | 176 | 32 | 0.443 | 34 | negative |
| 7 | 81 | 78 | 40 | 48 | 46.7 | 0.261 | 42 | negative |
| 3 | 83 | 58 | 31 | 18 | 34.3 | 0.336 | 25 | negative |
| 5 | 124 | 74 | 0 | 0 | 34 | 0.22 | 38 | positive |
| 0 | 105 | 84 | 0 | 0 | 27.9 | 0.741 | 62 | positive |

(b) External table

| name | age | mass | preg |
|---|---|---|---|
| Alice | 62 | 27.9 | 0 |
| Elizabeth | 21 | 19.1 | 1 |
| Suzanne | 34 | 32 | 4 |
| Cathy | 38 | 34 | 5 |
| Jennifer | 25 | 34.3 | 3 |
| Helen | 24 | 25 | 2 |
| Nicole | 42 | 46.7 | 7 |
| Marilyn | 28 | 35.6 | 3 |
| Samantha | 30 | 29.8 | 6 |

(c) WFS (C4.5)

| plas | pres | mass | age | class |
|---|---|---|---|---|
| 112 | 66 | 25 | 24 | negative |
| 80 | 55 | 19.1 | 21 | negative |
| 123 | 80 | 32 | 34 | negative |
| 81 | 78 | 46.7 | 42 | negative |
| 83 | 58 | 34.3 | 25 | negative |
| 124 | 74 | 34 | 38 | positive |
| 105 | 84 | 27.9 | 62 | positive |

(d) 3-anonymous (Mondrian) pima diabetes records

| preg | plas | pres | skin | insu | mass | pedi | age | class |
|---|---|---|---|---|---|---|---|---|
| [0.0:3.0] | 112 | 66 | 22 | 0 | [0:68) | 0.307 | [0:82) | negative |
| [0.0:3.0] | 80 | 55 | 0 | 0 | [0:68) | 0.258 | [0:82) | negative |
| (3.0:17.0] | 123 | 80 | 15 | 176 | [0:68) | 0.443 | [0:82) | negative |
| (3.0:17.0] | 81 | 78 | 40 | 48 | [0:68) | 0.261 | [0:82) | negative |
| [0.0:3.0] | 83 | 58 | 31 | 18 | [0:68) | 0.336 | [0:82) | negative |
| (3.0:17.0] | 124 | 74 | 0 | 0 | [0:68) | 0.22 | [0:82) | positive |
| [0.0:3.0] | 105 | 84 | 0 | 0 | [0:68) | 0.741 | [0:82) | positive |

(e) WFS (C4.5) + 3-anonymous (Mondrian)

| plas | pres | mass | age | class |
|---|---|---|---|---|
| 112 | 66 | [0.0:32.0] | [0:82) | negative |
| 80 | 55 | [0.0:32.0] | [0:82) | negative |
| 123 | 80 | [0.0:32.0] | [0:82) | negative |
| 81 | 78 | (32.0:68.0) | [0:82) | negative |
| 83 | 58 | (32.0:68.0) | [0:82) | negative |
| 124 | 74 | (32.0:68.0) | [0:82) | positive |
| 105 | 84 | [0.0:32.0] | [0:82) | positive |

Figure 5.3: Example of pima diabetes records.

Now, let us consider the impact of feature selection on the performance. Results in Table 5.1 show that feature selection does not have any negative impact on the performance. In fact, in the case of N.B., feature selection even improves the classification accuracy and this improvement is statistically significant (based on paired t-test at 0.05).

Table 5.1: Comparison of classification accuracy of models built using all attributes (original) vs. only selected attributes using WFS. ⊕ indicates statistically significant higher performance. Higher classification accuracy is shown in bold.

| Dataset | C4.5 | | | N.B. | | |
|---|---|---|---|---|---|---|
| | **Original** | **WFS** | ***p* value** | **Original** | **WFS** | ***p* value** |
| heart stat logs | 76.67 | **85.19** ⊕ | 0.0004 | 83.7 | **86.3** ⊕ | 0.0248 |
| pima diabetes | 73.83 | **75.78** | 0.0767 | 76.3 | **77.73** ⊕ | 0.0481 |
| German credit | 70.5 | **73.1** | 0.1100 | 75.4 | **76.1** | 0.4716 |
| liver patients | 68.78 | **71.012** | 0.1852 | 55.74 | **71.87** ⊕ | 1.651e-05 |
| CRX | 85.29 | **86.37** | 0.325 | 78.25 | **87.29** ⊕ | 0.0006 |
| CMC | 52.14 | **54.72** | 0.1078 | 50.78 | **55.39** ⊕ | 0.01261 |
| Winconsin Breast Cancer | 93.41 | **95.17** | 0.0051 | 97.36 | **97.80** | 0.1944 |

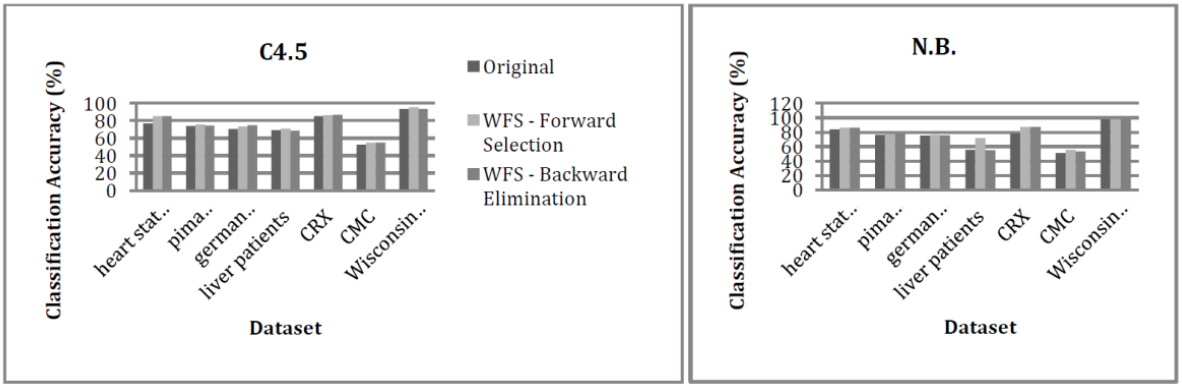



Figure 5.4: Comparison of the performance of original vs. WFS with BestFirst (both forward selection and backward elimination) when the base classifier is C4.5 and N.B.

It was mentioned in Chapter 4 that two commonly used search direction in the BestFirst search strategy are *forward selection* and *backward elimination*. It is beneficial to investigate whether the search direction has any impact on the outcome from the performance and privacy perspectives. We conducted an experiment to compare the impact of search direction on the classification accuracy and the number of retained QI attributes. These results are shown in Figure 5.4 and Table 5.2 respectively. The results show that performance of the models built using forward selection or backward elimination are comparable. Following the definition of QI attributes, we assume that all of these attributes have the same weight with respect to their impact on privacy. If that is the case, since the performance of forward selection and backward elimination is equivalent, out of these two search options we need to select the one that excludes a larger number of QI attributes. As such, our results are in favor of forward selection search direction. We may argue that the strategy of giving equal weights to all of the attributes that constitutes

the QI set may need to be reconsidered. This may lead to further investigation in the future in order to prioritize the attributes within a QI set according to their impact on privacy.

Table 5.2: Number of QI retained, the lower, the better. The lower value is shown in bold font.

| | | Heart stat logs | Pima diabetes | German credit | Liver patients | CRX | CMC | W breast cancer |
|---|---|---|---|---|---|---|---|---|
| C4.5 | Forward Selection | 0 | 2 | 2 | 0 | 1 | 4 | 2 |
| | Backward Elimination | 0 | 1 | 2 | 2 | 8 | 4 | 6 |
| N.B. | Forward Selection | 1 | 2 | 3 | 1 | 6 | 3 | 6 |
| | Backward Elimination | 1 | 2 | 4 | 2 | 7 | 5 | 8 |

### 5.2.2. Experimental Results and Discussions

We next compare the performance of the models built based on the original data vs. the anonymized data. We used paired *t*-test to compare the results obtained for different levels of anonymization. Let us consider the C4.5 results in Table 5.3. In there, Original and Mondrian refer to the classification accuracy of the models built using the original dataset and the dataset anonymized using the Mondrian technique respectively."WFS + Mondrian" refers to the classification accuracy of the models built using the TOP anonymized dataset in Figure 5.2.

The performance of Mondrian in the case of *heart stat logs* ($K$ = 10 and 20), *liver patients* ($K$ = 50), and *Winconsin breast cancer* ($K$ = 10) is higher than Original and this difference is statistically significant. In general, we expect anonymized dataset to result in lower performance due to the fact that generalization leads to hiding some of the details and potential information loss. This is, however, not always the case. In justifying such results, we follow the observation made in (Fung et al. 2005), i.e. when the ultimate analysis task involves classification, anonymization may actually lead to better classification accuracy, or at least may not negatively impact the classification accuracy as we see in our results. Data usually consists of redundant structures for classification. Although generalization eliminates some useful structures, other structures emerge and could be helpful (Fung et al. 2005).

There are, however, other datasets such as *pima diabetes*, *CRX*, and *CMC* which consistently show lower accuracies of Mondrian compared with Original. This is specially the case with *CRX* dataset where the lower performance is statistically significant for all

different anonymization levels. In the case of *Winconsin breast cancer*, anonymization with higher *K* values shows lower performance. *Liver patients* shows lower performance of anonymization at *K* values equal to 10 and 20, however, these results are not statistically significant.

Let us compare the accuracy of the models resulted from our approach "Mondrian + WFS". *Heart stat logs* dataset shows higher classification accuracy compared with Original and Mondrian for all different levels of anonymization and this increase in classification accuracy is statistically significant. Compared with Mondrian results, "Mondrian + WFS" shows statistically significant improvement even at *K* = 30 and 50. The same is applied for the *Wisconsin breast cancer* dataset. "Mondrian + WFS" outperforms Mondrian and Original. Such increase in performance is seen even for the cases where Mondrian resulted in lower classification accuracy compared with Original. Consider the *CRX* dataset in which Mondrian significantly reduces the performance. With "Mondrian + WFS" approach (for all *K* values), compared with Original, better accuracies are obtained. For the *CMC* dataset, although both Mondrian and "Mondrian + WFS" show lower performance than Original, however, the *p* values are in favor or "Mondrian + WFS".

Table 5.3: Comparison of classification accuracy of the original, WFS, Mondrian, and Mondrian + WFS using C4.5 induction algorithm. ⊕ and ⊖ indicates statistically significant results in favor of anonymized and original datasets respectively. Higher accuracy is shown in bold. *K* refers to the anonymization level.

| C4.5 | Original | WFS | Mondrian | | | | Mondrian + WFS | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | | | K =10 | K = 20 | K = 30 | K = 50 | K =10 | K = 20 | K = 30 | K = 50 |
| | Classification Accuracy (%) | | | | | | | | | |
| Heart stat logs | 76.67 | 85.18 | **80.37** ⊕ | **80.37** ⊕ | **80.00** | **79.63** | **85.18** ⊕ | **85.18** ⊕ | **85.18** ⊕ | **85.18** ⊕ |
| Pima diabetes | 73.83 | 75.78 | 72.66 | 73.44 | 71.74 | 72.53 | 69.4 | 73.44 | **74.22** | 73.83 |
| German credit | 70.7 | 73.1 | **70.1** | **71.4** | **70.9** | **73.3** ⊕ | 69.6 | **72.4** | **72.5** | **73.5** |
| Liver patients | 68.78 | 71.012 | 68.74 | 67.18 | **70.12** | **72.19** | **71.012** | **71.012** | **71.012** | **71.012** |
| CRX | 85.29 | 86.37 | 66.15 ⊖ | 61.10 ⊖ | 60.49 ⊖ | 58.81⊖ | **86.37** | **86.37** | **86.37** | **86.37** |
| CMC | 52.14 | 54.72 | 50.24 | 51.18 | 50.98 | 51.32 | **54.18** | 52.07 | 51.32 | 51.73 |
| Wisconsin breast cancer | 93.41 | **95.17** | **95.75** ⊕ | **94.29** | 93.85 | 90.63 ⊖ | **96.05** ⊕ | **94.29** | **95.46** ⊕ | **95.31** ⊕ |

Consider the results shown in Table 5.4 which corresponds to the N.B. classifier. With the exception of two cases, i.e. (*liver patients* when *K*=20) and (*CMC* when *K* = 50), Mondrian results in lower performance compared with Original. In the case of *CRX* this lower performance is statistically significant for all different levels of anonymization. The same is applied for *liver patients* (*K* = 10, 30, and 50), and *Winconsin breast cancer* (*K* = 30 and 50). *German credit* shows statistically significant lower performance for Mondrian for *K* = 20, 30, and 50.

Now, let us consider the results of "Mondrian + WFS" where we do not see such consistent lower performance. *Heart stat logs*, shows higher performance for all *K* values compared with Original. The same is applied to the *pima diabetes* and *German credit* datasets. Although this higher performance is not statistically significant it is consistently higher than the performance of Mondrian. In the case of *German credit*, we observe that while Mondrian performance is statistically lower compared with Original, "Mondrian + WFS" resulted in comparable accuracies. Similarly, in the case of *liver patient,* where Mondrian shows statically significant lower performance compared with Original, "Mondrian + WFS" shows statistically significant higher performance across all *K* values. We also obtain better performance of "Mondrian + WFS" compared with the Original and Mondrian for *CRX* and *CMC*. *Wisconsin breast cancer* shows relatively similar results.

In addition to the privacy gains from using feature selections (Section 5.2.1), the above results show that combining wrapper feature selection and anonymization in the context of TOP data publishing, has a positive impact on the performance of classification. By integrating the intended analysis task into the data publishing process we are able to customize the datasets so that they are best fitted to our analysis purpose. The resulting customized dataset lead to better classification models and our results support this. Many real world datasets consist of large number of records and when the QI set consists of large number of attributes, anonymization becomes a challenging process. By anonymizing the selected attributes in the QI set only, we reduce the amount of generalization required, and this eventually leads to less computational cost and a promising direction towards optimizing the anonymization process in the light of growing high dimensional datasets.

Table 5.4: Comparison of classification accuracy of the original, WFS, Mondrian, and Mondrian + WFS using N.B. induction algorithm. ⊕ and ⊖ indicates statistically significant results in favor of anonymized and original datasets respectively. Higher accuracy is shown in bold.

| N.B. | Original | WFS | Mondrian | | | | Mondrian + WFS | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| Dataset | | | K =10 | K = 20 | K = 30 | K = 50 | K =10 | K = 20 | K = 30 | K = 50 |
| | Classification Accuracy (%) | | | | | | | | | |
| Heart stat logs | 83.7 | 86.29 | 82.96 | 82.96 | 83.33 | 82.96 | **84.81** | **84.81** | **84.81** | **84.81** |
| Pima diabetes | 76.3 | 77.73 | 73.69 | 72.92 | 74.48 | 75.26 | **76.56** | **76.04** | **76.17** | **76.04** |
| German credit | 75.4 | 76.2 | 72.9 | 73.3 ⊖ | 74.20 ⊖ | 74.30 ⊖ | **76.4** | **76.4** | **76.4** | **76.4** |
| Liver patients | 55.74 | 71.87 | 55.61 ⊖ | **55.78** | 56.13 ⊖ | 55.96 ⊖ | **71.87** ⊕ | **71.87** ⊕ | **71.87** ⊕ | **71.87** ⊕ |
| CRX | 78.25 | 87.29 | 64.31 ⊖ | 60.49 ⊖ | 61.41 ⊖ | 58.96 ⊖ | **84.99** ⊕ | **82.54** | 77.03 | 77.03 |
| CMC | 50.78 | 55.39 | 50.37 | 49.08 | 50.03 | **51.73** | **52.74** | **51.19** | 50.03 | **51.05** |
| Winconsin Breast Cancer | 97.36 | 97.80 | 95.75 | 93.7 | 84.63 ⊖ | 81.11 ⊖ | 95.90 | 92.09⊖ | 90.33⊖ | 90.33⊖ |

## 5.3. TOP Data Publishing and Differential Privacy

Section 5.2 studied the impact of incorporating feature selection in the anonymization process for the *K*-anonymity model. In the current section we consider the impact of feature selection on the utility of differentially private datasets.

Recall from Section 2.2.5 that two main approaches for guaranteeing differential privacy can be categorized as *interactive* and *non-interactive* techniques. The current strategy in the *non-interactive* approach is to publish a contingency table (i.e. table of counts) or marginals of the raw data (which are smaller tables of counts with lower dimension) (Dwork 2008). The idea is to first derive a frequency matrix of the original data over the dataset domain. A frequency matrix for a contingency table is computed over all the attributes. After obtaining the counts, noise is added to each count in order to satisfy the privacy requirement.

The main issue with publishing contingency tables is that such an approach is not suitable for high-dimensional data with large domains. This is due to the fact that, in such a setting, the added noise becomes very large compared to the counts and, therefore, the utility of the data is substantially degraded to the level that it makes the data useless. In an alternative technique where low order marginals are sufficient for data analysis, the data holder may release a set of *M* marginals (Dwork 2008). However, the main issue with this approach is that, since noise is added to marginals independently, there will be inconsistency between different marginals. We propose that feature selection address these issues since it only retains the attributes that are relevant according to ultimate usage of the data, thus, reducing the dimensionality of the data. Consequently, it eliminates the need to release multiple marginals.

Let us consider the following sample dataset extracted (with some alternations) from the *pima diabetes* dataset. The dataset in Figure 5.5a consists of ten records. A table of counts constructed from this dataset consists of the same ten records (since they are unique) each having count = 1. Adding noise to such small count value (i.e. 1) substantially impacts the result and the large amount of distortion makes the data useless.

Recall that when the data is anonymized according to the *K*-anonymization model, *K* records are grouped together as they will share the same value with respect to their quasi-

identifier attributes. If all of the attributes are considered quasi-identifiers, such grouping will directly affect the value of counts. Let us examine such a scenario in our example. Consider a 3-anonymous version of Figure 5.5a and its corresponding table of counts (Figure 5.5b and Figure 5.5c respectively).

The table of counts constructed from Figure 5.5b shows a larger size of counts. This results in reducing the effect of noise required to achieve differential privacy (Soria-Comas et al. 2013). Now, consider the case where, feature selection is employed in order to identify the subset of attributes that are most relevant to the task at hand (e.g. building a C4.5 classifier).

| preg | plas | pres | skin | insu | mass | pedi | age | class |
|---|---|---|---|---|---|---|---|---|
| 2 | 112 | 66 | 22 | 0 | 25 | 0.307 | 24 | - |
| 1 | 80 | 55 | 0 | 0 | 19.1 | 0.258 | 21 | + |
| 4 | 123 | 80 | 15 | 176 | 32 | 0.443 | 34 | - |
| 7 | 81 | 78 | 40 | 48 | 46.7 | 0.261 | 42 | + |
| 3 | 83 | 58 | 31 | 18 | 34.3 | 0.336 | 25 | - |
| 5 | 124 | 74 | 0 | 0 | 34 | 0.22 | 38 | + |
| 0 | 105 | 84 | 0 | 0 | 27.9 | 0.741 | 62 | + |
| 4 | 146 | 85 | 27 | 100 | 28.9 | 0.189 | 27 | - |
| 2 | 100 | 66 | 20 | 90 | 32.9 | 0.867 | 28 | + |
| 5 | 139 | 64 | 35 | 140 | 28.6 | 0.411 | 26 | - |

Figure 5.5a: Sample dataset

| preg | plas | pres | skin | insu | mass | pedi | age | class |
|---|---|---|---|---|---|---|---|---|
| [0.0:3.0] | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | - |
| [0.0:3.0] | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | + |
| (3.0:8.0) | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | - |
| (3.0:8.0) | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | + |
| [0.0:3.0] | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | - |
| (3.0:8.0) | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | + |
| [0.0:3.0] | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | + |
| (3.0:8.0) | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | - |
| [0.0:3.0] | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | + |
| (3.0:8.0) | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | - |

Figure 5.5b: 3-anonymous version of Figure 5.5a

| preg | plas | pres | skin | insu | mass | pedi | age | class | Count |
|---|---|---|---|---|---|---|---|---|---|
| [0.0:3.0] | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | - | 2 |
| [0.0:3.0] | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | + | 3 |
| (3.0:8.0) | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | - | 3 |
| (3.0:8.0) | [80:147) | [55:86) | [0:41) | [0:177) | [19.1:46.8) | [0.189:0.868) | [21:63) | + | 2 |

Figure 5.5c: Table of Counts corresponding to Figure 5.5b

| plas | pres | mass | age | class |
|---|---|---|---|---|
| 112 | 66 | 25 | 24 | - |
| 80 | 55 | 19.1 | 21 | + |
| 123 | 80 | 32 | 34 | - |
| 81 | 78 | 46.7 | 42 | + |
| 83 | 58 | 34.3 | 25 | - |
| 124 | 74 | 34 | 38 | + |
| 105 | 84 | 27.9 | 62 | + |
| 146 | 85 | 28.9 | 27 | - |
| 100 | 66 | 32.9 | 28 | + |
| 139 | 64 | 28.6 | 26 | - |

Figure 5.5d: Sample dataset with Feature Selection

| plas | pres | mass | age | class |
|---|---|---|---|---|
| (105.0:147.0) | [55:86) | [0.189:0.868) | [21:63) | - |
| [80.0:105.0] | [55:86) | [0.189:0.868) | [21:63) | + |
| (105.0:147.0) | [55:86) | [0.189:0.868) | [21:63) | - |
| [80.0:105.0] | [55:86) | [0.189:0.868) | [21:63) | + |
| [80.0:105.0] | [55:86) | [0.189:0.868) | [21:63) | - |
| (105.0:147.0) | [55:86) | [0.189:0.868) | [21:63) | + |
| [80.0:105.0] | [55:86) | [0.189:0.868) | [21:63) | + |
| (105.0:147.0) | [55:86) | [0.189:0.868) | [21:63) | - |
| [80.0:105.0] | [55:86) | [0.189:0.868) | [21:63) | + |
| (105.0:147.0) | [55:86) | [0.189:0.868) | [21:63) | - |

Figure 5.5e: 3-anonymous dataset of Figure 5.5d

| plas | pres | mass | age | class | Count |
|---|---|---|---|---|---|
| (105.0:147.0) | [55:86) | [0.189:0.868) | [21:63) | - | 4 |
| [80.0:105.0] | [55:86) | [0.189:0.868) | [21:63) | + | 4 |
| [80.0:105.0] | [55:86) | [0.189:0.868) | [21:63) | - | 1 |
| (105.0:147.0) | [55:86) | [0.189:0.868) | [21:63) | + | 1 |

Figure 5.5f: Table of Counts corresponding to Figure 5.5e

Figure 5.5: A sample raw data example extracted from the pima diabetes dataset and its anonymized versions.

The corresponding table and its 3-anonymous version are shown in Figure 5.5d and Figure 5.5e respectively. The table of counts corresponding to Figure 5.5e shows even higher counts. This has important implication and leads to the conclusion that, when feature selection is incorporated into the anonymization process we obtain higher counts, and when noise is added to these counts (in order to achieve differential privacy), the distortion caused by adding noise is minimized.

Motivated by these observations, we propose a novel technique for privacy preserving data publishing satisfying a differential privacy model. We show that the fact that feature selection enhances both *K*-anonymity and differential privacy enables us to trade-off the level of anonymization and the amount of noise to obtain a dataset that satisfies our

privacy and utility requirements. Our algorithm is capable of handling both numerical and categorical features, and therefore adds the flexibility of being applicable to datasets that essentially include both types of attributes. A good example is the health care domain which consists of datasets that usually contain both numerical and categorical attributes.

### 5.3.1. The *TOP_Diff* Algorithm

The *TOP_Diff* algorithm incorporates the ultimate usage of the data and employs feature selection in order to achieve a differentially private dataset built from *K*-anonymous dataset that maintains high utility for further data analysis according to a given task. The output of this algorithm is an anonymized dataset satisfying differential privacy.

The inputs to this algorithm include a raw dataset *D* with *m* attributes and *n* records, a wrapper feature selection algorithm with a base classifier *Cls* (i.e. *WFS_Cls_Alg*), a given implementation of the *K*-anonymity technique (i.e. *K_Alg*) along desired level of anonymization, *K*, and a privacy budget $\varepsilon$. The choice of *Cls* depends on the analysis task and is dictated by the DR. The algorithm first applies wrapper feature selection algorithm to *D* and obtains the list of selected features $m_s$. It then creates $D_{fs}$ dataset with new feature vector {$m_s$, class} and the same number of records *n*. In other words, the number of records is intact, whereas, $m_s$ corresponds to a projection of *m* and indicates attributes selected by the feature selection algorithm. The algorithm then applies *K_Alg* to $D_{fs}$ and obtains $D_{fsk}$ which is a *K*-anonymous version of $D_{fs}$. In the next step, the algorithm groups similar records together and counts the number of records in each group and generates a table of counts *TC* which essentially consists of *u* unique records and their number of appearances in $D_{fsk}$. It then applies Laplace noise to each true count of *TC*. Since the goal is to reconstruct a new dataset from the noisy table of counts, post-processing is required (*PostProc*). This includes rounding up noisy counts to the nearest non-negative integer value (Mohammed et al. 2011). For example, if after adding noise the count is 23.342 the value will be rounded up to 24. Following the post-processing step, the algorithm obtains a differentially private table of counts *TCDiff*. From this table of counts, it duplicates the number of records based

on their noisy count and generates a differentially private dataset to be released. This algorithm is shown in Figure 5.6.

**Input**: Original dataset ***D*** with $m$ attributes and $n$ records
*WFS_CLs_Alg* (Wrapper Feature Selection and base classifier *Cls*)
*K_Alg* = K-anonymity Algorithm, Anonymization Level (K) , Privacy budget $\varepsilon$

$m_s$ = WFS_CLs_Alg(***D***)
Obtain $D_{fs}$ with {{$m_s$}, class} feature vector and $n$ records
$D_{fsk} = K\_Alg(D_{fs})$
Group similar records in $D_{fsk}$ together and
Obtain *TC* from $D_{fsk}$ with {{$m_s$}, class, Count} feature vector and $u$ records

**for** each $u \in TC$ **do**
    $C' = PostProc(Count + \mathbf{Lap}(2/\varepsilon))$
**end for**

Obtain *TCDiff* with {{$m_s$}, class, C'} feature vector and $u$ records

**for** each $u \in TCDiff$ **do**
    **for** (1 .. C') **do**
        $D_{TOP\_Diff}.add(u)$
    **end for**
**end for**
**Output**: Anonymized dataset $\boldsymbol{D_{TOP\_Diff}}$

Figure 5.6: The *TOP_Diff* Algorithm.

The *TOP_Diff* algorithm generates differentially private counts. However two questions remain. According to (Gehrke et al. 2012), "one of the key insights behind the notion of differential privacy was that privacy should be a property of the sanitization mechanism and not just the output of it". This means every step in the algorithm should also be differentially private. This raises the following two concerns:

First, In the current setting, the generalization procedure is deterministic. Therefore, the algorithm is not differentially private. Second, the discretization of numerical attributes is differentially private?

To address the first concern, one technique is to add a random sampling. The approach of adding random sampling in order to achieve differential privacy was referenced in Section 2.2.6. We add a sampling step similar to (Li et al. 2011). This is, we sample from input dataset with probability β. In other words, each tuple in the input dataset is chosen with probability β. We already use data-independent generalization in our approach. However, unlike (Li et al. 2011), we do not suppress any tuple that appears less than k times. The reason is that, in our approach we do add Laplace noise to the final counts.

Another approach to solve the issue arising from deterministic generalization is to add a a natural sampling step as suggested in (Gehrke et al. 2012).

To address the second concern, since we are using data-independent generalization and global recoding, the discretization of numerical attributes in our algorithm are already differentially private.

#### 5.3.1.1. Experimental Results

The closest work to our technique is the work presented in (Mohammed et al. 2011) which proposes the *DiffGen* algorithm. In order to compare our results with *DiffGen* we used the same evaluation settings used in that work. We build the classifiers using a 10-fold cross validation technique. For consistency in comparing the results, the privacy budget ε values were selected as 0.1, 0.25, 0.5, and 1.

Following the above evaluation process, we applied the *TOP_Diff* algorithm to the *Adult* dataset. We re-run the algorithm for different values of *K* and ε, obtained the anonymized dataset, built a classifier, and recorded the classification accuracy. Before comparing our results with *DiffGen*, we investigated the accuracy of our *TOP_Diff* anonymized dataset (which is differentially private) with that of anonymized datasets using Mondrian *K*-anonymity without feature selection (i.e., KW/OFS) and Mondrian K-anonymity with feature selection (i.e. KWFS). We used a *t*-test statistical significance test to compare our results. The results are shown in Table 5.5. The symbols ⊕ and ⊖ in each case indicate if the results are significantly higher and lower than KW/OFS respectively. For example, for

*K*=10, having ⊕ next to 83.280 KWFS value indicates that the classification accuracy of the model built using 10-anonymized dataset with feature selection is significantly higher than 81.468 (corresponding to KW/OFS or the classification accuracy of the model built using 10-anonymized dataset without feature selection). In another example, consider the value 74.774 (corresponding to KWFS with ε = 0.1, that is, the accuracy of the model built using 10-anonymized *TOP_Diff* dataset with privacy budge equals to 0.1). In this case, having ⊖ next to this value indicates that the classification accuracy compared with 10-anonymized dataset without feature selection is significantly lower than KW/OFS at *K*=10.

Table 5.5: The Classification accuracy of the TOP_Diff algorithm with feature selection.

| | K-anonymity without differential privacy | | K-anonymity combined with differential privacy | | | |
|---|---|---|---|---|---|---|
| | KW/OFS | KWFS | ε = 0.1 | ε = 0.25 | ε = 0.5 | ε = 1 |
| K = 10 | 81.468 | 83.280⊕ | 74.774⊖ | 80.042 | 81.396 | 83.160⊕ |
| K = 20 | 81.086 | 83.234⊕ | 74.703⊖ | 80.156⊕ | 81.636 | 83.208⊕ |
| K = 50 | 80.104 | 83.314⊕ | 78.431⊖ | 82.525⊕ | 83.075⊕ | 83.161⊕ |
| K = 100 | 79.030 | 82.737⊕ | 80.475⊕ | 82.644⊕ | 82.525⊕ | 82.664⊕ |
| K = 200 | 78.670 | 82.080⊕ | 81.808⊕ | 81.881⊕ | 82.080⊕ | 82.080⊕ |

Let us consider the results reported for each ε value. From differential privacy's point of view, a lower value of ε corresponds to a larger noise and hence, implies higher privacy protection. For *K*=10, 20, and 50, the effect of noise at ε = 0.1 is high and we get a lower performance if compared with KW/OFS. However, as the value of *K* increases, due to a larger count values the effect of noise is reduced. As such, the performance becomes significantly higher when *K* = 100 and 200. For the remaining values of ε, for different anonymization levels, in general, we obtained significantly higher accuracies if compared with KW/OFS. It follows that, for the given privacy budgets (i.e. 0.25, 0.5, and 1) the *TOP_Diff* algorithm generates datasets that satisfy the differential privacy with a utility higher than corresponding *K*-anonymous dataset at the same anonymization level.

An argument can be made that, what happens if the feature selection step is omitted from the *TOP_Diff* algorithm? Using the very notion of indistinguishablity and the fact that higher *K* values essentially lead to higher counts, we investigate whether the role of feature selection in the process is significant. In this experiment, we re-run the *TOP_Diff* algorithm without the feature selection step. In other words, we generate the *TC* based on $D_k$ and not

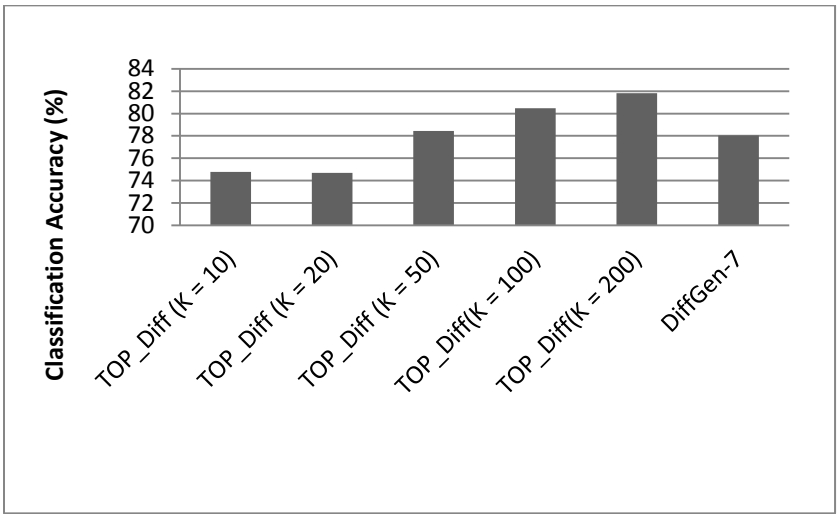
$D_{fsk}$. We select the same anonymization levels and ε values and compare the results. These results are shown in Table 5.6.

The results show that when the feature selection step is omitted, higher count values obtained due to increasing the value of *K* alone (without feature selection) do not improve the utility of differentially private dataset. This is shown by ⊖ next to most of the results (for different values of *K* and ε). In other words, achieving differential privacy comes at the cost of statistically significant lower classification accuracy compared with normal *K*-anonymization. This shows the essential role of feature selection in the process (see Table 5.5). As we reduce ε, we inject more noise into the dataset and therefore, we expect a reduction in its utility, i.e. classification accuracy. However, we can compensate for this by increasing the value of *K*. Increasing *K* means increasing the number of records that appear in each equivalence group and hence achieving higher counts. Higher counts at higher *K* values eliminate the negative impacts of the added noise. On the other hand, as we increase ε, we inject less noise in the dataset and we obtain results closer to the *K*-anonymized dataset without noise (i.e. our final anonymized dataset is similar to KWFS). Remember that increasing the value of *K* usually leads to a decrease in the classification accuracy. When we increase ε, higher values of *K* mean a coarser grained generalization which usually results in reduction in the classification accuracy. At the same time, lower *K* reduce generalization and consequently lead to a higher classification accuracy. To this end, we are able to choose the desired trade-off between *K* and ε to satisfy our privacy and utility needs.

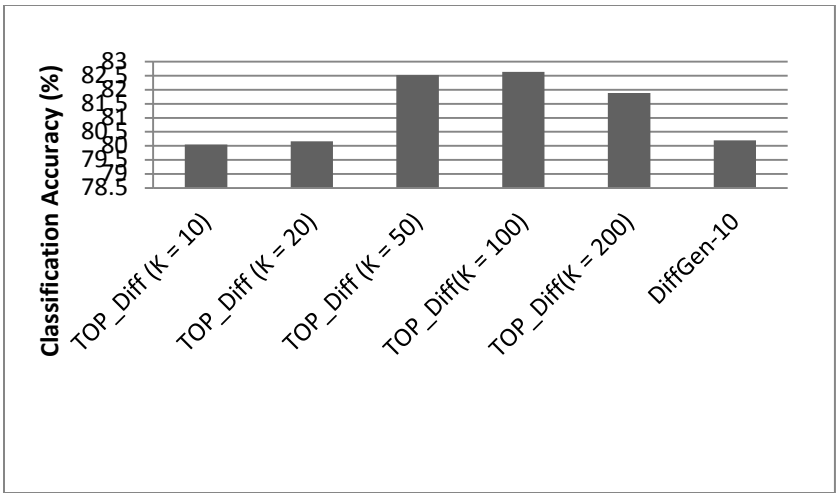
Table 5.6: The classification accuracy of the TOP_Diff algorithm without feature selection.

| | K-anonymity without differential privacy | K-anonymity combined with differential privacy | | | |
|---|---|---|---|---|---|
| | KW/OFS | ε = 0.1 | ε = 0.25 | ε = 0.5 | ε = 1 |
| K = 10 | 81.468 | 71.012⊖ | 73.925⊖ | 77.667⊖ | 78.425⊖ |
| K = 20 | 81.086 | 69.802⊖ | 76.840⊖ | 79.375⊖ | 80.675 |
| K = 50 | 80.104 | 73.490⊖ | 78.791⊖ | 79.534 | 79.859 |
| K = 100 | 79.030 | 76.409⊖ | 78.910 | 79.109 | 78.970⊖ |
| K = 200 | 78.670 | 76.460⊖ | 78.551 | 78.544⊖ | 78.305⊖ |

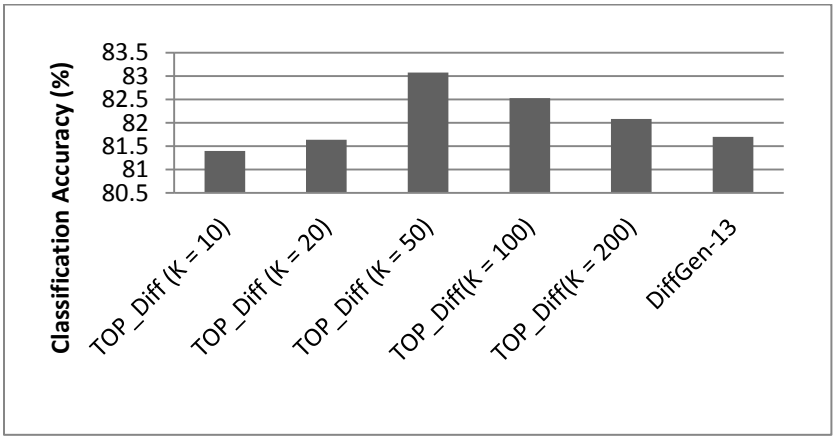
Next, we compare our results with the results of the *DiffGen* algorithm(Mohammed et al. 2011). The *DiffGen* algorithm initially generalizes all of the records into one group and iteratively applies a sequence of specializations. At each iteration, the algorithm probabilistically selects an attribute to specialize based on some score value (e.g. Max, and Information Gain). The algorithm terminates after a preset number of specializations. The work in (Mohammed et al. 2011) shows that, compared with Information Gain, Max score results in much higher accuracy across different number of specializations and for different values of ε. Therefore, we compare our results with those reported for the utility function Max. For each choice of ε (i.e. 0.1, 0.25, 0.5, and 1) we select the highest accuracy reported by (Mohammed et al. 2011). These values are marked as *DiffGen-x* where *x* refers to the number of specializations used to achieve that result. We use the same $K$ and $\varepsilon$ when making the comparison. The comparison results are shown in Figure 5.7 a, b, c, and d. The results, in general, show that when ε = 0.1, 0.25, and 0.5, higher values of $K$ lead to classification accuracy higher than obtained by *DiffGen*. As for ε = 1, we obtain higher classification accuracy with lower values of $K$.



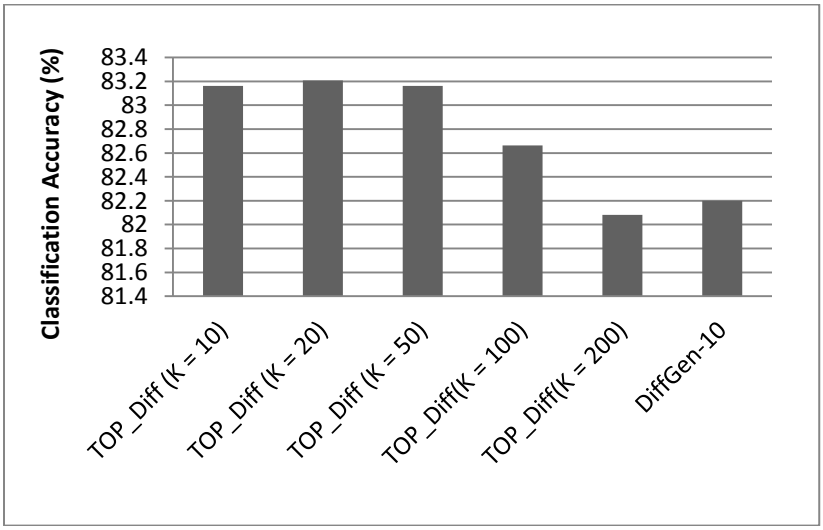

a (ε = 0.1).




b (ε = 0.25).




c (ε = 0.5).




d (ε = 1).

Figure 5.7: Comparison of the TOP_Diff and DiffGen algorithms at different values of ε.

In summary, recall that lower values of ε indicate a higher privacy protection due to the increased amount of noise. As we reduce ε, we inject more noise into the dataset and therefore, we expect a reduction in its utility, i.e. classification accuracy. However, we can compensate for this by increasing the value of *K*. Increasing *K* means increasing the number of records that appear in each equivalence group and hence achieving higher counts. Higher counts at higher *K* values eliminate the negative impacts of the added noise. On the other hand, as we increase ε, we inject less noise in the dataset and we obtain results closer to the *K*-anonymized dataset without noise. Remember that increasing the value of *K* usually leads to a decrease in the classification accuracy. When we increase ε, higher values of *K* mean a coarser grained generalization which usually results in reduction in the classification accuracy. On the other hand, lower *K* leads to less generalization and consequently to a higher classification accuracy. To this end, we are able to choose the desired trade-off between *K* and ε to satisfy our privacy and utility needs.

## 5.4. The Impact of Feature Selection on the Performance and Efficiency

In this section, we investigate the impact of feature selection on the time required to perform anonymization and the accuracy of the resulting classification models.

The implication of incorporating feature selection in the syntactic anonymization techniques (e.g. *K*-anonymity) on the privacy and the performance was studied earlier (Section 5.2). In that work, we showed that well-known machine learning techniques such as feature selection can be used to solve privacy preserving data publishing problem. Another important implication is the fact that feature selection results in selecting the most relevant attributes for the analysis task at hand (i.e. classification). Recall from Section 5.2.1 that, it is very likely that some of the eliminated attributes are quasi-identifiers which otherwise needed to be anonymized. One immediate conclusion is that, with less attributes to anonymize less time is required to perform anonymization. With large amount of data and datasets with high dimensionality such role of feature selection is extremely beneficial.

To study the impact of feature selection on performance and efficiency, we impose a very strict condition by considering all of the attributes to be quasi-identifiers. In doing so, we follow the very assumption made in differential privacy, i.e., the attacker could have

arbitrary background knowledge. We first compare the impact of feature selection on the time required to perform anonymization. We apply different levels of anonymization (i.e. considering different K values in the K-anonymity model) with and without feature selection and compare the resulting performance. We evaluate our results using t-test statistical significant test.

### 5.4.1. Experimental Results and Discussions

We used the *Adult* dataset which has become the benchmarking data set for academic research in the area of privacy. This dataset (after removing the records with missing values) contains 45,222 records with 6 numerical attributes, 8 categorical attributes, and a binary class attribute which represents two income levels, i.e., <=50 K and >50 K.

In Section 5.2, we showed that when feature selection is incorporated into the anonymization process, the classification accuracy of models built from anonymized data was increased. That work, however, did not address the impact of feature selection on the efficiency of anonymization. The fact that feature selection leads to smaller set of attributes especially when all of the attributes are considered quasi-identifiers has important implication from the efficiency point of view. In fact, such a strict assumption about the quasi-identifiers becomes practical only if feature selection is in place and eventually results in selecting a subset of attributes which are most relevant to our task. Without feature selection, considering all attributes as quasi-identifiers, especially when the data is high dimensional is neither practical nor realistic due to the curse of dimensionality (Aggarwal 2005).

We recorded the time required to anonymize the dataset along the classification accuracy of the anonymized dataset at each anonymization level and used a statistical *t*-test to compare the results. We first applied the Mondrian anonymization technique to the *Adult* dataset without applying feature selection and recorded both the time required for anonymization and the classification accuracy of the resulting anonymized dataset. We evaluated the classification accuracy using 10-fold cross validation. We then applied wrapper feature selection with C4.5 as the base classifier and obtained the set of most relevant attributes for building a C4.5 classifier. After identifying the most relevant attributes, we constructed a new dataset with the same records but with a smaller feature

vector projecting the list of selected features. We then applied the Mondrian algorithm to this new dataset using the same anonymization levels as before.

Figure 5.8 shows that when feature selection is applied to the dataset, the time required to anonymize the dataset has decreased substantially. The proportion of such decrease ranges from being 3.8 times faster (K=10) to being 2.5 times faster (K=200).

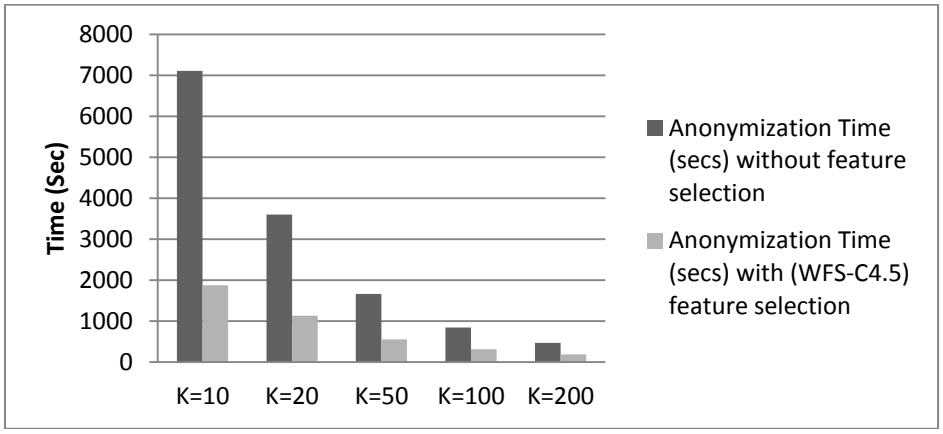



Figure 5.8: Comparison of the anonymization time required with and without FS.

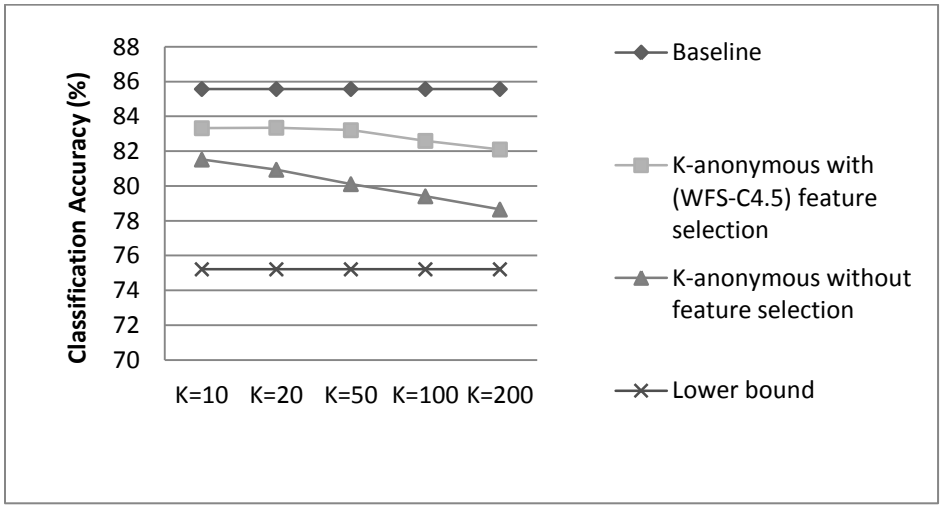



Figure 5.9: Comparison of classification accuracy with feature selection, without feature selection, baseline, and lower bound.

Figure 5.9 compares the classification accuracy of the models built using the anonymized dataset with and without feature selection. *Baseline* accuracy refers to the accuracy of the C4.5 model built using the original dataset intact, and *Lower bound* accuracy refers to the accuracy of the C4.5 model when all attributes except the target attribute are removed (Mohammed et al. 2011). By incorporating feature selection into the

anonymization process the performance has increased in statistically significant way across different K values.

## 5.5. Summary

In this chapter we presented our results in using feature selection as an add-on in order to enhance existing privacy preserving techniques such as k-anonymity (Section 5.2) and differential privacy (Section 5.3). We showed that when feature selection precedes applying anonymization operations, we obtain datasets that maintain high utility for analysis purpose while achieve privacy protection. We also presented the impact of feature selection on enhancing performance and time required to apply anonymization.

# Chapter 6

# Privacy-aware Filters

This chapter summarizes our work in (Jafer et al. 2014a). It was mentioned earlier in Chapter 4 that one of the main advantages of filters was their speed. This is mainly due to the fact that in filters there is no interaction with the learning algorithm and only the characteristics of data itself taken into account when selecting relevant features. Filters become indispensible specially when dealing with large amount of data. We aim at incorporating privacy into feature selection inspired by the privacy-by-design standard (Cavoukian 2009). Recall from Section 1.2 that we consider two principles of this standard in our work, namely, "*privacy as the default setting*" and "*privacy embedded into design*".

As such, we turn filter-based feature selection into a privacy-aware process. Without privacy considerations, filter-based feature selection would rank individual features according to their relevance to the class and then depending on the selected threshold low-ranked attributes get eliminated. Some of these eliminated attributes happen to be part of

the quasi-identifier set. When filters are privacy-aware, correlation between features is exploited in order to remove some of the QI attributes that are above the threshold and would have otherwise been included in the final selected set. We propose a privacy-aware filter-based feature selection method (*PF-IFR*). This method enables data custodians to define a trade-off measure for controlling the amount of privacy and efficacy using filter-based feature selection techniques.

This chapter is organized as follows: Section 6.1 shows the methodology behind our privacy-aware filter-based feature selection. The experimental results are shown in Section 6.2. Finally, Section 6.3 summarizes this chapters and discusses some potential future directions.

## 6.1. Privacy-aware Filter-based System

Recall from Chapter 4 that, feature selection is based on the notion that redundant and/or irrelevant variables bring no additional information about the data classes and can be considered noise for the predictor. As a result, the total feature set of a dataset could be minimized to only few features containing maximum discrimination information about the classes. In other words, feature selection aims to measure the relevance of each feature with the output class/labels and eliminate those with a minimal input.

The *filter* model uses general characteristics of the data in order to evaluate attributes. This approach attempts to assess the merits of features from the data before rule induction is applied by the algorithm. The methods use ranking techniques in order to score the attribute; attributes that do not satisfy a threshold are removed. The idea is to determine the usefulness of a feature in discrimination of different classes. This feature property is referred to as feature relevance (Kohavi and John 1997). In other words, a relevant feature cannot be independent of the class labels (though can be independent of other features).

Filters are categorized based on their evaluation procedure. This includes individual feature evaluation and subset evaluation (Yu and Liu 2004). Individual evaluation determines a weight for an attribute according to its relevance to the target class. A main advantage of feature ranking is its speed. Disadvantages of this approach are that (i) the selected subset might not be optimal and (ii) the correlation among the attributes within the subset is ignored (Guyon and Elisseeff 2003).

Subset evaluation, on the other hand, generates candidate features subsets according to a certain search strategy (Yu and Liu 2004). Each candidate subset is evaluated using a given evaluation measure and then is compared with the previous best subset with respect to the same measure. One main advantage of subset evaluation is that it can handle both feature redundancy and feature relevance. The main issue with subset evaluation is the inevitable problem of searching through feature subsets which is required during the subset generation step (Bolón-Canedo et al. 2013).

Since computational efficiency is a premium when we deal with large and high-dimensional datasets (where scalability becomes a very important factor) we use filter feature selection with individual feature ranking approach in the current work. Furthermore, when the ultimate analysis goal is pre-determined (e.g. building a classifier) but the user has not yet decided which classification algorithm to use or wants to try different algorithms (e.g. NetFlix Prize) privacy-aware filters become very useful since feature selection is done in a privacy-aware environment while being independent of any given classifier.

Our privacy preserving feature selection system consists of two main blocks, namely, the *Evaluation* block and the *Correlation* block. Having separate blocks for evaluation and correlation enables the user to choose different rankers and correlation measures independently from a specific filter-based feature selection technique.

Our system is illustrated in Figure 6.1. The *Evaluation* block realizes the Evaluation Function *EF*(*D, T, e, FS*) which takes four input parameters: the dataset *D*, the ranker threshold *T* (defined by the user), the evaluation criteria *e*, and the type of feature selection *FS*. *D* consists of *m* tuples and *n* attributes {$a_1$, $a_2$,..., $a_{n-1}$, *C*} where *C* refers to the class attribute. Examples of evaluation criteria are Information Gain, Chi-square, Relief, etc. *FS* refers to the type of feature selection (e.g. filters with individual ranking, filters with subset ranking, wrapper), and *T* refers to the ranking threshold which is chosen by the user.

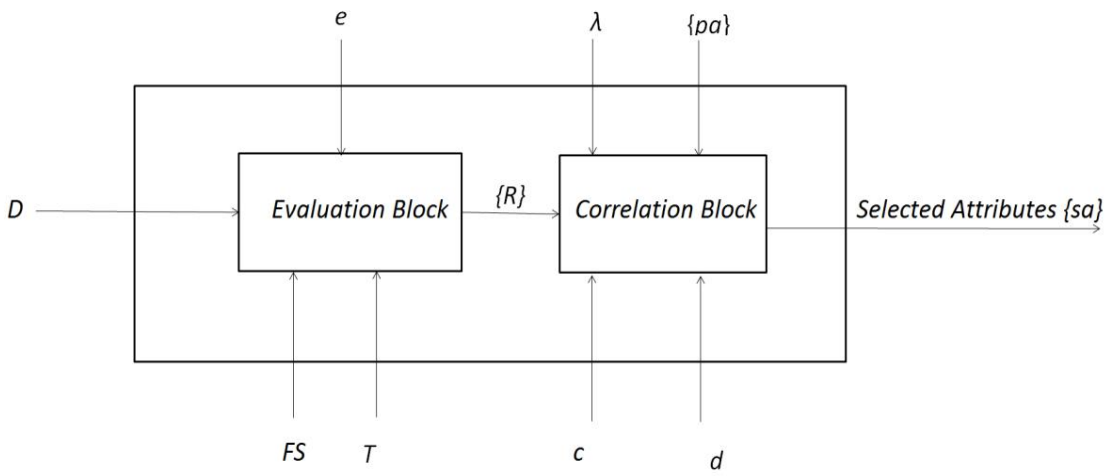



Figure 6.1: The Privacy-aware Filter-based Feature Selection System.

The *Correlation* block realizes the Correlation Function *CF*({R},{*pa*}, *c, d*, *λ*) which takes five parameters: {*R*} refers to the ranked list of attributes obtained from the *Evaluation Block*. {*pa*} refers to the set of potentially privacy-breaching attributes identified by the user, i.e., QI. *c* is the correlation criteria such as Mutual Information (*MI*), Symmetric Uncertainty (*SU*), etc and is considered the choice of the user. *d* is the discretization factor and is required when the numerical attributes need to be discretized as a requirement before applying correlation; this variable needs to be set to either true of false, and if set to true the type of discretization technique is specified. Finally, *λ* is the blending (or the trade-off) parameter which controls the level of privacy vs. the level of accuracy. When *λ* is increased, in order to remove QI attributes, higher correlation between QI and non-QI attributes is required and vice versa. *λ* ranges from 0 to 1 and is chosen by the user. (*λ* = 1) refers to the case where none of the QI attributes are removed except for the case where there is at least one perfect correlation between QI and non-QI attribute. (*λ* = 0) refers to the case where all of the QI attributes are removed.

In this work, the system's output is obtained through Privacy-aware Filter with Individual Feature Ranking (*PF-IFR*) algorithm. The algorithm functions as follows: First (in the *Evaluation Block*) the attributes are ranked according to a given evaluation criterion, *e.* Then (in the *Correlation Block*), starting with the bottom of the ranked list and going upward, the correlation between the encountered QI attribute $q_i$ and other attributes (which essentially have higher ranking) is calculated using the chosen correlation criteria, *c*.

Following this step, if there is an attribute that is correlated with $q_i$ and this correlation is higher than the trade-off parameter $\lambda$, $q_i$ is removed since the other attribute has higher relevance with respect to the target class anyways; otherwise it is included in the list of selected attributes. This process is repeated iteratively. This algorithm is presented in Figure 6.2.

**Input**: *D*, *e*, *T*, *FS*, $\lambda$ , *c*, *d*, {*pa*}

{*R*} = *EF*(*D, T, e, FS*)
/* Returns a ranked list of attributes {*R*} = {$R_1$, $R_2$, …, $R_n$} where $R_1$ refers to the attribute with highest rank and $R_n$ refers to the attribute with the lowest rank. The class attribute is excluded */

*CB*($\lambda$ , *c*, *d*, *D*,{*pa*}, {*R*})
/*Implementation of the Correlation Block*/

**if** (*d*) *discretize*(*D, d*)
/* If discretization is required, discretize using the discretization technique specified in *d* */

{*sa*} = {*R*}
/* First initialize {*sa*} to {*R*}
**while** (){
  **if** ($R_n \subset$ {*pa*}) {
    *max_corr*= MAX(*find_correlation*($R_n$, {$R_1$, .., $R_{n-1}$}, *c*))
    /* obtain the maximum correlation between $R_n$ and other attributes that have higher rank. Return the value of max correlation and the other attribute $R_x$ */

    **if** (*max_corr* >= $\lambda$) {
      {*sa*} = {*sa*} – {$R_n$}
      /* if the correlation is higher than $\lambda$ *r*emove this {*pa*} attribute *
    }
  }
}

**Output**: list of selected attributes {*sa*}

Figure 6.2: The PF-IFR algorithm.

## 6.2. Experimental Results and Discussions

In the *Evaluation Block*, we selected *e* (i.e. the evaluation criteria) to be InfoGain (IG) (Hall and Smith 1998) and ReliefF (Kononenko 1994). *IG* is one of the most commonly used attribute evaluation method. This filter provides a ranked list of all attributes and a threshold is used to determine which attributes are to be included. ReliefF is an extension of the original Relief algorithm (Kira and Rendell 1992). Relief randomly samples an instance from the data. It then locates its nearest neighbor from the same and opposite class. Following this step, the values of the attributes of the nearest neighbors are compared with sample instance and then the relevance score for each attribute is updated.

The attributes are ranked according to their *IG* and ReliefF score and the ranked list {*R*} is obtained; the higher value, the higher rank and the more relevant the attribute is with respect to the class.

As for the *Correlation Block*, we select *c* to be the Symmetric Uncertainty (*SU*) metric. *SU* is a modified information gain measure in order to estimate the degree of association between discrete features (Press 1988). It measures the correlation between two features and more specifically the amount of information a given attribute provides about another attribute.

$$SU = 2.0 \times \left[\frac{H(X)+H(Y)-H(X,Y)}{H(X)+H(Y)}\right] \qquad (6.1)$$

In formula (6.1), *H(X)* and *H(Y)* refer to the entropy of attributes *X* and *Y* respectively, and *H(X,Y)* refers to the joint entropy of *X* and *Y*.

Since *SU* is symmetric it can be used to measure feature-feature correlations in which there is no notion of one attribute being the "class" attribute (Hall 1999). It is required for the *SU* measure that the attributes be categorical. Therefore, we set *d* to true and first discretize the numeric features using the technique of Fayyad and Irani (Fayyad and Irani 1993).

{*pa*} in our example refers to {QI}. From the privacy perspective, we treat all the QI attributes similarly, i.e., we assume that all attributes in the QI set have the same privacy risk and therefore it is not important which particular QI attribute is removed.

We consider different values for $\lambda$, although the choice of $\lambda$ is made by the user and depends on the correlation among the attributes in a given dataset. In each round, we change $\lambda$ and obtain a set of selected attribute and then use those attributes only to build the classification model.

We worked with three real life datasets from the UCI repository (http://archive.ics.uci.edu/ml/). The QI attributes in this dataset include {*marital_status*, *age*, *education*, *occupation*, *sex*, *workclass*, *native_country*, and *race*}. We selected {preg, mass, age} in the *pima diabetes* dataset, and {purpose, credit_amount, personal_status, residence_since, age, job} in the *german credit* dataset as the QI attributes (Keng-Pei and Ming-Syan 2011).

To run our experiments, we selected three classifiers, namely, C4.5, N.B., and KNN. C4.5 is considered a logical model whereas N.B. is considered a probabilistic model. KNN, on the other hand, is a geometric model (Flach 2012). Recall that our algorithm is filter-based, and filters consider general characteristics of the data in order to evaluate attributes, and attempts to assess the merits of features from the data independent of the induction algorithm.

The results (Figure 6.3) show a consistency between using InfoGain vs. ReliefF as feature ranking measures. This provides the user with greater flexibility of choosing different individual feature ranking filter-based methods.

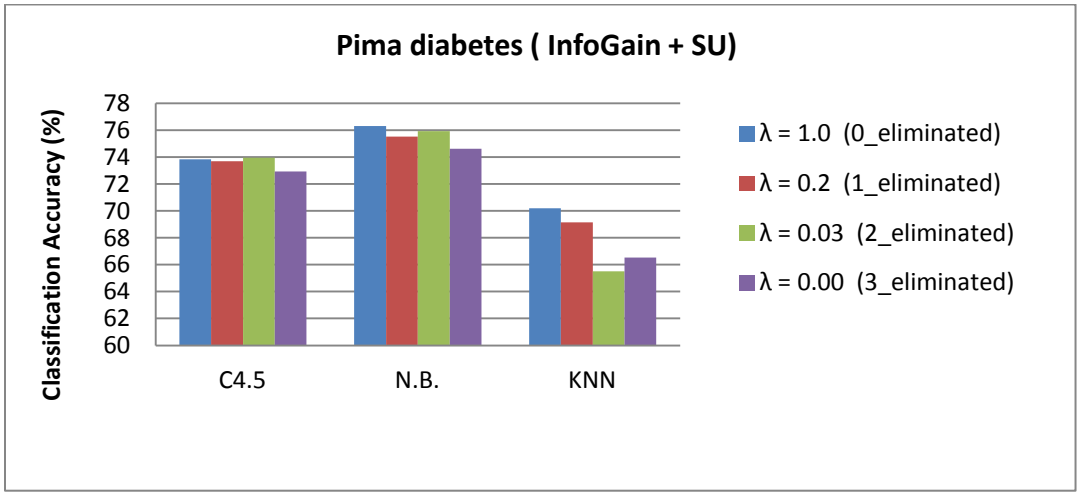
Following the *PF-IFR* algorithm, we identify the QI attributes that have *max_corr* with other attributes (with higher rank) greater than $\lambda$. For each dataset, we decrease $\lambda$ and eliminate QI attributes, build a classifier, and record the classification accuracy.

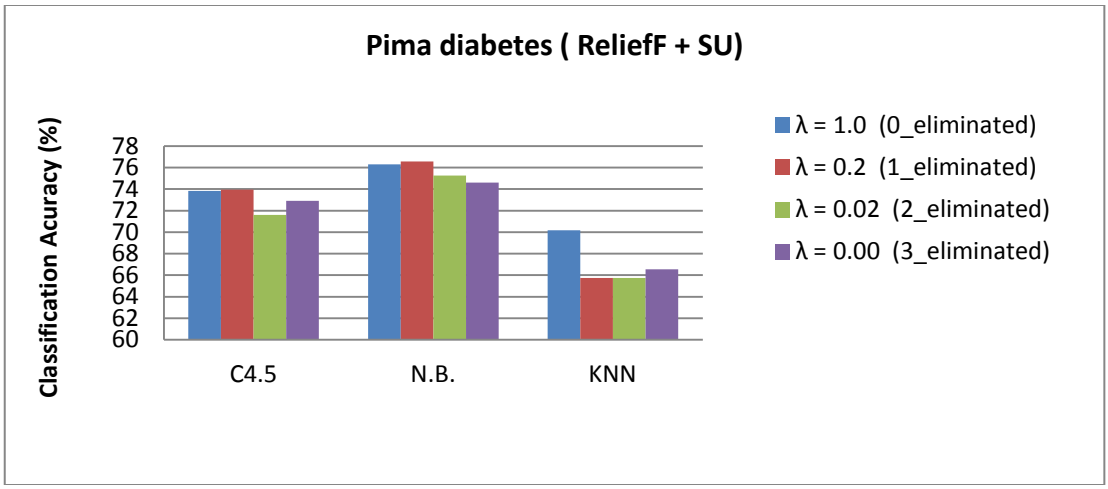
We ran a statistical significance test (*t*-test) in order to investigate if the change in the reported accuracy due to elimination of QI attributes is significant. In each case, the classification accuracy of the original model (i.e. $\lambda$ = 1.0 and 0 eliminated QI attribute) is compared with models corresponding to different values of $\lambda$. For the *Adult* dataset, in the case of C4.5 and KNN classifiers, changing $\lambda$ does not impact the accuracy significantly. The case for N.B. classifier is different; as any elimination of QI attributes due to decreasing of $\lambda$ reduces the accuracy significantly. For the *pima diabetes* dataset, while C4.5 and N.B. do not show any significant change of classification accuracy due to eliminated of the QI attributes,

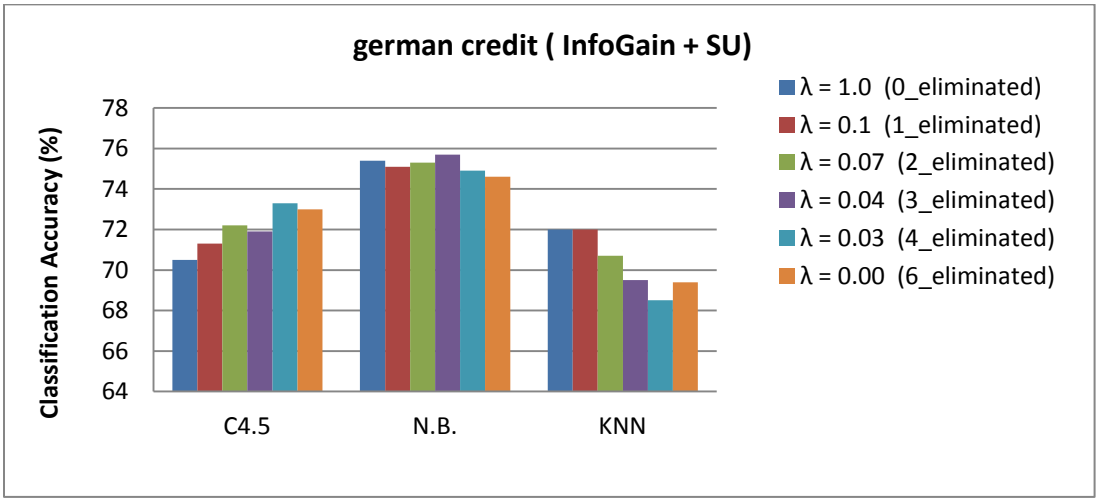
KNN classifier (ReliefF + SU case) shows significant reduction in accuracy at $\lambda$ = 0.2 and 0.02.

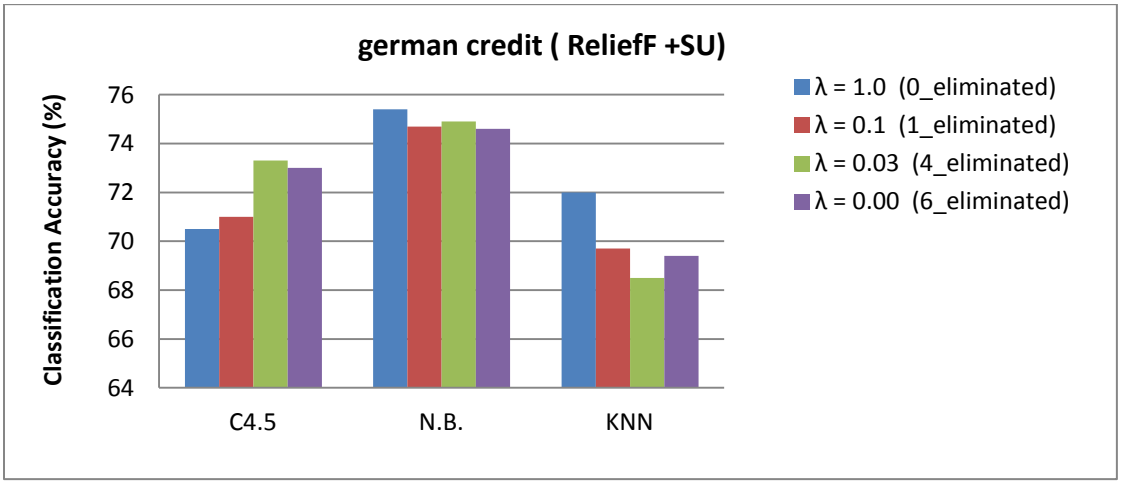
The *german credit* dataset does not show any significant change in the classification accuracy when $\lambda$ is greater than 0.03. This is the case for all of the three examined classification algorithms. In the case of C4.5 and KNN, reducing $\lambda$ below 0.03 results in significant increase and decrease in the accuracy respectively. In this dataset, N.B. does not show any significant change in the classification accuracy.

The goal of this work is to enforce privacy protection during the feature selection process. The degree of correlation between QI and non-QI attributes plays an important role. When such correlation is high, there is more confidence that QI attributes could be removed without deteriorating the performance. One existing challenge is making the distinction between QI attributes and non-QI attributes. This leads us to the very attempt of choosing QI attributes which is an open area of research and beyond the scope of this thesis. In (Motwani and Xu 2008), a method was proposed to determine the minimal set of quasi-identifiers in a data table. The idea is to identify a minimal set of attributes from the dataset which is capable of almost distinctly identify a record and capable of separating two data records. However, according to (Fung et al. 2010b) identifying the minimal set of QI does not result in the most appropriate privacy protection setting since the method does not consider the attributes that an attacker could potentially get access to.




Pima diabetes ( ReliefF + SU)
Classification Acuracy (%)
78
76
74
72
70
68
66
64
62
60
C4.5
N.B.
KNN
λ = 1.0 (0_eliminated)
λ = 0.2 (1_eliminated)
λ = 0.02 (2_eliminated)
λ = 0.00 (3_eliminated)


german credit ( InfoGain + SU)
Classification Accuracy (%)
78
76
74
72
70
68
66
64
C4.5
N.B.
KNN
λ = 1.0 (0_eliminated)
λ = 0.1 (1_eliminated)
λ = 0.07 (2_eliminated)
λ = 0.04 (3_eliminated)
λ = 0.03 (4_eliminated)
λ = 0.00 (6_eliminated)


german credit ( ReliefF +SU)
Classification Accuracy (%)
76
74
72
70
68
66
64
C4.5
N.B.
KNN
λ = 1.0 (0_eliminated)
λ = 0.1 (1_eliminated)
λ = 0.03 (4_eliminated)
λ = 0.00 (6_eliminated)

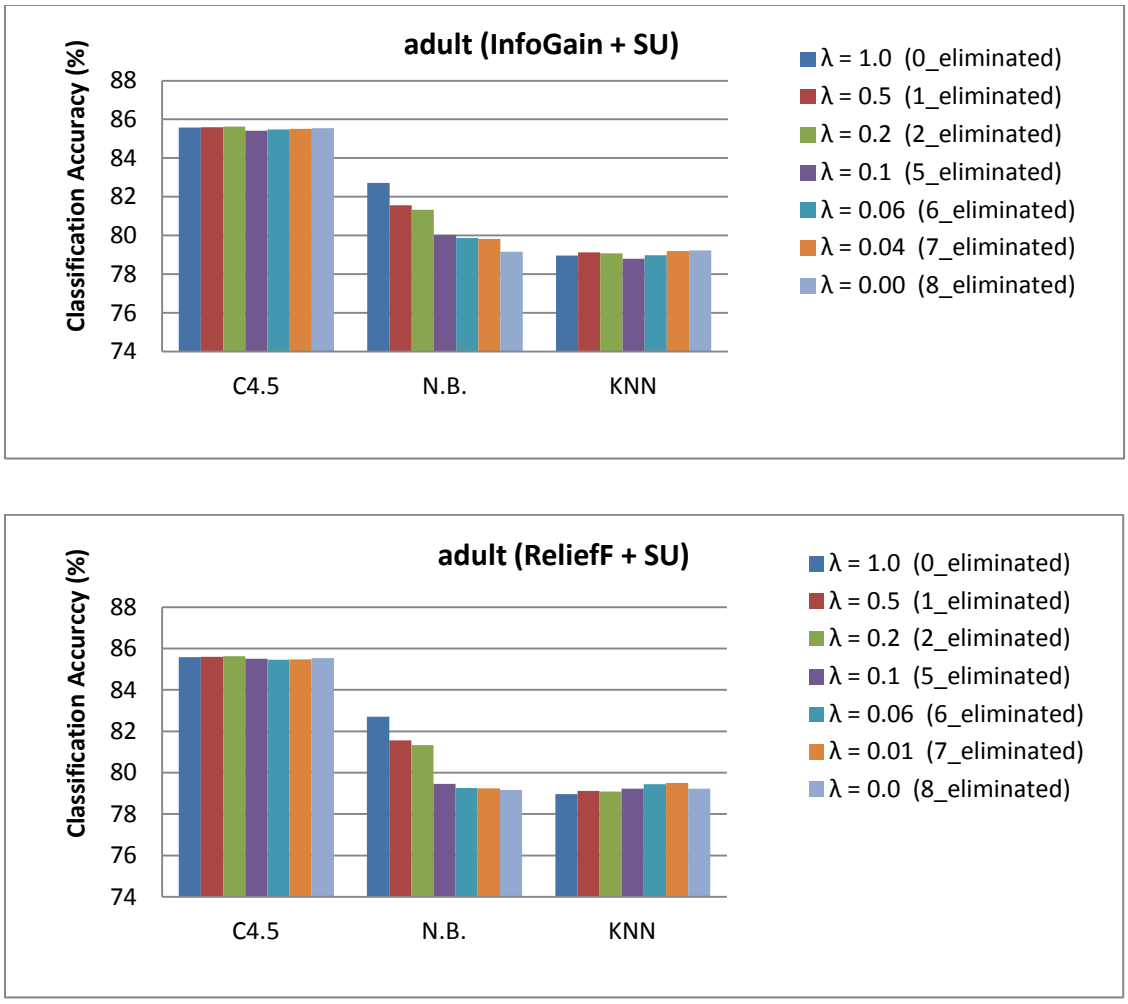



Figure 6.3: The Privacy-aware Filters results corresponding to selected datasets.

A more generic solution is to divide attributes into higher risk attributes and lower risk attributes. There are many practical examples from real datasets which show that such distinction is realistic. For example, an attribute such as the level of blood pressure certainly has a lower (if it has any) privacy risk value compared with the gender of a given patient. Identifying the potential privacy breaching attributes as QI attributes is a special case where the privacy risk of all attributes in the QI set is assumed to be equal.

Our proposed approach in this thesis is limited to in its dependency on the correlation among attributes. Therefore, in the extreme cases where all QI attributes are relevant and there is a very low correlation between attributes, any reduction of QI attributes may result in a significant reduction in the performance.

## 6.3. Summary

In this chapter we showed that filter-based feature selection with individual feature ranking can be modified to address privacy consideration without impacting the performance of the classifiers. We proposed the *PF-IFR* algorithm which exploits the correlation between attributes in order to eliminate the QI attributes. In this work, we considered equal importance is given to all attributes in the QI set from privacy point of view given. In future we will consider the case where different risk factors are associated with the QI attributes. We will also consider more datasets with different inter-correlation among their attributes. We will further consider privacy-aware filter-based feature selection methods with subset ranking as well as privacy-aware wrappers.

# Chapter 7

# Privacy-aware Wrappers

This chapter summarizes our work in (Jafer et al. 2015c). It was mentioned in Chapter 4 that, the main advantage of wrappers is that, they usually result in higher performance compared with filters, and in cases where performance is the main objective they become very practical. Such higher performance provides us with a space of maneuver and could be exploited for privacy preserving purposes. In other words, wrappers provides the user with a space of performance gain that can be used up towards achieving more privacy as this is shown and discussed in this chapter.

In general, the goal of feature selection is to obtain the most optimal feature set that achieves best performance. Without privacy considerations, an optimal feature subset is a set that would maximize the performance metric, and yet has the minimum cardinality, i.e. has the minimum number of attributes. Therefore, in the default setting, features get selected according to their positive impact/relevancy with respect to the performance

measure (i.e. the classification accuracy). In this work, we want to incorporate privacy considerations into the very functionality of wrappers. Recall from Chapter 4 that, in the wrapper approach, subsets of variables are assessed based on their usefulness to a given predictor. Wrappers conduct a search for a good subset using the learning algorithm itself as part of the evaluation function.

Similar to privacy-aware filters, we aim at incorporating privacy into wrappers by design. In doing so, we turn wrapper-based feature selection into a privacy-aware process. Without privacy considerations, wrappers would select a subset of the original features whose classification performance is no worse than that of the original feature set. To make a privacy-aware wrapper, we pay special attention to the quasi-identifier (QI) set which refers to a set of attributes that are not privacy breaching per se but if they are linked to external sources could potentially identify individuals via identity disclosure (Chapter 2). We show that by making wrappers privacy-aware, more quasi-identifier attributes can be eliminated without negatively impacting the achieved classification accuracy in any significant way. We also identify a privacy region that can be exploited in order to achieve privacy vs. utility trade-off. Our empirical results show that the high performance achieved by wrappers provides us with a space in which we can obtain valuable privacy gains.

To this end, we introduce a Privacy-aware Wrapper (*PW*) system which incorporates privacy into the functionality of wrappers. It ensures that privacy gain is achieved through the selected features without negatively impacting the performance of the models if compared with the performance of the original dataset.

This chapter is organized as follows.. The methodology of the proposed PW algorithm is discussed in Section 7.1. Section 7.2 shows the experimental results and discusses them. Finally Section 7.3 summarizes this chapter and discusses some of the potential future directions.

## 7.1. Methodology of Privacy-aware Wrappers

In discussing the logic behind this proposed technique, it is beneficial to re-consider different levels of feature relevancy which were discussed in the previous section.

Consider Figure 7.1, adapted from (Kohavi and John 1997). In this figure, the feature subset is divided into strongly relevant attributes (*SR*) shown in dark grey, weakly relevant attributes (*WR*) shown in light grey, and irrelevant attributes (*IR*) shown in white. When privacy is our main concern, our goal becomes eliminating as many QI attributes as possible as long as this process does not degrade our performance significantly. It is possible that attributes in the QI attribute set may fall into *IR*, *WR*, and *SR* regions. Recall from Section 4.1 that an optimal feature set consists of all strongly relevant attributes and those weakly relevant attributes that are non-redundant. In general, if the QI attributes fall into the *IR* region and the *WR* region (where attribute are redundant), they will not be selected by the feature selection algorithm anyways because they are not part of the optimal feature subset. Recall also that strongly relevant attributes provide unique information about the target class and cannot be removed or replaced by other attributes. Our focus is rather on the portion of quasi-identifier attributes in the *WR* region that are non-redundant but could be ignored provided that their elimination does not (significantly) impact the classification accuracy.

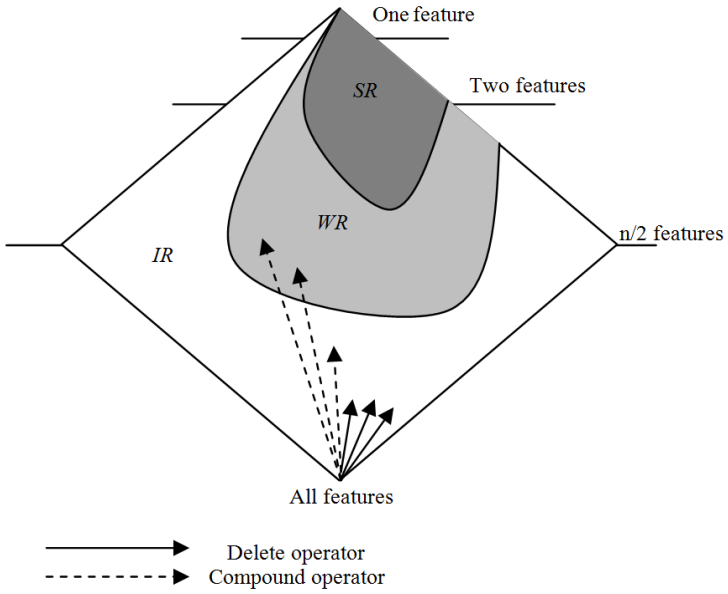
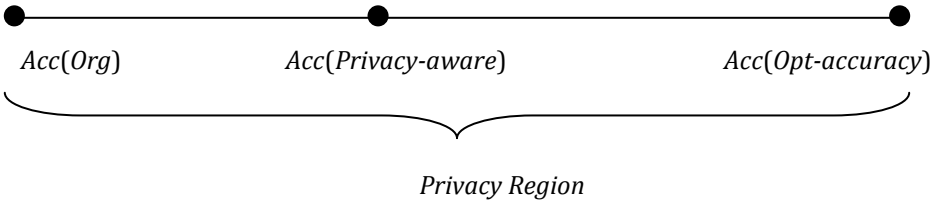



Figure 7.1: The strongly relevant, weakly relevant, and irrelevant attributes (Kohavi and John 1997).

Applying wrappers improves performance over the data set, thus creating a margin of added accuracy in which the data holder could implement his/her privacy preferences. We call this margin a *privacy region* shown in Figure 7.2. *Acc*(*Opt-accuracy*) in Figure 7.2 represents the accuracy corresponding to the optimal feature set. That is, the classifier that is built using only selected attributes by wrapper feature selection. At the other end of the bar, *Acc*(*Org*) represents the accuracy of building classifiers with the complete feature set (without feature selection). *Acc*(*Opt-accuracy*) – *Acc*(*Org*) represents a range of acceptable accuracy of a privacy-aware feature set that is obtained via privacy-aware wrappers. Without privacy considerations, the goal of wrapper-based feature selection is to obtain a feature set which achieves best classification accuracy. However, by incorporating privacy into feature selection, the goal becomes obtaining a feature set that is not necessarily optimal from the classification accuracy's point of view; however, it is privacy-aware. In other words, we give up some of the achieved classification accuracy to gain privacy. The gain in privacy is identified as removing more QI attributes and as long as *Acc*(*Privacy-aware*) is not significantly lower that *Acc*(*Org*) , the result is acceptable. In other words, in Figure 7.2, the choice of (*Privacy-aware*) becomes a trade-off between gaining more/less privacy and achieving lower/higher accuracy respectively.




*Org*: Original feature set
*Opt-accuracy*: Optimal feature set with respect to accuracy
*Privacy-aware*: A privacy-aware feature set considering utility vs. privacy trade-off

Figure 7.2: Illustration of the privacy region.

Before getting into further details, let us consider Figure 7.3 which illustrates the concepts behind our technique. For simplicity, in this figure, {*A*} corresponds to the complete set of attributes, {*B*} represents to the set of QI attributes, and {*C*} refers to the

selected set of attributes obtained by applying Wrapper-based Feature Selection (WFS) to the original dataset.

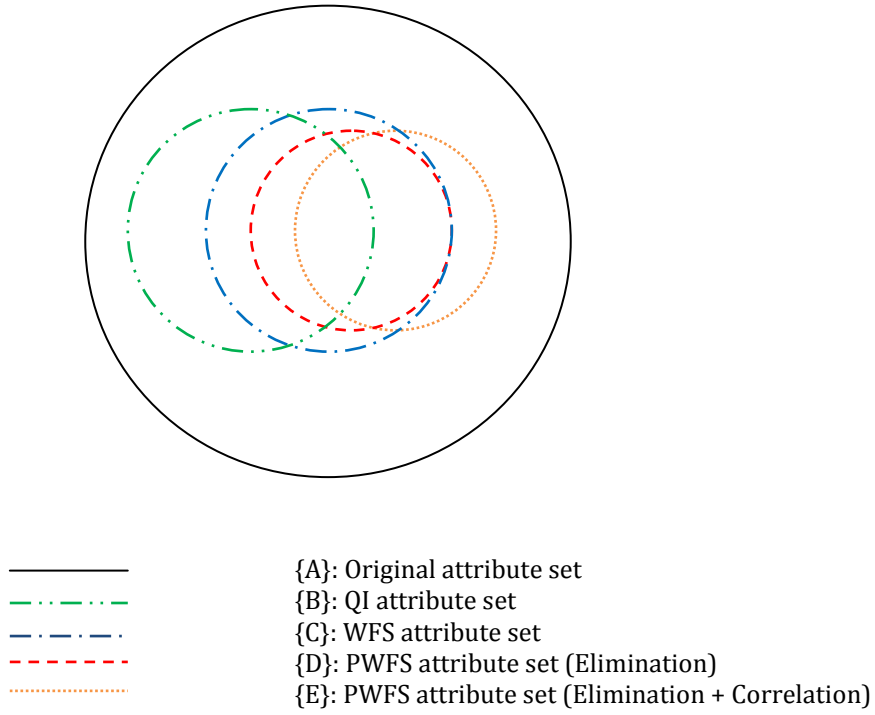



Figure 7.3: Illustration of interactions between different subsets (i.e. Original, QI, WFS, PWFS).

The intersection between {$B$} and {$C$} (i.e. {$B$} ∩ {C} shows the set of quasi-identifier attributes that are selected by WFS. To this end, one observation is that even without any privacy consideration, wrapper feature selection results in elimination of some of the QI attributes. Such an observation is in accordance with the work in (Jafer et al. 2014b).

The proposed technique here is that, by incorporating privacy into the wrapper feature selection further elimination of QI attributes is achieved at two levels. This is shown by subset {$D$} and {$E$} in Figure 7.3. In the first level, further QI attributes are eliminated (i.e. attribute set {$D$}| {$D$} ⊂ {$C$}). In the second level, additional QI attributes are replaced by other non-QI attributes from the original attributes set to obtain subset {$E$}. We notice that {$E$} may contain features in {$A$} that were excluded by wrapper feature selection and were not included in {$C$}. This is a backtrack step confirming that our goal is no longer obtaining

an optimal feature set solely from classification point of view. Rather, the privacy element becomes another factor to be taken into consideration.

In general, the goal of classification is to minimize the test error and maximize the classification accuracy. Therefore, many of the classification techniques implicitly solve optimization problems. In our method, our aim becomes maximizing the number of eliminated QI attributes. Our constraint is that in such process the classification accuracy should not be degraded below a specified threshold significantly.

Given the dataset $\boldsymbol{D}$ consists of a set of $n$ training instances where each instance $\boldsymbol{X}$ is an element of the set $F_1 \times F_2 \times \ldots \times F_m$, and $F_i$ is the domain of the $i$th feature. The original feature vector consists of $\boldsymbol{X}$ attributes $\{X_1, X_2, \ldots, X_m\}$ in addition to the class attribute $\boldsymbol{C}$. The accuracy of the original attribute set is denoted as $Acc(\boldsymbol{D}[\boldsymbol{X}, \boldsymbol{C}])$. This is the classification accuracy of a model (e.g. C4.5, N.B., etc) built using dataset $\boldsymbol{D}$ with complete attribute set. Assume that ($\boldsymbol{QI'}$) refers to a set of QI attributes that are eliminated/replaced; $MAX(\boldsymbol{QI'})$ refers to a set of maximum number of QI attributes that get eliminated from the final selected feature set. Our objective is therefore to: *find the maximum number of QI attributes which could be removed (via elimination and/or correlation) such that, the model built with the resulting dataset (with fewer projected features) does not degrade the classification accuracy significantly.* That is,

$$\textbf{find } MAX\,(\boldsymbol{QI'}) \mid (Acc(\boldsymbol{D}[(\boldsymbol{X}\text{-}\boldsymbol{QI'}), \boldsymbol{C}]) \geq Acc(\boldsymbol{D}[\boldsymbol{X}, \boldsymbol{C}]) \;||\; Diff\,(Acc(\boldsymbol{D}[\boldsymbol{X}, \boldsymbol{C}]), Acc(\boldsymbol{D}[(\boldsymbol{X}\text{-}\boldsymbol{QI'}), \boldsymbol{C}])) \; != \text{statistically significant}) \tag{7.1}$$

Recall from Chapter 2 that, a table is considered $K$-anonymous if given a record that has some QIs values, there are at least $K$-1 other records which have the same QIs values. Assuming that we have a dataset with $\{QI_1, QI_2, \ldots, QI_j\}$ and we want to $K$-anonymize the dataset with $K = n$. Let $M$ represents the difference between the anonymized dataset $A$ and the original dataset $D$. That is, the more generalization and suppression, the more difference between the two. $M$ is a function of $K$ and $\varphi$ which represents the number of eliminated QI attributes i.e. $M = f(\frac{K}{\varphi})$. From this formula, the higher is the $K$, the larger is the equivalence class. Equivalence class refers to the set of tuples that share the same QI values. Such larger equivalence class implies more modification, i.e. higher $M$. However,

with increased number of eliminated QI attributes at the same privacy level, (i.e. the same *K*) within each new equivalence class, more details are preserved which eventually results in datasets that are less modified, i.e. having smaller *M*.

Our proposed Privacy-aware Wrapper (PW) system is depicted in Figure 7.4. The system consists of three main blocks, namely, the *Evaluation* block, the *Elimination* block, and the *Correlation* block.

The inputs to the system include the dataset ***D***, the evaluation criteria ***e***, the feature selection type ***FS***, the list of potentially privacy breaching attributes ***PA***, the correlation measure ***c***, and the discretization parameter ***d***. These inputs are discussed in more details as we explain the functionality of the system. ***PA*** = QI is a special case where all potentially privacy breaching attributes are given equal weight and therefore it is not important which attribute in the QI set gets eliminated. We follow this assumption in our technique since such approach is commonly used in privacy preserving data publishing. The evaluation criteria ***e*** refers to the classification accuracy of a chosen classifier such as C4.5, N.B, KNN, etc. ***FS*** in our case refers to the specific type of wrapper and the search technique employed (e.g. best first + forward selection, best first + backward elimination, etc). The initial evaluation of the dataset is performed in the *Evaluation* block.

The output of the *Evaluation* block includes two sets of attributes, namely ***WS*** and ***R***. ***R*** refers to a ranked list of attributes in the original feature space. This rank is obtained by individually considering the ability of features in predicting the target class. ***WS*** refers to the list of selected attributes due to applying wrapper-based feature selection to the dataset ***D***. In general, ***WS*** is a subset of the original attribute set, i.e. $\boldsymbol{WS} \subset \boldsymbol{R}$. At this point, two possible outcomes may occur. If $\boldsymbol{WS} \cap \boldsymbol{PA} = \emptyset$, this is an indication that, none of the attributes in the ***PA*** set were selected by the wrapper and the algorithm would terminate since there are no additional potentially privacy breaching attributes to be eliminated. In most scenarios this is not the case. That is, wrappers eliminate some of the ***PA*** attributes, but some others will still be present in the ***WS*** feature set which need to be handled in the upcoming blocks.

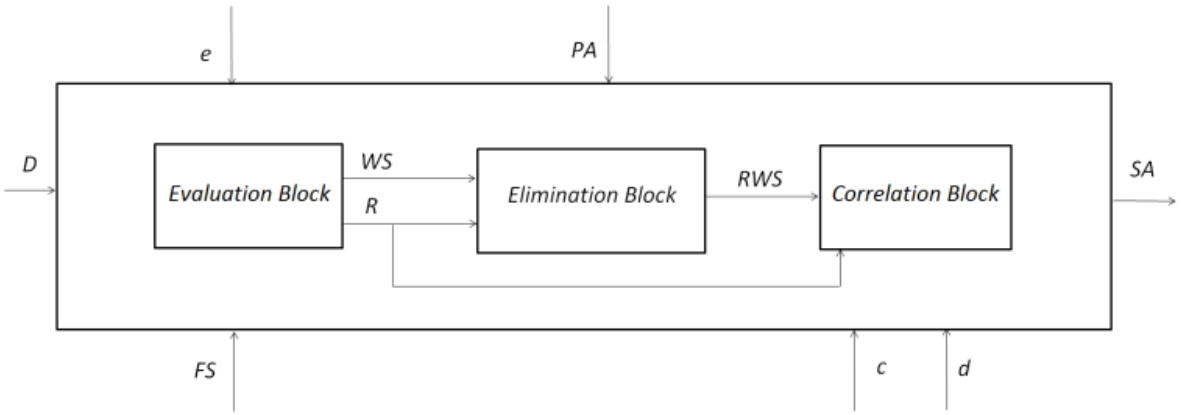



Figure 7.4: Privacy-aware Wrapper (PW) System.

The next block in the system is the *Elimination* block and its goal is to remove more QI attributes as long as such removal does not degrade the performance of the resulting models significantly. The input to this block includes ***WS*** and ***R***. Using the rank of individual attributes in the original set, a ranked list of ***WS*** is obtained, i.e., ***RWS***. Then using the ***PA*** list, the QI attributes in the ***RWS*** set are identified. In the following step, starting from the bottom of the ***RWS*** list, the QI attributes are removed and the classification model is built with the remaining attributes. If the new accuracy is higher, equal, or (not significantly lower) than the original accuracy, then it is safe to remove this particular QI attribute. The ***RWS*** list is updated accordingly until all of the QI attributes are visited.

Following the above step, if $\boldsymbol{RWS} \cap \boldsymbol{PA} = \emptyset$, it indicates that all of the QI attributes were eliminated in the *Elimination* block. The algorithm terminates and, ***SA*** which is equal to ***RWS*** is released. In other words, the *Correlation* block is executed only if after the logic of *Elimination* block has been executed, $\boldsymbol{RWS} \cap \boldsymbol{PA} \neq \emptyset$.

Now let us consider the *Correlation* block in more detail. Similar to the *Elimination* block, our goal is to reduce the number of QI attributes that could not be eliminated previously. One immediate observation is that removing of any of these QI attributes at this step would result in a significant reduction in performance. After all, this is the main reason these attributes could not be removed in the previous step. However, it is possible that these QI attributes are highly correlated with other attributes in the original feature space, i.e. ***X***.

The correlation measure ***c*** can any correlation criteria such as Mutual Information (*MI*), Symmetric Uncertainty (*SU*), and so on, and is considered the choice of the user. Some of

these correlation measures require that the numerical attributes be discretized first. In this work, we choose ***c*** to be the Symmetric Uncertainty (*SU*) metric. As it was mentioned in Section 6.2, *SU* is a modified information gain measure in order to estimate the degree of association between discrete features (Press 1988). It measures the correlation between two features and more specifically the amount of information a given attribute provides about another attribute. For a reminder, the formula (6.1) is shown below again:

$$SU = 2.0 \times \left[\frac{H(X) + H(Y) - H(X,Y)}{H(X) + H(Y)}\right]$$

*H(X)* and *H(Y)* refer to the entropy of attributes *X* and *Y* respectively, and *H(X,Y)* refers to the joint entropy of *X* and *Y*. Since *SU* is symmetric it can be used to measure feature-feature correlations in which there is no notion of one attribute being the "class" attribute (Hall 1999). The discretization factor, *d* is used and is set to either true or false. When it is set to true the type of discretization technique needs to be specified. It is required for the *SU* measure that the attributes be categorical. We set ***d*** to true and discretize the numeric features using the technique of (Fayyad and Irani 1993).

The inputs to the *Correlation* block include ***R*** and ***RWS***. Similar to the pervious block, we start with the bottom of the ***RWS*** list searching for the QI attributes. Whenever such an attribute is found, we find its correlation with all of the attributes that have higher rank in the ***R*** list and are neither in the ***RWS*** set nor in the ***PA*** set.

That is, ***NR*** = ***R*** – ***PA*** – ***RWS***. We then find the attribute(s) in the ***NR*** set that have higher rank compared with the rank of an encountered QI attribute in the ***RWS*** set, and construct the ***HNR*** set. Following this step, we find the attribute in ***HNR*** that has maximum correlation with our QI attribute. In the following step, we replace our QI attribute with that attribute *xn'* and build the classifier with the updated ***RWS***. The feature vector in this case would be *D*[*RWS*-{*candidate QI attribute to be removed*}+{replacement attribute from the HNR list, i.e. *xn'*}]. If the new accuracy is higher, equal, or (not significantly lower) than the original accuracy, then it is safe to finalize the removal of the given QI attribute and the replacement is considered safe. If not, we repeat the process with the next attribute in the ***HNR*** set that has the highest correlation with the QI attribute.

*PA* = A set of potentially privacy breaching attributes (e.g. the QI set); *X* = Original feature set;
*C* = Target class; *R* = Ranked *X* i.e. {$X_1, X_2, \ldots, X_n$} (1: highest rank and n: lowest rank)
*WS* = Selected attributes as a result of applying wrapper feature selection on the original dataset (with *X* feature set);
*RWS* = Ranked *WS* i.e. {$rws_1, rws_2, \ldots, rws_m$} (1: highest rank and *m*: lowest rank)

**Input** = *D, e, PA, FS, c, d*

*Original_Acc* = *Acc*(*D*[*X*], *C*)

**if** (*RWS* ∩ *PA*) = 0 **go to** A

```
for (i = m, i > 0, i--) /* in RWS */
{
    if ({rws_i} ⊂ PA)
    {
        New_Acc = Acc(D[RWS-{rws_i}], C)
        if ((New_Acc ≥ Original_Acc) || Diff(New_Acc, Original_Acc) != Statistically Significant)
        RWS = RWS – { rws_i}
    }
}
```

**if** (*RWS* ∩ *PA*) = 0 **go to** A

*NR* = *R* – *PA* – *RWS*

```
for (i = m, i > 0, i--) /* in RWS */
{
    if ({rws_i} ⊂ PA)
    {
     /* Identify NR attributes with higher rank compared with the rank of rws_i */
    HNR = NR | NR_r >= {rws_i}_r
    /* HNR has p attributes */
    /* Find the attribute xn' in HNR set that has highest correlation with {rws_i} */
        for (j = p, j>0, p--) { /* in HNR */
            [max_corr, xn']= MAX(find_correlation({rws_i}, {hnr_j}, c))
            New_Acc = Acc(D[RWS-{rws_i}+xn'], C)
            if ((New_Acc ≥ Original_Acc) || Diff(New_Acc, Original_Acc) !=
        Statistically Significant)
            {
                RWS = RWS – {rws_i} + xn'
                break
            }
        }
    }
if (RWS ∩ PA) = 0 go to A
A: SA = RWS
```

**Output** = SA

Figure 7.5: The PW algorithm.

If the replacement was successful we stop the search, break out of the loop, and continue the procedure with the next QI attribute in the ***RWS*** set. If after scanning all of the attributes in the ***HNR*** set we could not replace the QI attribute, such QI attribute will remain in the final subset. The system's output is obtained through Privacy-aware Wrapper (*PW*) algorithm which is formally presented in the algorithm presented in Figure 7.5.

## 7.2. Experimental Results and Discussions

This section summarizes the application of our *PW* algorithm on a number of real datasets obtained from the UCI repository (http://archive.ics.uci.edu/ml/). In most of these datasets, a subset of features was selected as the QI attributes set (Keng-Pei and Ming-Syan 2011). For two datasets, i.e. *CMC* and *Wisconsin breast cancer*, all of the attributes except the target class were selected as the QI set. In many cases, such an assumption is not realistic and it is made for illustration purposes only.

We used four classifiers, namely, C4.5, N.B., KNN, and SVM. We applied a *t*-test to compare our results. All the models were built using 10-fold cross validation. Our results are shown in Table 7.1. The accuracy values in the baseline column refer to the results of the classifiers built using the complete original attribute set, i.e., prior to applying any feature selection on the dataset. They provide us with a reference point.

Recall from Figure 7.2 that the baseline (or original) accuracy corresponds to *Acc*(*Org*). WFS refers to the classification accuracy of a dataset that consists only of attributes selected by wrapper-based feature selection. This accuracy corresponds to *Acc*(*Opt-accuracy*) in Figure 7.2. It was mentioned earlier that wrapper-based feature selection (by itself) is not privacy-aware and therefore QI attributes may be eliminated due to their characteristics and their relevancy/irrelevancy to the target class. Finally, *PW* refers to the accuracy of the dataset consists of attributes selected by the privacy-aware wrapper. This corresponds to *Acc*(*Privacy-aware*) in Figure 7.2.

We show the number of QI attributes in the baseline dataset, in the dataset resulting from applying WFS, and in the dataset resulting from applying PW. The *p*-value corresponding to the t-test is shown in the final column as well. The *p*-value corresponds to the *PW* algorithm results compared to the baseline.

Let's consider the results associated with the *pima* dataset, case of C4.5. Initially there are 3 QI attributes in this dataset. The baseline accuracy is 73.82%. This is the accuracy of building a C4.5 classifier from the original dataset without any modification. If we apply WFS, one of the three QI gets eliminated. WFS results in a dataset with optimal accuracy (75.787%) which is higher than the baseline accuracy. By applying *PW* we are able to further reduce the number of QI attributes without significant reduction in the accuracy compared with the baseline accuracy. With PW we achieve lower accuracy compared with WFS which is expected. Our algorithm reduces the QI attributes in a step-by-step approach until any further reduction of the QI attributes results in an accuracy that is significantly lower than the baseline accuracy. In our case we notice that even if all QI attributes are removed the resulting accuracy is acceptable. In such a case, with PW algorithm we can remove all QI attributes. The privacy implication is that the resulting dataset can be released without any anonymization or modification. We see the same behavior in the case of N.B., KNN, and SVM. In fact such behavior is seen for other datasets/classifiers combinations in Table 7.1.

Table 7.1: Comparison results of the performance and privacy obtained from the original dataset, WFS, and PW. ⊕/⊖ corresponds to statistically significant increase/decreases.

| **Dataset (No. tuples)** | **Alg.** | **Baseline** | | **WFS** | | **PW** | | |
|---|---|---|---|---|---|---|---|---|
| | | **No. QI** | **Acc%** | **No. QI** | **Acc %** | **No. QI** | **Acc %** | ***p*-value** |
| Pima (768) | C4.5 | 3 | 73.8 | 2 | 75.7 | 1 | 74.3 | 0.6306 |
| | | | | | | **0** | **73.1** | 0.7133 |
| | N.B. | 3 | 76.3 | 2 | 77.7 | 1 | 77.2 | 0.2964 |
| | | | | | | **0** | **75.9** | 0.6970 |
| | KNN | 3 | 72.6 | 2 | 73.5 | 1 | 70.7 | 0.0808 |
| | | | | | | **0** | **71.7** | 0.6445 |
| | SVM | 3 | 77.3 | 3 | 77.3 | 2 | 76.8 | 0.3732 |
| | | | | | | 1 | 75.6 | 0.0613 |
| | | | | | | **0** | **75.9** | 0.1197 |
| German credit (1000) | C4.5 | 6 | 70.7 | 2 | 73.1 | 1 | 73.5 | 0.1636 |
| | | | | | | **0** | **73.0** | 0.2382 |
| | N.B. | 6 | 75.4 | 3 | 76.2 | 2 | 76.0 | 0.5462 |
| | | | | | | 1 | 76.0 | 0.5203 |
| | | | | | | **0** | **75.9** | 0.5366 |
| | KNN | 6 | 73.3 | 1 | 75.1 | **0** | **72.9** | 0.7828 |
| | SVM | 6 | 75.1 | 2 | 75.8 | 1 | 74.8 | 0.4344 |
| | | | | | | **0** | **74.6** | 0.6141 |
| liver patients (583) | C4.5 | 2 | 68.7 | 0 | 71.0 | N/A | N/A | N/A |
| | N.B. | 2 | 55.7 | 0 | 71.8 | N/A | N/A | N/A |
| | KNN | 2 | 64.6 | 2 | 71.5 | 1 | 71.3 | 0.5864 |
| | | | | | | **0** | **67.9** | 0.1702 |
| | SVM | 2 | 71.3 | 0 | 71.3 | N/A | N/A | N/A |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| heart stat logs (270) | C4.5 | 2 | 76.6 | 0 | 85.1 | N/A | N/A | N/A |
| | N.B. | 2 | 83.7 | 1 | 86.2 | **0** | **85.5** | 0.0957 |
| | KNN | 2 | 78.8 | 0 | 84.4 | N/A | N/A | N/A |
| | SVM | 2 | 84.0 | 0 | 84.0 | N/A | N/A | N/A |
| CMC (1473) | C4.5 | 9 | 52.1 | 3 | 55.5 | **2** | **50.3** | 0.1703 |
| | | | | | | 1 | 47.5⊖ | 0.0046 |
| | N.B. | 9 | 50.7 | 3 | 55.3 | 2 | 50.7 | 0.9648 |
| | | | | | | **1** | **47.8** | 0.1690 |
| | KNN | 9 | 45.2 | 3 | 52.6 | 2 | 51.9⊕ | 0.0001 |
| | | | | | | **1** | **47.998** | 0.1273 |
| | SVM | 9 | 48.2 | 9 | 48.2 | 8 | 45.8⊖ | 0.0443 |
| | | | | | | 7 | 45.6⊖ | 0.0443 |
| | | | | | | 6 | 45.5⊖ | 0.0345 |
| Wisc. breast cancer (683) | C4.5 | 9 | 93.4 | 2 | 95.1 | **1** | **91.7** | 0.1776 |
| | N.B. | 9 | 97.3 | 6 | 97.8 | 5 | 97.9⊕ | 0.0367 |
| | | | | | | 4 | 97.3 | 0.9966 |
| | | | | | | **3** | **96.6** | 0.2126 |
| | | | | | | 2 | 93.5⊖ | 0.0005 |
| | | | | | | 1 | 91.7⊖ | 0.0005 |
| | KNN | 9 | 95.3 | 7 | 96.3 | 6 | 96.3⊕ | 0.0438 |
| | | | | | | 5 | 95.9 | 0.2214 |
| | | | | | | 4 | 95.6 | 0.6372 |
| | | | | | | 3 | 95.4 | 0.7826 |
| | | | | | | **2** | **94.2** | 0.1536 |
| | | | | | | 1 | 92.5⊖ | 0.0007 |
| | SVM | 9 | 96.1 | 6 | 97.0 | 5 | 96.4 | 0.5095 |
| | | | | | | **4** | **95.6** | 0.3965 |
| | | | | | | 3 | 94.7⊖ | 0.0229 |
| | | | | | | 2 | 94.7⊖ | 0.0497 |
| | | | | | | 1 | 92.6⊖ | 0.0029 |
| adult (45222) | C4.5 | 8 | 85.5 | 3 | 85.7 | 2 | 85.7 | 0.1043 |
| | | | | | | 1 | 85.7 | 0.2637 |
| | | | | | | **0** | **85.5** | 0.8894 |
| | N.B. | 8 | 82.7 | 8 | 82.7 | 7 | 82.5⊖ | 0.0148 |
| | | | | | | 5 | 81.3⊖ | 9.4e-09 |
| | | | | | | 4 | 80.3⊖ | 3.6e-13 |
| | | | | | | 3 | 80.0⊖ | 1.5e-12 |
| | KNN | 8 | 81.3 | 1 | 85.2 | **0** | **82.7** ⊕ | 4.0e-05 |

The *German* credit dataset shows similar behavior as of the *pima* dataset. Once again, it is possible to remove all QI attributes and obtain an acceptable accuracy. It is possible, according to the *PW* algorithm, that all of the QI attributes get removed by feature selection anyways. See the results for the liver patients' dataset. In three cases of C4.5, N.B, and SVM, WFS does not select any of the two QI attributes identified for this dataset. Therefore, there are no more QI attributes to be removed by the PW method. This is shown by N/A in the corresponding cells. KNN, however, is different. In this case, none of the QI attributes get removed by WFS; However, the PW technique is able to remove/replace the QI attributes

without degrading the accuracy. In fact accuracy is higher than the baseline even if all of the QI attributes are removed from the dataset.

In the *heart stat logs* dataset, similar to the liver patient dataset, WFS eliminates the two QI attributes identified for this dataset (i.e. for C4.5, KNN, and SVM classifiers). For N.B., while PWS removes one of the QI attributes

Earlier in this section, it was mentioned that in the cases of *CMC* and *Wisconsin breast cancer* datasets the assumption is that all attributes except the target class are considered QI attributes. Once again, while WFS removes some of the QI attributes, PWFS provides room for additional removal of QI attributes and implicitly resulting in higher privacy.

Let us consider the *CMC* dataset. With the exception of SVM, for other three classifiers WFS excludes six attributes. PW eliminates more QI attributes while maintaining acceptable accuracy. In the case of C4.5 removal of one extra QI attribute is possible. However, removing of additional attribute reduces the accuracy significantly. This is shown by $\ominus$ next to the accuracy of 47.5%. Therefore, the best acceptable result is 50.3%. For the N.B. classifier, the best achieved result is 47.9% where two additional attributes get excluded. As for the KNN classifier, exclusion of two more extra attributes does not have a negative impact on accuracy significantly. As for SVM, it is not possible to remove any QI attribute without significantly degrading the classification accuracy. We only show three attempts of removal. Wisconsin breast cancer dataset shows similar results.

For the *Adult* dataset, in the case of C4.5, by applying PW it is possible to remove/replace three extra QI attributes and achieve accuracy of 85.5% which does not significantly different from the baseline accuracy. We compare our results for the case of C4.5 classifiers with other workload-aware anonymization techniques discussed in Section 2.5.2 and Section 3.1. Most of these works run their experiments using *Adult* dataset and C4.5 classifier. We notice that, with our technique, due to the fact that all identified QI attributes are removed, no anonymization is needed. Furthermore, the resulting accuracy is comparable to the baseline accuracy. This comparison is shown in Table 7.2.

Table 7.2: Comparison results of the performance of different workload-aware anonymization techniques and privacy-aware wrappers (C4.5 classifier).

| C4.5 Classifier – 10 fold cross validation | Attributes Used | Parameter/Best Achieved Result | privacy model |
|---|---|---|---|
| (Iyengar 2002) | Only 8 attributes + class attribute | K=10 (82.7%) | K-anonymity |
| (Wang et al. 2004) | Only 7 categorical attributes + class attribute | K=10 (81.6 %) | K-anonymity |
| (Fung et al. 2005) | Only 8 attributes + class attribute | K=10 (82.8 %) | K-anonymity |
| (Fung et al. 2005) | All attributes | K= 10 (85.2 %) | K-anonymity |
| (Fung et al. 2007) | All attributes | K= 10 (85.2 %) | K-anonymity |
| (Mohammed et al. 2009) | All attributes | C = 20%, L = 2, K=20 (85.2 %) | *LKC-privacy* |
| (Mohammed et al. 2011) | All attributes | ε =4 (83.8%) | Differential Privacy |
| Our Method | All attributes | All QI attributes are eliminated (85.5 %) | K-anonymity (where applicable) |

The same is applied for the case of KNN classifier. One exception is the N.B. classifier. In this case, all attributes are selected by the wrapper feature selection. Therefore WFS does not eliminate any of the QI attributes. Any removal of the attributes in the QI set significantly degrades the classification accuracy. The correlation block cannot be implemented since all attribute are selected in the first place, i.e., R = *RWS, PA* ⋂ *RWS* = *PA*, and *NR* is empty.

Privacy-aware wrappers inherit the complexity and the performance issues associated with wrapper-based feature selection since this method is an extension of wrapper-based feature selection in which privacy considerations lead to further refinement of the ultimate selected subset. We address this shortcoming by introducing the privacy-aware evaluation measure.

In classification algorithms such as decision tree, some attributes may help separating the classes in one branch while other attributes might help separating the classes in other branches. Wouldn't removing the attributes eliminate such possibility? It should be mentioned that our technique, at each step, considers the impact of removal or replacement of given attribute by another on the classification accuracy. We know that wrapper-based feature selection is adjusted to data distribution in terms of importance of attributes. The datasets that we are dealing with are not completely balanced. Therefore, it

is not the case that we will remove an attribute that is good for  example, for 90% of the data, but the one which impacts 10% of the data. We remove the attributes that may slightly contribute to better accuracy, however, as we mentioned at the beginning of this chapter, we aim to exercise controlled reduction of classification accuracy in order to gain more privacy.

Another question is, why do not we publish the model rather than publishing the dataset. It should be noted that in bioinformatics the datasets have too large number of attributes that we need to decrease. Also, in medical research it is a requirement that the data itself is to be published as well and not only the model. In other words, even if the model is published, the data should be published along the model.

In the *PW* system, the complexity of the blocks following the *Evaluation* block (which implements a standard non privacy-aware wrapper) is considered minimal. A future direction is to empirically study the added complexity of the system.

It was mentioned in Section 7.2 that, reducing QI enhances the models that support identity disclosure via reducing the risk of re-identification. Another future direction is to study the impact of reducing the number of QI attributes, hence changing the size of equivalence class (number of tuples that share the same QI values) in the context of attribute disclosure. This includes applying *PW* prior to anonymization techniques such as *l*-diversity (Machanavajjhala et al. 2006) and *t*-closeness(Li et al. 2007). In all of the above models the assumption is that the attacker knows the QI of the victim. An existing challenge, however, is determining the QI attributes which is/remains an open issue in PPDP.

## 7.3. Summary

In this chapter, we incorporated privacy into the very process of wrapper-based feature selection. Our results showed that, compared with basic wrappers, our proposed *PW* system, was either able to eliminate all of the QI attributes or was able to exclude more QI attributes. These objectives were attained while maintaining a good utility that does not differ significantly from the utility of the original dataset.

# Chapter 8

# A Multi-dimensional Privacy-aware Evaluation Function in Automatic Feature Selection

In Chapters 5, 6, and 7, we proposed techniques (e.g., PF-IFR and PW) that turn the feature selection into a privacy-aware process. These techniques had one common characteristic. In all of them, the used evaluation measure was classification accuracy. In other words, the evaluation measure of feature selection technique was not modified. The goal was that to use feature selection and while using it aiming to reduce the number of QI attributes as much as possible without degrading the resulting classification accuracy.

In this chapter, we aim at turning feature selection into a privacy-aware process by changing its very evaluation measure. This chapter summarizes the work in (Jafer et al. 2015a) and (Jafer et al. 2015b). In there, we incorporate privacy considerations into the very evaluation measure that is used to evaluate and select feature subsets.

Generally, the goal of automatic feature selection is to obtain the most optimal feature set that provides for algorithm's best performance. In our proposed technique, feature selection process is guided such that the data holder can predict the amount of inference of an individual's sensitive attribute by the attacker who has access to sophisticated data mining tools.

We consider privacy during the feature selection process and as such, introduce a two-dimensional measure in automatic feature selection that takes into account both objectives of privacy and efficacy (e.g. accuracy) simultaneously and provides the data user with the flexibility of trading-off one for another. We introduce a two-dimensional measure in wrappers. Our technique generates a list of candidate privacy-aware attribute subsets and enables the data user to trade-off privacy and efficacy.

In the second part of this chapter, we also introduce $E(S)$ a multi-dimensional privacy-aware evaluation function in automatic feature selection that enables the data holder to select and eventually release a best subset according to its desired efficacy, privacy, and dimensionality of the resulting dataset.

This chapter is organized as follows. In Section 8.1, Section 8.2, and Section 8.3, we introduce the PBI measure of privacy and its application. Section 8.4 introduces the candidate privacy-aware attribute subset generating system. This system mainly consists of two subsystems, namely, correlation-aware attribute ranking (Section 8.4.1) and candidate attribute subset generator (Section 8.4.3). Section 8.6 introduces the multi-dimensional privacy-aware evaluation function. The experimental results, discussions and future directions are discussed in Section 8.7 and Section 8.8 consequently.

## 8.1. Measuring Privacy-preserving Feature Selection

We setup the stage for introducing the new evaluation function by some background informaiton about the steps that are taken in order to generate privacy-aware candidate attribute subsets. We first start with the case there C!=SA (Jafer et al. 2015a). We then show how the algorithms intrduced in (Jafer et al. 2015a) could be slightly modified in order to adapt the case where C=SA.

### 8.1.1. Basic Notations

Let $D$ be a dataset with tuple set $T=\{t_1, t_2, \ldots, t_n\}$ and attribute set $A=\{a_1, \ldots, a_m\}$. The assumption is that $D$ is a subset of some larger population $P$. Therefore, the distribution of the attribute values in $D$ represents the distribution of the population as whole. The sensitive attribute (e.g. medical condition of patients) is denoted as SA $\in A$. Usually, the assumption is that the sensitive attribute will be released without any modification. However, the association between individuals and their sensitive attributes should be kept secret. Attribute C $\in A$ represents the target class attribute. Baseline attribute subset $BL$ refers to a subset of attributes in $A$ excluding SA and C. That is, $BL = A\backslash(C \cup SA)$. In our work, we assume that C and SA are different. In other words, we assume that the goal of the adversary differs from the legitimate user of the dataset (e.g. researcher). The case where C=SA is discussed later in Section 8.5. Baseline accuracy subset $BLC$ refers to a subset of all attributes except SA (i.e. $A$ - SA). Baseline privacy subset $BLP$ refers to a subset of all attributes except C (i.e. $A$ - C). Figure 8.1 illustrates these subsets. A dataset $DBLP$ is the projection of the tuple set $T$ onto the attributes in $BLP$ and a dataset $DBLC$ is the projection of tuple set $T$ onto the attributes in $BLC$. These notations will be used later when we discuss the algorithm.

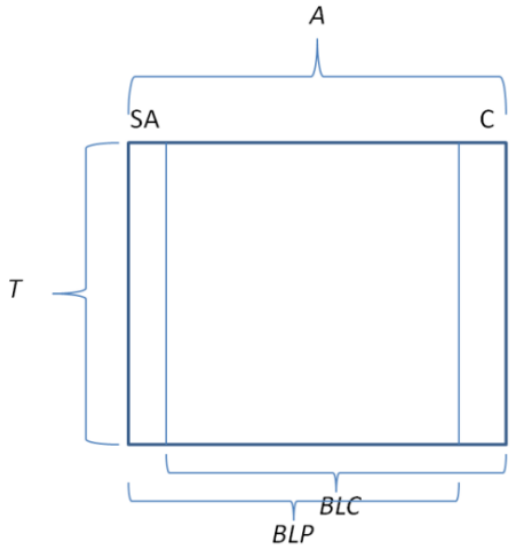



Figure 8.1: Dataset D and the projection of S attributes (case C!=SA).

## 8.2. Defining the PBI Measure of Privacy

The measure of privacy, i.e. *Privacy Breach Increase* (*PBI*) (Jafer et al. 2015a) is inspired by the notion of "empirical privacy" proposed in (Cormode et al. 2013). Similar to "empirical privacy", our measure represents the precision with which the attacker can infer the sensitive values of individuals from released data. It is inspired by a widely adopted notion of privacy breach which refers to the correct posterior inference of an adversary about sensitive values in the data. Such measure considers a sophisticated attacker who will use data mining tools for his/her attack. The empirical privacy studies the increase/decrease of attacker's inference ability as a function of the amount of anonymization (Cormode et al. 2013). For example, how much such inference changes when the dataset is $K$=100 anonymized vs. $K$=1000 anonymized. In the *PBI* measure, privacy breach increase (or decrease) is considered as a function of the selected attributes in the final dataset (or implicility, the number of eliminated attributes). The work in (Cormode et al. 2013) considers an adversary who would use the output $A(D)$ (where $A$ refers to an anonymization mechanism) to build a classifier in order to attack anonymized data. With the assumption that feature selection is done prior to anonymization, we consider an attacker who will use only the selected features (compared with the complete feature set i.e. baseline dataset) in order to build a classifier to predict the sensitive attribute. We study the impact of reduction of selected attributes on the overall ability of the attacker in correctly predicting the value of the sensitive attribute.

Let *DS* represents the projection of a selected attributes set $S$ on the dataset $D$ (where $S \subset A \backslash C$). We use *DS* to build a model of the data in order to classify SA. Assume that $Acc(DS)$ and $Acc(DBLP)$ refer to the accuracy of correctly predicting the value of SA when the classifier is built using *DS* and *DBLP* respectively.

**Definition 1**: Given dataset $D$, a projected dataset *DS*, and the baseline dataset $DBLP$, we compute *Privacy Breach Increase (PBI)* as

$$PBI(DS) = \left(\frac{Acc(DS)}{Acc(DBLP)}\right) - 1 \qquad (8.1)$$

*PBI* is the characterization of privacy in terms of ease/difficulty of correctly predicting the sensitive attributes using a given attribute set. From formula (8.1), if *PBI* is positive, it indicates that the privacy risk of *DS* is higher than the baseline dataset, and likewise, a negative value of PBI indicates that, the privacy risk of *DS* is lower than the baseline dataset. We consider PBI to be computed as percentage of the baseline privacy. The minimum privacy breach is referred to as distribution privacy *DistP*. The amount of allowed *PBI* of a given dataset is chosen by DH. For example, if the *PBI* associated with a given attribute subset is (-10%), it means that it becomes (10%) more difficult for the attacker to correctly predict the correct value of SA for a given individual when compared with attacking the baseline dataset. The distribution privacy is simply the distribution of the sensitive attributes which represents the population distribution and is known to the adversary (since the assumption is that the sensitive attributes will be released). If we are able to publish a dataset in which the prediction of the sensitive attribute is not more than the sensitive attribute distribution *DistP*, even if the attacker uses the same dataset in order to infer the values of sensitive attribute, his/her posterior belief will not change as the result of releasing the dataset. Although not the focus of this thesis, this observation could be related to the notion of ε-differential privacy in which the disclosure of the privacy of any individual should not substantially (bounded by $\varepsilon$) disclosed as a result of participation in a statistical dataset.

One remaining question is that, to what extend the proposed method is robust against malicious attacks in which the attacker would explore in-depth understanding of the data (e.g. correlation between attributes, etc)? Our privacy objective of a selected subset is achieved proportional to the baseline. In other words, we ensure that the privacy breach increase due to release of a given dataset (with selected features) is lower than that of the baseline and below a given threshold. To this end, the aim is that to minimize the power of the attacker in correctly predicting the correct value of SA as a result of dataset release. We insist that, the purpose here is to limit and control the amount of inference of sensitive attribute(s) due to the release of the dataset. Our aim is to control the difference between prior and posterior belief of the attacker (i.e. before and after seeing the released dataset). If the attacker is able to infer SA using external datasets, such inference will be made anyways and regardless of releasing/not releasing the dataset held by DH. In other words,

it is DH's responsibility to manage the dataset it owns and to be released. Furthermore, we assume that the recipient of the dataset (DR) is a well-known and trusted entity for a given analysis purpose.

## 8.3. Application of *PBI*

**Objective:** We want to find an attribute subset $S \subset A \backslash (C \cup SA)$ such that (1) the probability of achieving correct posterior inference about the sensitive attribute (i.e. SA) of an individual based on $S$ is below a given threshold $\alpha$ (privacy) and (2) the difference between the performance of the baseline attribute subset *BL*, represented by *DBL* dataset, and the selected attribute subset $S$ (in terms of predicting the target attribute C), represented by *DS* dataset, is not statistically significant or is in favor of the selected attribute subset (efficacy). Formally, we want to find:

$$S| \ (PBI(DS) \leq \alpha) \& \ \Big( \big( Perf(DS) \ \geq Perf(DBL) \big) \ || \ \ (|Perf(DBL) - \ Perf(DS)|)! = \\ statistically \ significant \Big) \qquad (8.2)$$

From this objective, it is possible to obtain more than one attribute subset which satisfies the above constraints. We refer to these attribute subsets as **candidate privacy-aware attribute subsets**. Depending on the user's preferences, within the space of candidate privacy-aware subsets four possible regions could be identified which are shown in Figure 8.2. We want to enable the user to select a trade-off between performance (i.e. classification accuracy) and privacy (i.e. *PBI*) based on his/her requirements and priorities. The user might be willing to give up some utility in order to gain more privacy and vice versa. Each point in Figure 8.2 represents a selected attribute subset with two associated values (*Perf*(*DS*), *PBI*(*DS*)).

*Perf*(*DS*) refers to the accuracy of the attribute subset in predicting the target class C i.e. classification accuracy. *PBI*(D*S*) refers to the increase in the likelihood of privacy breach of the attribute subset S in predicting the sensitive attribute SA compared with the baseline. *DistP_PBI* refers to the *PBI* associated with the distribution of the sensitive attribute in the whole dataset w.r.t the baseline. This is the maximum privacy guarantee and is unaffected by the records in the dataset.

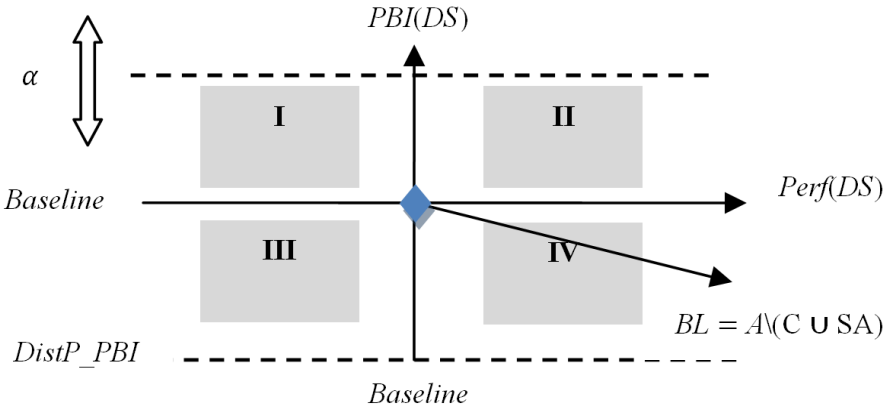



Figure 8.2: Illustration of performance vs. privacy trade-off

These regions are listed as follows:

- **Region I (NW)**: The attribute subsets in this region have a *PBI*(*DS*) and *Perf*(*DS*) higher and lower than the baseline respectively**.** In other words, any other candidate attribute subset in other regions will surpass the attribute subsets in this region because provide either better performance or better privacy or both. NW, in other words is the worst region to be at**.**
- **Region II (NE)**: The attribute subsets in this region result in higher *PBI*(*DS*). However, they achieve higher performance compared with the baseline.
- **Region III (SW)**: The attribute subsets in this region have performance that is lower than the baseline performance; however, their *PBI*(*DS*) is also lower than the baseline.
- **Region IV (SE)**: The attribute subsets in this region achieve better performance compared with the baseline while incur less potential privacy breach since their *PBI*(*DS*) is lower than that of the baseline. In this region, the best region/location to be at *PBI*(*DS*) = *DistP_PBI*.

## 8.4. Candidate Privacy-aware Attribute Subset Generating System

The initial privacy-aware feature selection system consists of two subsystems, namely, *Ranker* and *Candidate Subset Generator* (Figure 8.3). The details of each of these subsystems are provided in the following subsections.

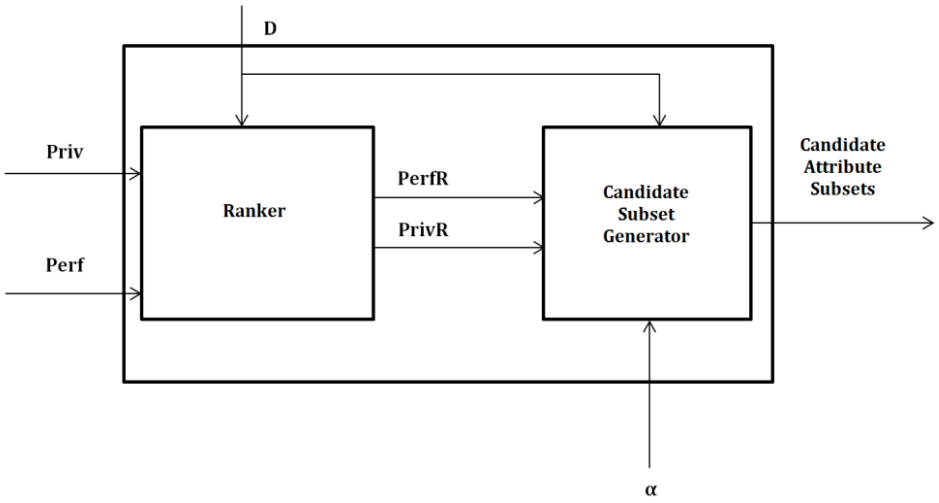


Figure 8.3: Candidate Privacy-aware Attribute Subset Generating System.

### 8.4.1. Correlation-aware Attribute Ranking

The privacy risk associated with each attribute may be different. Our system first generates two ranked list of attributes, namely, a privacy rank (*PrivR*) and a performance rank (*PerfR*). Consider the privacy rank (*PrivR*). Our aim is to identify the impact of each attribute in correctly predicting the sensitive attribute SA. We start with the assumption that SA differs from the target attribute C. We first build a classifier using attributes in *BL* ∪ SA w.r.t. SA, i.e. to obtain *BLP* (See Section 8.1.1). In order to find the privacy risk of each attribute, only one attribute is removed at a time and then the projection of tuple set *T* onto the attributes in the remaining attribute set is obtained and the accuracy of the projected dataset w.r.t. SA is calculated. The difference between accuracies of *DBLP* and the projected dataset is calculated. This difference indicates the impact of removing a given attribute on increasing/decreasing the accuracy of predicting the SA attribute. We then rank the attributes from lower to higher. The higher the rank, the higher the attribute's contribution in correctly classifying SA, i.e. the more harmful the attribute is. As such, *PrivR* will consist of attributes that are ranked from lower to higher privacy risk.

The same logic is followed in order to obtain *PerfR*. The difference is that, in this case the baseline attribute set becomes *BLA*. Starting from the complete list of attributes, and after eliminating attributes one-by-one we build a classifier using the remaining attributes in order to predict the target class attribute C. Finally, the difference between accuracies of *DBLA* and the projected dataset is calculated and the ranked list of attributes according to

their performance attribution is obtained. *PerfR* rank the attributes from higher to lower. This algorithm is shown in Figure 8.4.

The output of this algorithm consists of two ranked list of attributes *PrivR* and *PerfR* which will be used as input to the next algorithm discussed in the next section.

The reason for ranking privacy and performance lists differently (lower to higher, higher to lower respectively) is that when the ranked list is used in the next algorithm, in backward elimination, starting from the last attribute in the ranked list, we tend to, first, remove attributes that have higher privacy risk and lower impact on performance.

**Input:** Dataset *D* (*A* attributes and *T* tuples, sensitive attribute SA, target attribute C)

$BL = A \setminus (SA \cup C)$  $BL = \{BL_1, BL_2, ..., BL_k\}$
$BLP = BL \cup SA$
$BLC = BL \cup C$

*DBLP = proj(D, BLP)*
*DBLC= proj(D, BLC)*
*Perf_BLP = Acc(DBLP)*
*Perf_BLC = Acc(DBLC)*

//This function returns a list of all attributes ranked according to the required order.
Rank((attribute, value), order); order{high_to_low, low_to_high}:

**for** (i = 1, i ≤ k, i++) // *k* refers to the number of attributes in *BL*
{
  $BL_{temp\_i} = BLP - \{BL_i\}$
  $DBL_{temp\_i} = proj(D, BL_{temp\_i})$
  $BLP_i = Perf\_BLP - Acc(DBL_{temp\_i})$
  *PrivR* ←({BLi}, $BLP_i$)
}

*PrivR* = Rank(*PrivR*, low_to_high)
**for** (j = 1, j ≤ k, j++) // *k* refers to the number of attributes in *BL*
{
  $BL_{temp\_j} = BLC - \{BL_j\}$
  $DBL_{temp\_j} = proj(D, BL_{temp\_j})$
  $BLC_j = Perf\_BLC - Acc(DBL_{temp\_j})$
  *PerfR* ←({BLj}, $BLC_j$)

}
*PerfR* = Rank(*PerfR*, high_to_low)

**Output**: privacy-based ranked attributes (*PrivR*), performance-based ranked attributes (*PerfR*)

Figure 8.4: Privacy-based and Performance-based Ranking Algorithm.

## 8.4.2. Searching for Candidate Attribute Subsets

Before categorizing our search for candidate privacy-aware subsets in the context of existing search techniques we briefly consider these techniques/strategies as follows.

#### *8.4.2.1. Search Strategies*

Search strategies are, in general, divided into two categories, namely, *Uninformed* and *Informed* (Russell et al. 2010). The *Uninformed* search strategies do not have any additional information about the states beyond what is provided in the problem definition. The *Informed* (a.k.a. heuristic) search strategies use problem-specific knowledge beyond the definition of the problem itself. As such, these strategies can find solutions more efficiently compared with the uninformed strategies. In heuristic search strategies additional knowledge of the problem is imparted to the search algorithm.

Three approaches to greedy heuristic search are *best-first*, *hill-climbing*, and *beam search*.

In *best-first search* nodes are expanded one at a time according to some definition of best. Best-first search family is the least oriented towards solving a problem as quickly as possible (Wilt et al. 2010). Best-first algorithms are complete. The reason for completeness is that they have an open list of unbounded size and therefore they terminate only when they have found a solution or when they emptied the complete open list (Wilt et al. 2010).

In a basic *hill climbing algorithm* only one node is expanded at a time beginning with the root. It expands only the best child of the previously expanded node. Therefore, it can be said that these algorithms do no work more that what they really need to do. As such, one can expect hill climbing algorithms to find solutions with very little effort and allow them to be very competitive. There are, however, two major drawbacks associated with hill climbing: (1) some of them lack a tunable parameter so that a trade-off between speed and quality could be selected, and (2) they frequently fail to find solutions (Wilt et al. 2010).

*Beam searches* are grouped into two categories, namely, best-first beam search and breadth-first beam search. The *best-first beam* search is similar to a best-first search, the difference is that when the open list grows beyond a predetermined size limit, the lowest

quality nodes get removed from the open list until the open list is within its size bound (Rich and Knight 1991). The other kind of beam search is a *breadth first beam* search which is similar to breadth first search. Breadth-first search traverses a graph (or a tree) by starting at the root and then, first, explores the neighbor nodes prior to moving to the next level neighbors. The breadth first beam differs from a regular breadth-first is such a way that, at each depth layer a fixed number of nodes are expanded and the remaining are pruned (Bisiani 1992).

In many search problems, the path to the goal is irrelevant (Russell et al. 2010). As such, it is possible to consider different types of algorithms that do not give importance to the path at all. In other words, what is important is the solution state. Local search algorithms operate using a single node and move only to neighbors of that node. There are two main advantages associated with the local search algorithms as it is stated in (Russell et al. 2010): (1) they need very little memory because the paths followed are not retained. (2) they usually find reasonable solutions in large or infinite state spaces where it is not suitable to use systematic algorithms. In addition to finding goals, a main benefit of using local search algorithms is that they are very useful in order to solve optimization problems. In optimization problems the goal is to find the best state according to given objective function. Objective function is also considered a heuristic cost function. In addition to local hill-climbing search, there are other local search algorithms such as simulated annealing, local beam search, and genetic algorithms which are beyond the scope of this work.

#### *8.4.2.2. A Search Guided by PrivR and PerfR Ranks*

To this end, in the context of search, we identify our search space to be the space of all subsets. Obviously, with $n$ attribute such search will include $2^n$ possibility which is NP-complete. After all, this is the main reason that wrappers usually follow a heuristic approaches (e.g. best-first with forward selection or backward elimination). Our search for privacy-aware candidate subset is informed or heuristic in the sense that the subsets are selected based on an evaluation function. This evaluation function is guided by the rank of attributes in the *PrivR* and *PerfR* list since the order in which the attributes are eliminated is dictated by these ranks. Our search operators remove one attribute at a time and therefore we conduct a local search. Therefore, we essentially follow a stepwise backward

elimination procedure. We use backward elimination in order to preserve features whose usefulness requires other features. The very definition of stepwise backward elimination is that, we start with full set of attributes and then, at each step, remove the worst attribute that remains in the set. Such selection of the "worst" attribute is dictated by the *PerfR* and *PrivR*. Our search is guided by both *PerfR* and *PrivR*. In the case of *PerfR* we rank the attributes based on their relevance w.r.t. the target class. Therefore, worst attributes refer to the ones that are the least predictive. When search is guided by the *PrivR* list, worst attributes refer to the ones that are most predictive of the sensitive attribute. After each removal the remaining subset is tested against the accuracy and the privacy requirements. If both requirements are met the given subset is added to the list of candidate subsets. If not, the last removal is cancelled and the next attribute in the ranked list is removed.

In our search methodology, similar to hill climbing, only one node is expanded at a time with the goal of finding a local maxima. This is different than best first search where the goal is to find a global maxima and either to find an optimal solution or to empty the complete open list. Such an inability to find an optimal solution is well understood in the context of privacy-aware feature selection. The very definition of optimal feature subset in the case of regular wrapper, is to obtain a feature subset with the minimum number of attributes that has highest predictive accuracy. From formula (8.2), in a privacy-aware feature selection, we conduct a search for subsets that satisfy our efficacy and privacy requirements simultaneously. In most of the cases, however, it is not possible to find a given subset with the highest efficacy and lowest *PBI* (our privacy measure) simultaneously. This objective becomes less realistic if we add a third requirement of having minimum number of attributes. Therefore, our search aims to obtain a list of subsets (hence candidate subsets) that satisfy the efficacy and privacy requirements.

Obtaining a single subset is an extra step beyond generating candidate privacy-aware attribute subsets by means of incorporating accuracy, privacy, and dimensionality weights based on the DH's preferences. As such, it becomes possible to narrow down the search for feature subset to a single subset according to the data holder's preferences. Introducing a multi-dimensional evaluation function is the main contribution of this work. We start with the background information that precede the definition of the proposed multi-dimensional evaluation function.

### 8.4.3. Candidate Attribute Subsets Generator

After obtaining ranked list of attributes in *PerfR* and *PrivR*, our next step is to follow a backward elimination search technique that is guided by these ranked lists. The reason for having two ranked list *PerfR* and *PrivR*, as mentioned in the previous section, is to get the spectrum of subsets of attributes that have both performance and privacy as their main priority. This approach provides the data holder/publisher with greater flexibility in selecting the final subset of attributes.

**Input:** Dataset *D* (*A* attributes and *T* tuples, sensitive attribute SA, target attribute C), {*PrivR*|*PerfR*}, *Perf_BLC* (Performance of baseline w.r.t. C), *Perf_BLP* (Performance of baseline w.r.t. SA )

*RankedList*= {$R_1$, $R_2$, …, $R_k$} // depending on the input either {*PrivR*|*PerfR*}
n= 0;
**for** (i = k, i>0, i--) // *k* refers to the number of attributes in *BL*
{
*S* = *RankedList* – {$R_i$}
S_C = *S* ∪ C
S_SA = *S* ∪ SA
*DBS_C* = *proj*(*D*, *S_C*)
*DBS_SA*= *proj*(*D*, *S_SA*)
**if** ((Acc(*DBS_C*) ≥ *Perf_BLC* ||($|Perf(DBS\_C) - Perf\_BLC|$)! = $statistically\ significant$)
{
*PBI*(*DBS_SA*) = ((1- Err(*DBS_SA*))/ *Perf_BLP*) -1
**if**(*PBI*(*DBS_SA*) ≤ $\alpha$ ) {
Candidate_Privacy_Aware_Subset[n] ← (*S*, Acc(*DBS_C* ), PBI(*DBS_SA*))
n++
}
} else { // the elimination of {Ri} violates the utility requirements, we need
// to put it back and continue with the next attribute in the ranked list
*RankedList* = *RankedList* ∪{$R_i$}
}
}
**Output:** Candidate_Privacy_Aware_Subset

Figure 8.5: Privacy-aware Candidate Subset Generator.

The input to privacy-aware candidate attribute subset generator algorithm includes dataset *D*, the *PrivR* list, the *PerfR* list, the value of *Perf_BLC* (Performance of baseline attribute set *BL* with respect to the target class C), and the value of *Perf_BLP* (Performance of baseline attribute set *BL* w.r.t. SA ).We start with the bottom of a given list (being *PrivR* or *PerfR*) and eliminate the attributes one-by-one. At each elimination, we check both privacy and performance constraints. If both constraints are satisfied the given attribute is removed and the search proceeds with the next one. If one of the constraints will be

violated, do not eliminate this attribute and go to the next attribute in the list. Repeat this process until any further elimination would violate one of the constraints or when the attribute list is used up. The algorithm is shown in Figure 8.5.

Using the output of privacy-aware candidate attribute subset generator algorithm we populate the PBI(DS) vs. Perf(DS) diagram. Therefore, each point in that diagram represents a subset of attributes. It was mentioned earlier in this section that our favorite subsets of attributes are in region IV. Presumably, we are interested in a candidate attribute subset in IV with the minimum number of features. However, we need to ensure that there is no other candidate set S' such that it is a subset of S and has higher PBI. In other words we must ensure that,

S| if (S'$\subset$ S ) => PBI(DS') $\geq$ PBI(DS). DH is responsible to controlling the amount of privacy prior to release of a dataset (with selected feature vector).

We select the *Pima* dataset in order to show our approach in detail. For the remaining datasets/algorithms only the final candidate attribute subsets is shown. We consider that the goal is to build a C4.5 classifier. The attacker goal is to build a classifier in order to infer the sensitive attributes of the individuals. The *Pima* dataset consists of nine attributes, namely, *Preg*, *Plas*, *Pres*, *Skin*, *Insu*, *Mass*, *Pedi*, *Age*, and *Class*. We assume that *Preg* is the S.A. attribute. These attributes are defined in the UCI repository.

First step includes obtaining *PerfR* and *PrivR* ranks following the algorithm in Figure 8.4. These ranks are as follows:

*PerfR* = {Plas,Mass,Pres,Skin,Age,Pedi,Insu}

*PrivR* = {Skin,Insu,Pres,Mass,Pedi,Plas,Age}

After obtaining the ranks, following backward elimination strategy guided by these ranks, we eliminate the attributes step-wise and obtain the associated Perf(DS) and PBI(DS) for the remaining attribute subset. These results corresponding to *PerfR* and *PrivR* are shown in Table 8.1 and Table 8.2 respectively.

Following a feature's elimination we run a t-test and record the difference in the performance of the projected dataset from the remaining attribute subset and the baseline. Hereafter, when we refer to the PBI(DS) and Perf(DS) of attribute subset we implicitly refer to the projected dataset where its feature vector includes the subset S. The baseline performance for this dataset is 73.70%. If the performance is higher than the baseline or

does not differ significantly, this is an indication that this attribute subset satisfies our performance requirements. In other words, the combination of the attribute subsets in Table 8.1 and Table 8.2 refers to the list of attribute subsets that yield acceptable performance. ⊕/⊖ refers to significantly higher/lower performance compared to the baseline performance.

Table 8.1: Pima - C4.5. PerfR-based candidate attribute subsets.

| Subset S | Perf(DS) | *p*-value | PBI(DS) |
|---|---|---|---|
| Plas_Mass_Pres_Skin_Age_Pedi (A) | 75.26⊕ | 0.044 | 0.97 |
| Plas_Mass_Pres_Skin_Age (B) | 75.52⊕ | 0.009 | 1.29 |
| Plas_Mass_Pres_Skin (C) | 73.83 | 0.931 | -3.08 |
| Plas_Mass_Pres (D) | 74.35 | 0.599 | -3.08 |
| Plas_Mass (E) | 74.48 | 0.525 | -2.75 |
| Plas (F) | 73.04 | 0.626 | -2.75 |

Table 8.2: Pima - C4.5. PrivR-based candidate attribute subsets.

| Subset S | Perf(DS) | *p*-value | PBI(DS) |
|---|---|---|---|
| Skin_Insu_Pres_Mass_Pedi_Plas (G) | 73.96 | 0.8544 | -3.24 |
| ~~Skin_Insu_Pres_Mass_Pedi~~ | ~~66.15⊖~~ | ~~0.0015~~ | N/A |
| Skin_Insu_Pres_Mass_Plas (H) | 73.57 | 0.9318 | -3.08 |
| Skin_Insu_Pres_Plas (I) | 72.91 | 0.6115 | -3.08 |
| Skin_Insu_Plas (J) | 73.30 | 0.7542 | -3.08 |
| Skin_Plas (K) | 73.30 | 0.7542 | -3.08 |

Note that in Table 8.2, elimination of the Plas attribute significantly reduces the performance compared with the baseline. In such a case, following the algorithm, Plas is put back and the next attribute in the PrivR rank i.e. Pedi is eliminated and candidate attribute subset *H* is obtained.

The corresponding two-dimensional diagram of the candidate attribute subsets is shown in Figure 8.6. Each point in this diagram refers to an attribute subset. For consistency and simplicity an alphabetic letter is given to each attribute subset. It is possible to identify the four regions discussed earlier in Figure 8.2. The dotted line in Figure 8.6 refers to $\alpha$. We assume that $\alpha$ is equal to the majority class distribution of S.A. i.e. *DistP_PBI.* The value of *DistP_PBI* is equal to (-2.75%) compared to the baseline and is obtained using formula (8.1). Technically, the most ideal attribute subsets include *H*, *C*, *G*, *D*, and *E*. We are interested in the attribute subset with fewer numbers of attributes. The

number of attributes associated with *H*, *C*, *G*, *D*, and *E* are 5, 4, 6, 3, and 2 respectively. Therefore, we may select *E* = {Plas, Mass} as the final attribute subset.

For comparison purposes, let us consider a generic wrapper feature selection that is not privacy-aware. The selected features using best-first greedy search (with both forward and backward elimination directions) includes {Plas, Pres, Mass, Age}. The performance of the projected dataset using these features is 75.78% which is higher than that of E (i.e. 74.48 %). However, the PBI associated of WFS is 2.43% which is beyond our accepted $\alpha$. In fact, when generic wrapper is used, the DH cannot have any control over the PBI of the resulting dataset. However, with the privacy-aware evaluation measure the amount of privacy breach increase can be controlled and managed by DH.

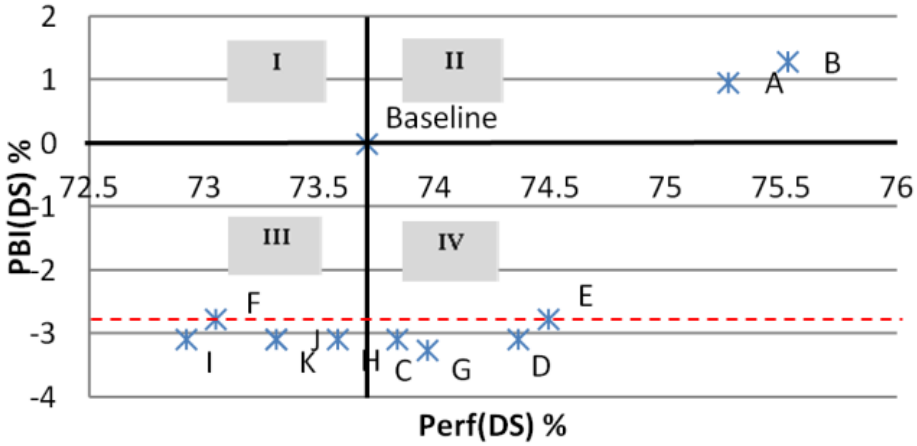



Figure 8.6: Pima dataset - C4.5. - PBI(DS) vs. Perf(DS).

By closely observing the results in Table 8.1 and Table 8.2 we notice that existence of the attribute Age is crucial in predicting the value of S.A. It follows that, whenever this attribute is included in the attribute subset, PBI(S) becomes positive and once it is removed PBI(S) becomes negative or near baseline.

One main challenge is how to ensure beforehand which classifier will be used by the attacker to attack the released dataset. In fact, we do not have any control over the data mining techniques that will be used by the adversary to attack the released dataset. We know that due to the no free lunch theorem there is no such classifier that will perform best on all datasets. There are two cases to consider here: First, if the maximum the attacker could gain is the majority class distribution of the sensitive attribute (i.e. *DistP* =

$Acc(DS)$), it means no matter which classification technique is used, the amount of inference of the sensitive will be bounded by $DistP$ since the majority class is independent of any classifier. Second, if $Acc(DS) > DistP$, we obtain the final candidate attribute subset to be released and build multiple classifiers against SA using the same candidate attribute subsets and check the resulting $Acc(DS)$ and check if $Acc(DS)$ of different classifiers is significantly higher than our main $Acc(DS)$. We adjust $\alpha$ accordingly. Let us consider this case in the context of our running example. Although our selected dataset $E$ belongs to the first category, however, for illustration purposes we use the selected candidate attribute subset E in order to mimic the attackers attempt to build a classifier with highest $Acc(DS)$. We chose some of the most commonly used classifiers such as SVM, N.B., LogisticR, KNN for our purpose. In each case attribute subset $E$ is used to build a classifier against the sensitive attribute. The results are shown in Figure 8.7.

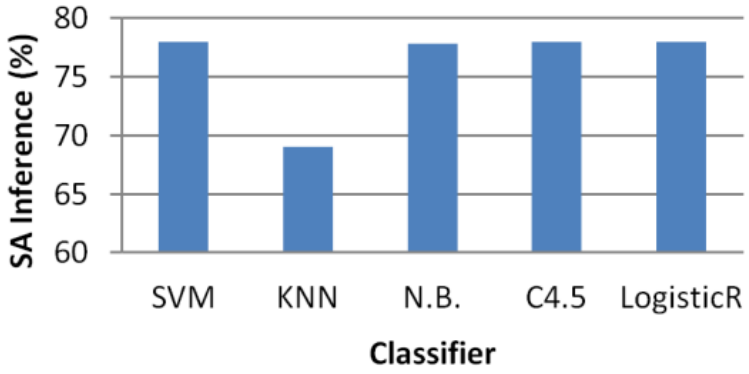



Figure 8.7: SA Inference (%) of different classifiers corresponding to E = {Plas, Mass}.

## 8.5. Extension of PBI and Ranker for the case where (C=SA)

So far, we assumed that the goal of the attacker and the legitimate user of the dataset is different (hence C!=SA). It is possible that these goals are the same in which both try to predict the value of C, one for legitimate purposes and the other for malicious purposes. The algorithms discussed so far could be modified slightly to accommodate such a change. For simplicity we assume that our dataset format is represented in Figure 8.8.

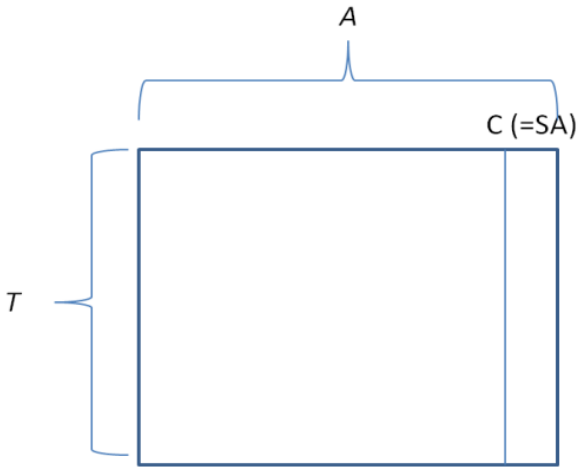


Figure 8.8: Dataset D and the projection of S attributes (case C=SA).

As before, the assumption is that the sensitive attribute will be released without any modification. However, the association between individuals and their sensitive attributes should be kept secret. Attribute C $\in A$ represents the target class attribute. Baseline attribute subset *BL* refers to a subset of attributes in $A$ excluding SA and C. That is, $BL = A \backslash C$.

Our modified measure of privacy i.e. *Privacy Breach Increase* is a slight alteration of the *PBI* measure introduced earlier for the case where C = SA.

Let *DS* represents the projection of a selected attributes set $S$ on the dataset $D$ (where $S \subset A \backslash C$). We use *DS* to build a model of the data in order to classify C. Assume that $Acc(DS)$ and $Acc(DBL)$ refer to the accuracy of correctly predicting the value of SA when the classifier is built using $S$ and *BL* respectively.

**Definition 2:** Given dataset $D$, a projected dataset *DS*, and the baseline dataset $DBL$, we compute the modified *PBI* as

$$\text{Modified } PBI(DS) = \left(\frac{Acc(DS)}{Acc(DBL)}\right) - 1 \tag{8.3}$$

Furthermore, contrary to the case of C!=SA (in which *PrivR* and *PerfR* are independent), when C=SA, *PerfR* and *PrivR* will be dependent and will have a reverse relation. The methodology of generating privacy-aware candidate attribute subsets in both cases is the same. The main idea being that a two-dimensional plot of Perf(DS) vs. PBI(DS) is obtained

and the four regions explained are populated as it is discussed in Section 8.3 and experimentally shown in Section 8.7.

## 8.6. Towards A Multi-dimensional Privacy-aware Evaluation Function

The *Evaluator* Subsystem (which is an extension of the subset shown in Figure 8.3) implements the proposed multi-dimensional privacy-aware evaluation function *E*(S). This extension is shown in Figure 8.9. We consider three factors; namely, the performance associated with a given subset *S*, i.e. *Perf*(S), the privacy associated with *S*, i.e. *Priv*(S), and the number of attributes in *S*, i.e. *Num*(*S*).

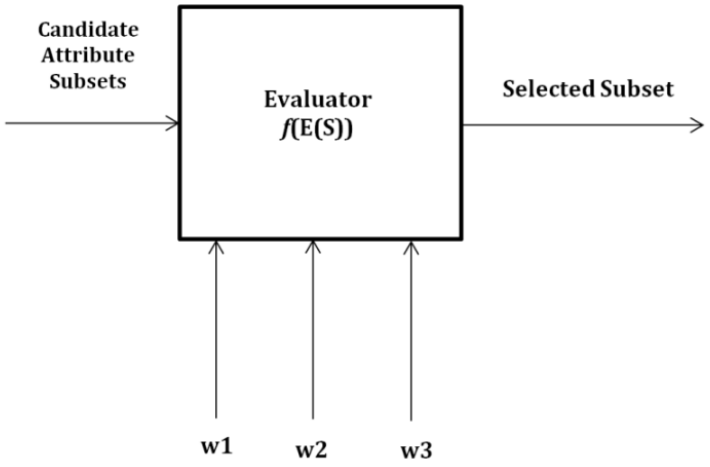



Figure 8.9: The Evaluator subsystems.

Ideally, within the list of candidate subsets, we are looking for a subset which has the best *Perf*(S) (i.e. **argmax**(Perf(S))), the best privacy (in the case of *PBI* as the privacy measure, the worst *PBI*) (i.e. **argmin**(PBI(S))), and the least number of attributes (i.e. **argmin**(Num(S))).

In practice, a given subset might satisfy none, some, or all of these combinations. We can list the possible combination of these factors being either true or false as follows. In binary encoding, (**$2^3$ =**) eight combinations may exist. These combinations are shown in Table 8.3.

Table 8.3: $2^3$ combinations of the three identified factors.

| Combination | argmax(Perf(S)) | argmin(PBI(S)) | argmin(Num(S)) |
|---|---|---|---|
| A | F | F | F |
| B | F | F | T |
| C | F | T | F |
| D | F | T | T |
| E | T | F | F |
| F | T | F | T |
| G | T | T | F |
| H | T | T | T |

It is very likely that we obtain three datasets each satisfying only one condition. In fact, the majority of cases belong to case 'A'. Let us use the running example (the Pima dataset from the UCI repository and the case of C4.5 classifier) in order to discuss this observation. After obtaining a list of candidate subsets of attributes, the number of attributes *Num*, the privacy breach increase *PBI*, and the performance *Perf* associated with each of the candidate subsets are recorded. The results are shown in Table 8.3. We keep baseline dataset for reference and comparison purposes.

Table 8.4: The n, PBI, and Perf associated with candiate subsets (Pima-C4.5). The corresponding combination category is shown in the last column.

| Subset | n | PBI | Perf | Combination |
|---|---|---|---|---|
| **baseline** | 8 | 0 | 73.835 | A |
| **plas_mass_preg_skin_pedi_age_pres** | 7 | 1.409 | 74.875 | E |
| **plas_mass_preg_skin_pedi_age** | 6 | 1.231 | 74.744 | A |
| **plas_mass_preg_skin_pedi** | 5 | 0.889 | 74.491 | A |
| **plas_mass_preg_skin** | 4 | 0.711 | 74.359 | A |
| **plas_mass_preg** | 3 | 0.711 | 74.359 | A |
| **plas_mass** | 2 | 0.882 | 74.486 | A |
| **plas** | 1 | -1.072 | 73.043 | B |
| **insu_pres_age_pedi_skin_preg_plas** | 7 | -1.942 | 72.401 | C |
| **insu_pres_age_pedi_skin_plas** | 6 | -0.708 | 73.312 | A |
| **insu_pres_age_pedi_plas** | 5 | -1.238 | 72.920 | A |
| **insu_pres_age_plas** | 4 | -1.241 | 72.919 | A |
| **insu_pres_plas** | 3 | -0.893 | 73.175 | A |
| **insu_plas** | 2 | -1.072 | 73.043 | A |

We highlight the cells with **argmax**(*Perf*(*S*)), **argmin**(*PBI*(*S*)), and **argmin**(*Num*(*S*)). Therefore, we identify three subsets that each satisfy only one ideal requirement i.e. having the best performance, the worst PBI, or the least number of attributes. These datasets belong to combinations 'E', 'C', and 'B' respectively (see Table 8.3). However, the remaining datasets belong to combination 'A' where none of the requirements are satisfied.

Our proposed evaluation function $E(S)$ relaxes such restrictions and provide the data holder with the flexibility of choosing a given factor (being performance or privacy or dimensionality of the data) over other(s). Since the main goal of feature selection is to reduce the dimensionality of data, we consider the number of attributes in the selected subset to be another factor. Our evaluation function is independent of a specific performance measure or privacy measure. It is a function that combines/blends the two measures in such a way that the data holder could exercise its privacy/utility preferences. The proposed measure is called $E(S)$ and is obtained as,

$$E(S) = w1.prf + w2.prv + w3.num \quad (8.4)$$

where *prf*, *prv*, and *num* are the rank of a given subset within the set of candidate subsets with respect to $Perf(S)$, $PBI(S)$, and $Num(S)$ respectively. In other words, rather than considering only the maximum or the minimum value of performance, privacy, or number of attributes of a given subset (which is highly limited as it was discussed earlier), we consider *prf*, *prv*, and *num* associated with each candidate subset within the set of candidate subsets.

$w1$, $w2$, and $w3$ refer to the weights given to each of these factors (ranks). We assume that, $\sum w = 1$, i.e. $w1 + w2 + w3 = 1$. Associating weights with the aforementioned factors has two benefits: First, it makes the combinatory evaluation function generalizable. In other words, we can ignore any of the factors by setting its corresponding weight to 0. Second, it allows the data holder to evaluate/select the subsets based on his/her preferences. For example, by setting $w2 = 4 * w1$, the privacy of the attribute subset is given four times of importance compared with the performance of the subset, and so on. When no preference is given to any of the factors, the default case refers to $w1 = w2 = w3 = 1/3$.

Technically, the goal of the *Evaluator* subsystem is to apply the evaluation function and to obtain a single selected subset with the best $E(S)$ given the selected weights:

$$f(E(S)) = \textbf{argmax}(E(S)) \mid w1, w2, w3) \quad (8.5)$$

Table 8.5: The candidate attribute subsets, their corresponding Perf(S), PBI(S), and Num(S) and their ranks (pima-C4.5). prf w.r.t. performance (lowest being the worst and highest being the best), prv w.r.t. PBI (highest being the worst and lowest being the best).

| Subset | n | Num. (R) | PBI | PBI (R) | Perf | Perf (R) |
|---|---|---|---|---|---|---|
| **baseline** | 8 | 1 | 0 | 7 | 73.835 | 8 |
| **plas_mass_preg_skin_pedi_age_pres** | 7 | 2 | 1.409 | 1 | 74.875 | 14 |
| **plas_mass_preg_skin_pedi_age** | 6 | 4 | 1.231 | 2 | 74.744 | 13 |
| **plas_mass_preg_skin_pedi** | 5 | 6 | 0.889 | 3 | 74.491 | 12 |
| **plas_mass_preg_skin** | 4 | 8 | 0.711 | 5 | 74.359 | 9 |
| **plas_mass_preg** | 3 | 10 | 0.711 | 5 | 74.359 | 9 |
| **plas_mass** | 2 | 12 | 0.882 | 4 | 74.486 | 11 |
| **plas** | 1 | 14 | -1.072 | 10 | 73.043 | 4 |
| **insu_pres_age_pedi_skin_preg_plas** | 7 | 2 | -1.942 | 14 | 72.401 | 1 |
| **insu_pres_age_pedi_skin_plas** | 6 | 4 | -0.708 | 8 | 73.312 | 7 |
| **insu_pres_age_pedi_plas** | 5 | 6 | -1.238 | 12 | 72.920 | 3 |
| **insu_pres_age_plas** | 4 | 8 | -1.241 | 13 | 72.919 | 2 |
| **insu_pres_plas** | 3 | 10 | -0.893 | 9 | 73.175 | 6 |
| **insu_plas** | 2 | 12 | -1.072 | 10 | 73.043 | 4 |

Continuing with our running example (the Pima dataset – C4.5), for each of the candidate subsets (and the baseline dataset) we obtain the *prf*, *prv*, and *num* ranks. To examine our evaluation function $E(S)$, we consider three cases. In each of these cases, we want to use $E(S)$ in order to obtain a single subset according to our preferences (applied via weights). In case 1 we assume that ($w1 = w2 = w3$). In case 2, we give performance twice the importance of privacy ($w1 = 2 * w2$), and in case 3, we give privacy three times of importance compared with performance ($w2 = 3 * w1$) . The $E(S)$ results associated with the candidate attribute subsets is shown in Table 8.6. We distinguish between two categories of subsets, those obtained based on *PerfR* (shown in the top bracket in Table 8.6) and those obtained based on *PrivR* (shown in the top bracket in Table 8.6).

Table 8.6: The E(S) results associated with candidate subsets corresponding to the selected weights (Pima dataset - C4.5).

| | Subset | E(S) (w1 = w2 = w3) | E(S) (w1 = 2w2) | E(S) (w2=3w1) |
|---|---|---|---|---|
| | **baseline** | 5.328 | 6 | 6 |
| *PerfR* | **plas_mass_preg_skin_pedi_age_pres** | 5.661 | 7.75 | 3.8 |
| | **plas_mass_preg_skin_pedi_age** | 6.327 | 8 | 4.6 |
| | **plas_mass_preg_skin_pedi** | 6.993 | 8.25 | 5.4 |
| | **plas_mass_preg_skin** | 7.326 | 7.75 | 6.4 |
| | **plas_mass_preg** | 7.992 | 8.25 | 6.8 |
| | **plas_mass** | 8.991 | 9.5 | 7 |
| *PrivR* | **plas** | 9.324 | 8 | 9.6 |
| | **insu_pres_age_pedi_skin_preg_plas** | 5.661 | 4.5 | 9 |
| | **insu_pres_age_pedi_skin_plas** | 6.327 | 6.5 | 7 |
| | **insu_pres_age_pedi_plas** | 6.993 | 6 | 9 |
| | **insu_pres_age_plas** | 7.659 | 6.25 | 9.8 |
| | **insu_pres_plas** | 8.325 | 7.75 | 8.6 |
| | **insu_plas** | 8.658 | 7.5 | 9.2 |

From Table 8.6 we notice that in general when performance is given higher importance compared with privacy, the subsets corresponding to *PerfR* result in higher $E(S)$. On the other hand, when privacy is given more importance compared with performance, subsets corresponding to *PrivR* have higher $E(S)$. This is in line with the observation that, that the subsets obtained based on *PerfR* ranking have higher accuracy compared with those which are based on *PrivR* ranking. On the other hand, the *PBI* of the corresponding subsets based on *PrivR* ranking is lower than the corresponding subsets that are obtained based on associated with *PerfR* ranking.

We plot the diagrams associated with subset/$E(S)$ combinations and show the results in Figure 8.10.

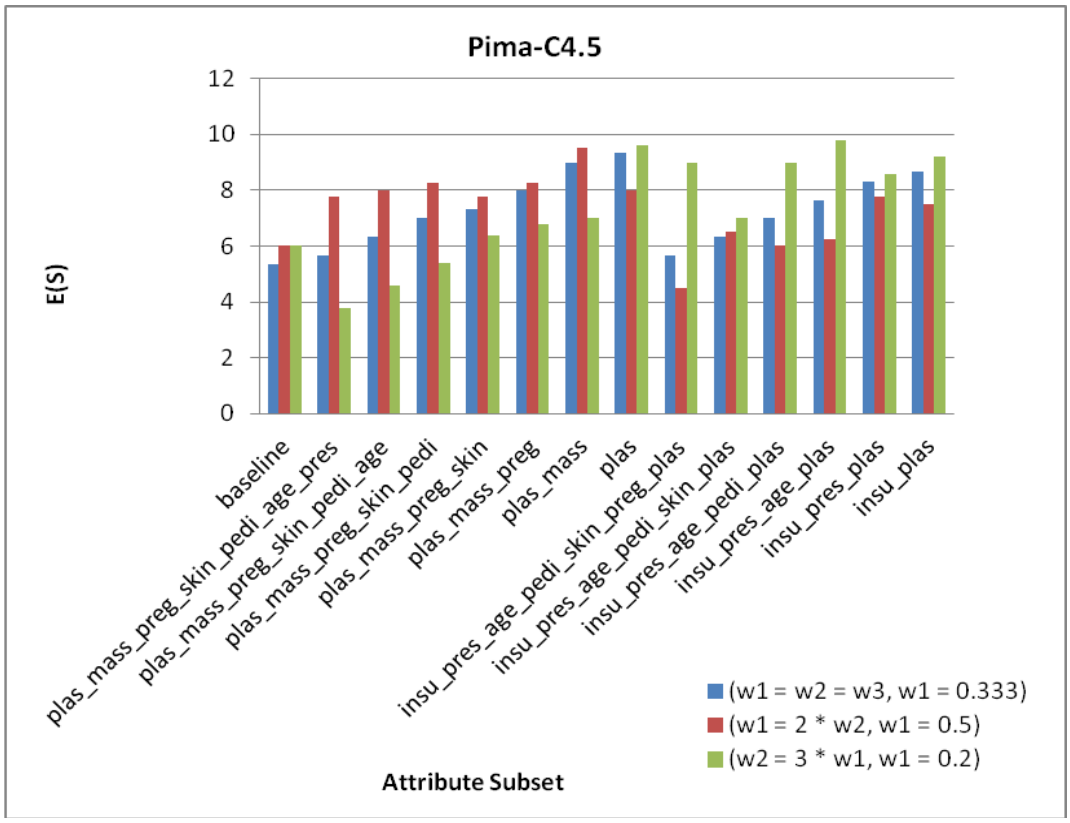


Figure 8.10: The E(S) corresponding to different weight ratios w.r.t. candidate subsets (Pima-C4.5).

From this diagram, when $w1 = w2 = w3$, subset {plas} is selected. When $w1 = 2 * w2$, and $w2 = 0.5$, subset {plas, mass} is selected, and when $w2 = 3 * w1$, and $w1 = 0.2$, subset {insu, pres, age, plas} is selected.

## 8.7. Experiments

We used four datasets from the UCI repository. Two relatively large datasets include the *Adult* dataset (consists of 45,222 records) and the *Diabetes* dataset (consists of 101,766 records). We also considered two smaller datasets; i.e., the *Pima* dataset and the *Liver Patient* dataset. The records with missing attribute values were eliminated. We obtained the results using two classification algorithms: namely, C4.5. and N.B. 10-fold cross validation was performed for evaluation and a statistical significant *t*-test was used to compare the results.

The corresponding results and analysis is summarized in the following two subsections.

### 8.7.1. Results and Analysis of the Two-dimensional PBI(DS) vs. Perf (DS) plot

For the case of C!=SA, a PBI(DS) vs. Perf(DS) plot is obtained and the four regions discussed in Section 8.3 are shown. We obtain the list of candidate attribute subsets and identify the location of each selected attribute subset within the PBI(DS) vs. Perf(DS) diagram.

The two-dimensional plot corresponding to the Pima dataset (C4.5 classifier) was shown and discussed in Section 8.4.3. We show, below in Figure 8.11, the results for the N.B. classifier.

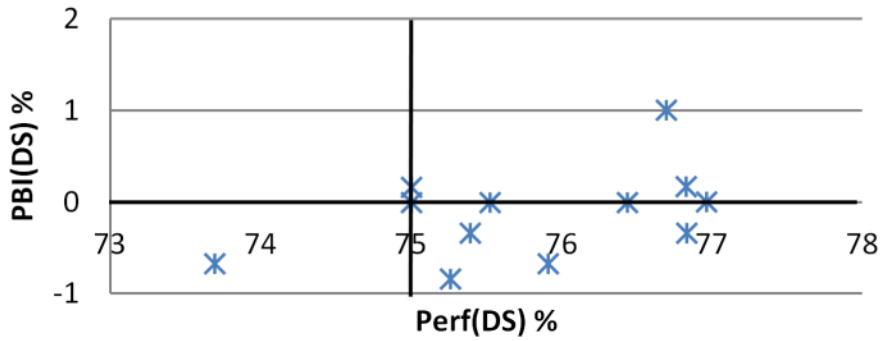



Figure 8.11: Pima dataset - N.B. - PBI(DS) vs. Perf(DS).

From this diagram, the candidate attribute subsets are spread over regions II, III, and IV. Ideally region IV is the best since it has a PBI and Perf values that are lower and higher than the baseline attribute set respectively. We notice that the density of distribution of the points on the plot in region IV is higher than other regions. This gives the data holder different options to select the final subset to be released. The spread of the points (each representing subset of attributes) is a function of the PBI and Perf associated with the given subset.

We conducted our experiments with the other datasets as well. Similar behaviour was observed; though, the distribution of the points was different for a variety of dataset/classifier combinations. These results, corresponding to C4.5 and N.B. classifiers are shown in Appendix A.

These graphs show a visual representation of PBI and Perf values associated with the selected subsets and enable DH to trade-off one for another when releasing a given dataset.

In general, it is possible to populate the plot with different subsets of attributes since each point on the plot represents a given subset. Region IV is the preferred region for the aforementioned reasons of having better utility and privacy compared with the baseline dataset. From the plots we notice that sometimes one or more regions are empty (i.e. no attribute subset point falls in that region) and this is related to the very characteristics of data and is dataset dependent. It is important to mention that, all of the subsets in the two-dimensional plot satisfy the minimum utility requirement. In other words, their accuracy is either higher than the baseline or if it is lower, this difference is not statistically significant. The *t*-test was used to validate this point. It is finally up to the data holder to choose appropriate $\alpha$ which represents the privacy breach increase/decrease factor, and identify the legitimate attribute subset from the list of available attribute subsets. It might be the case that DH provides a list of feature subsets (not the dataset) and the corresponding classification accuracy and communicate those to DR. DH only provides those attributes that satisfy the privacy requirements.

### 8.7.2. Results and Analysis of the Evaluation Function *E*(S)

For the case of C=SA, the evaluation function is implemented and the corresponding *E*(S) values for each combination of the weights are obtained. The case of Pima-C4.5 was discussed in detail earlier in Section 8.6. It was mentioned that, the proportion of weights specially between *w*1 and *w*2 is the choice of the data holder. Having a proportional relations between weights such as w1 = 4 * w2 , w2= 7 * w1 is one way of representing the relative weights. However, it is possible to consider any value for each of the weights as long as their sum is equal to 1.0. For instance the data holder might decide to consider 0.34, 0.42, and 0.24 for *w*1, *w*2, and *w*3 respectively.

Let us consider the results associated with Pima-N.B. case. After obtaining *PerfR* and *PrivR* and generating the candidate-subsets, we record the *Perf*(*S*), *PBI*(*S*), and *Num*(*S*) of each of these subsets and rank them based on the methodology discussed in Section 8.6. The final results are shown in Figure 8.12.

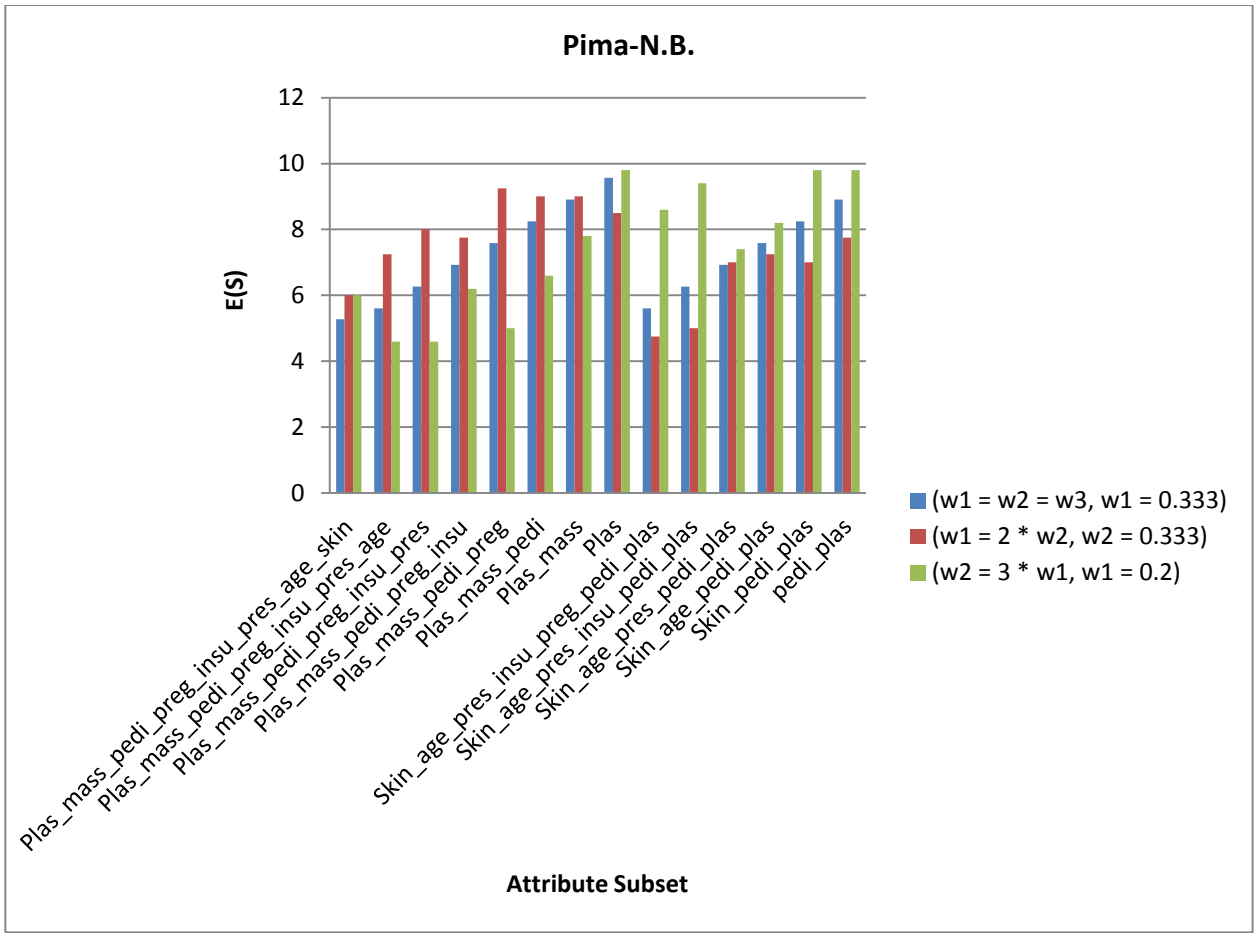



Figure 8.12: The E(S) corresponding to different weight ratios w.r.t. candidate subsets (Pima-N.B.).

From Figure 8.12, when $w1 = w2 = w3$, subset {plas} is selected. When $w1 = 2*w2$, and $w2 = 0.333$, subset {plas, mass, pedi, preg} is selected. When $w2 = 3*w1$, and $w1 = 0.2$, subset {plas} is selected. We could have selected subset {skin, pedi, plas} or {pedi, plas} since they have the same E(S) value of 9.8. However, in such cases where there is a tie (and more than one subset have the maximum $E(S)$ associated with them) , we may select the subset with least number of attributes i.e. {plas}.

We consider the *Adult* dataset next. In this case, different proportions of the weights are assumed. The corresponding results are shown in Figure 8.13.

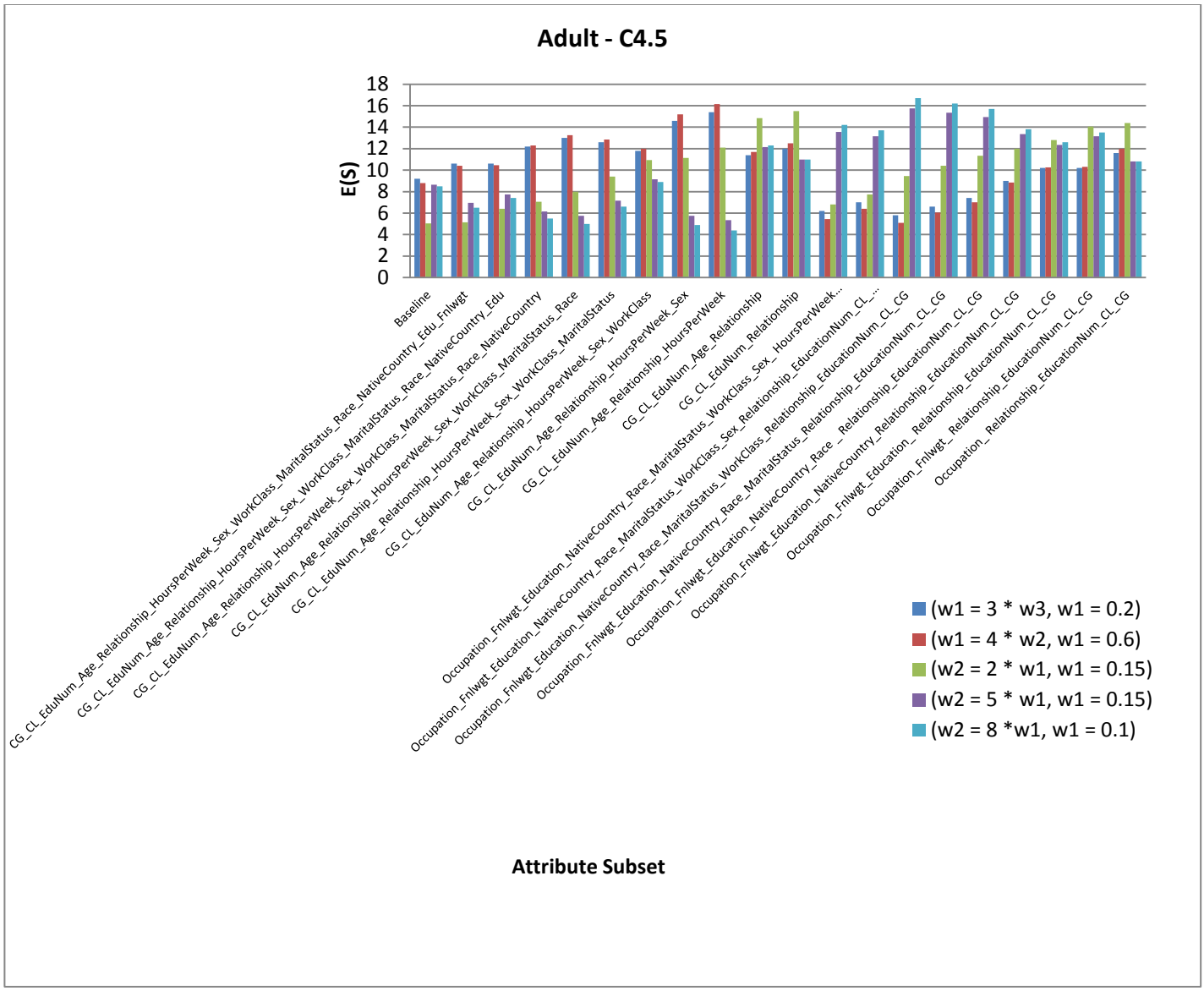



Figure 8.13: The E(S) corresponding to different weight ratios w.r.t. candidate subsets (Adult-C4.5).

Depending on the chosen weights for performance, privacy, and number of attributes, different values of $E(S)$ per candidate subsets are obtained. We record the $E(S)$ corresponding to **argmax**($E(S)$) | $w1$, $w2$, $w3$) associated with each case (i.e. selected weight ratio) and list the results along best subset in Table 8.7

Table 8.7: Best f(E(S)) and its corresponding selected attribute subset given the weight selected by the data holder corresponding to Adult-C4.5.

| Weights | | | Best Subset | E(S) |
|---|---|---|---|---|
| ***w*1** | ***w*2** | ***w*3** | | |
| 0.60 | 0.20 | 0.20 | **CG, CL, EduNum, Age, Relationship, HoursPerWeek** | 15.4 |
| 0.60 | 0.15 | 0.25 | **CG, CL, EduNum, Age, Relationship, HoursPerWeek** | 16.5 |
| 0.15 | 0.30 | 0.45 | **CG, CL, EduNum, Relationship** | 15.5 |
| 0.15 | 0.75 | 0.1 | **CG, CL, EduNum, Relationship, Workclass, MaritalStatus, Race, NativeCountry, Education, Fnlwgt, Occupation** | 16.7 |
| 0.1 | 0.8 | 0.1 | **CG, CL, EduNum, Relationship, Workclass, MaritalStatus, Race, NativeCountry, Education, Fnlwgt, Occupation** | 15.75 |

Following the same methodology, similar observations are made in the case of Adult-N.B. dataset/classifier combination. The results associated with different weights are shown in

Figure 8.14. Once again, it is possible to select an attribute subset with the best $E(S)$ according to a given selected weights i.e. to obtain **argmax**($E(S)$) | $w1$, $w2$, $w3$). For instance, consider the case where $w1$, $w2$, and $w3$ are 0.15, 0.3, and 0.55 respectively. In such case, the best $E(S)$ corresponds to the subset {CG, Relationship, MaritalStatus, Occupation, Education, EducationNum} which implies that given those weights it represents the best selected subset. Similarly, with the same weights the second best attribute subset would correspond to { CG, Relationship, MaritalStatus, Occupation, Education, EducationNum, Age}.

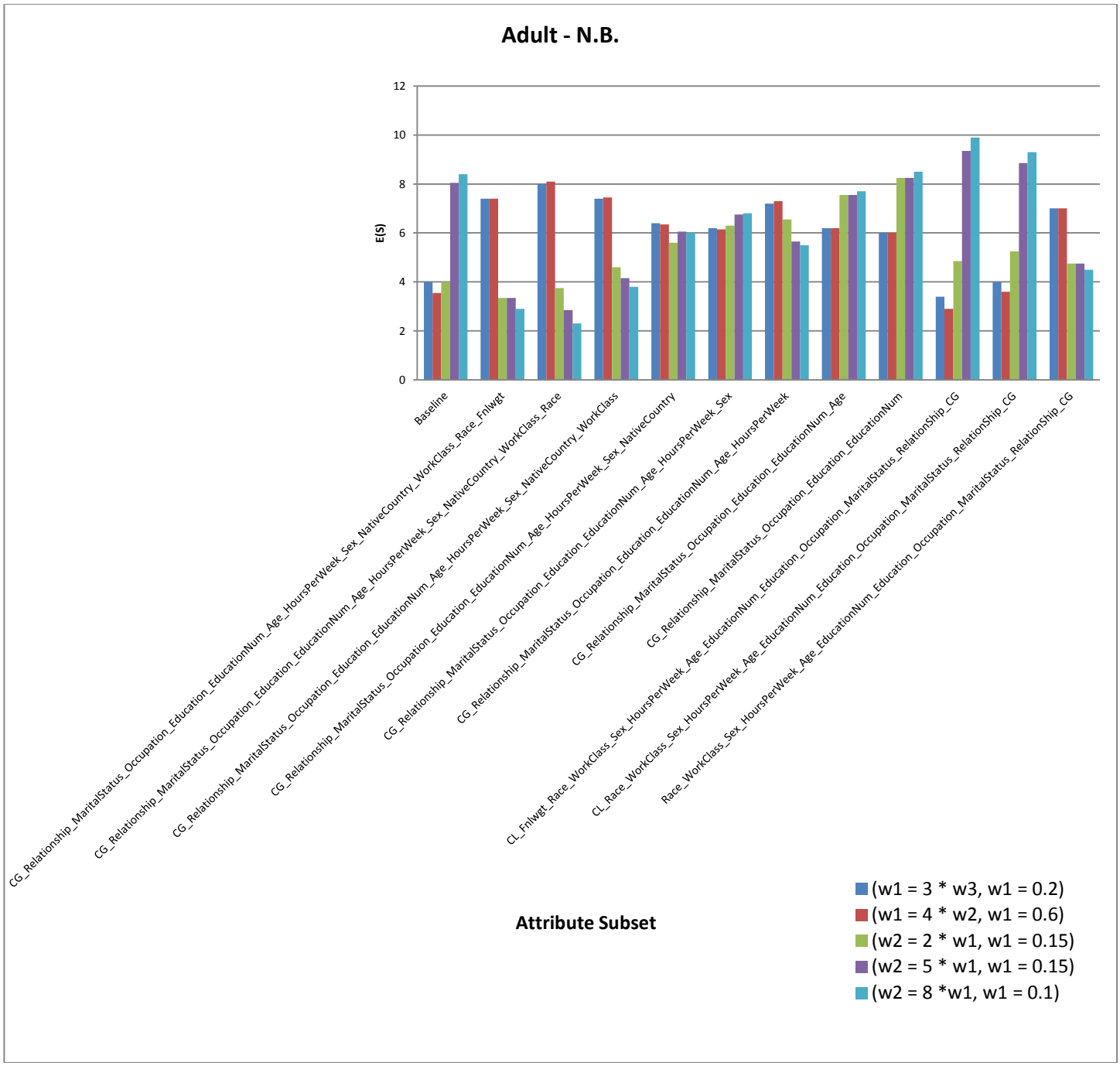



Figure 8.14: The E(S) corresponding to different weight ratios w.r.t. candidate subsets (Adult - N.B.).

We conducted further experiments with other datasets i.e. Diabetes and Liver Patients datasets (both case of N.B. and C4.5 classifiers). The results are shown in Appendix B.

In general, the results show that by introducing the $E$(S) evaluation function it is possible to identify a single attribute subset based on the data holder preferences. In each case, with different combinations of privacy, utility, and dimensionality weight it becomes possible to make decisions about the accepted/desired trade-off between these three factors and specially the privacy and utility factors. It should be outlined that, the proposed evaluation function is generalizable. That is, by setting any of the weights (or a combination of two of them) to zero, it is still possible to use the measure to evaluate the attribute subsets. All possible combinations are listed in Table 8.8. x, y, z are the weights chosen by the data holder.

Table 8.8: Possible combinations when one weight or two weights are set to zero.

| ***w*1** | ***w*2** | ***w*3** | ***E*(*S*)** |
|---|---|---|---|
| **0** | **0** | **1** | $\boldsymbol{w}3.num$ **(or *num*)** |
| **0** | **1** | **0** | $\boldsymbol{w}2.prv$ **(or *prv*)** |
| **1** | **0** | **0** | $\boldsymbol{w}1.prf$ **(or *prf*)** |
| **0** | **x** | **y** | $\boldsymbol{w}2.prv + w3.num$ |
| **x** | **0** | **y** | $\boldsymbol{w}1.prf + w3.num$ |
| **x** | **y** | **0** | $\boldsymbol{w}1.prf + w2.prv$ |
| **x** | **y** | **z** | $\boldsymbol{w}1.prf + w2.prv + w3.num$ |

From Table 8.8 we notice that, $E(S)$ = *prf* is a special case where the evaluation is based on the performance factor only. Such $E(S)$ (which dependents only on the performance (e.g. accuracy)), reminds us of the evaluation used in a regular non privacy-aware wrappers. In wrappers, the impact of addition/removing of a given attribute is evaluated by calculating the accuracy of the resulting attribute subset compared with accuracy of the preceding subset.

## 8.8. Discussions and Future Directions

In empirical studies, we employ *PBI* to ensure that the minimal set of the released attributes would not allow an attacker to increase the breach above $\alpha$. The ideal case is when the inference of SA is limited to the SA majority class distribution i.e. *DistP*. This guarantees that, while the released dataset maintain high utility for a given workload, it satisfies the privacy requirements.

One legitimate question is: what if the same data recipient requires the same dataset for another purpose, e.g. different target class. This results in releasing different sets of

attributes. Then, how we can ensure that multiple view publishing of the same dataset does not collectively breach the privacy requirement? In such case a new threshold can be introduced and imposed by the data holder $\beta$ which ensures that the PBI(DS) of the union of selected attributes in multiple releases is bounded by the $\beta$.

Another question is: what if the data recipient, given the same target class, wants to use a different classifier? The answer to this question is two-fold. Recall, that our feature selection goal is not to obtain the most optimal attribute subset. We may re-consider the previously obtained candidate privacy-aware attribute subsets. We re-use the attribute subsets that satisfy the performance and privacy constraints using the new classifier. We could also find the candidate privacy-aware attribute subsets using the new classifier and then only to choose/release from those that fall in the intersection of the candidate attribute subsets.

Building a classifier against the SA requires that the class attribute (being the SA attribute) to be discrete. There are, however, different cases where SA is a numerical attribute. A remedy to handle this case is to discretize the numerical attribute prior to classification.

There are cases where it is mandated that some attributes should be included in the released dataset (due to their scientific/financial importance). With the proposed framework this requirement could be addressed. By generating a candidate attribute subset to choose from, we could address cases where the data recipient wants to select which attributes should be included in the final attribute subset when more than one possibility exist. The list of attributes in the candidate attribute subsets could be communicated to the data recipient. Rather than aiming for a given attribute subset in region IV (Figure 8.2) with the least number of attributes, the data recipient might be interested in including some attributes in the released dataset. As long as the performance and privacy requirements are satisfied, this should be allowable.

The computational complexity of our algorithm is a factor of the number of features in the dataset. The fundamental difference between this privacy-aware measure and the non privacy-aware measure used in typical wrappers is that, contrary to the later, the goal of privacy-aware evaluation measure is not obtaining the optimal attribute subset w.r.t. the accuracy. This reduces the search effort. Moreover, the search in our technique is not

random, rather, it is guided by *PrivR* and *PerfR* and at each step one attribute is eliminated in a greedy (without backtracking) strategy which enhances the search performance.

One legitimate question is that, if we know the goal is to use a specific classifier, why don't we just publish the final classification model instead. In many cases, e.g. bioinformatics, datasets have too large a number of attributes that needs to be reduced. Furthermore, in areas such as medical research, it is a requirement that (even if the model is published) the data itself to be published as well. Moreover, even in the extreme case of publishing the model, we know that the model itself may be attacked. Therefore, a model that is built using privacy-aware feature selection as a pre-processing step, consists not only of fewer attributes (and all of the associated benefits w.r.t. performance), but those that were selected taking privacy considerations into account.

Another point is that, how to choose $\alpha$ such that it provides enough privacy. In selecting $\alpha$, the very characteristics of the dataset should be taken into consideration. In general, any value of $\alpha < 0$ (i.e. being in regions III and IV) indicates more privacy when compared with baseline's PBI i.e. 0. Controlling the amount of privacy is ultimately the choice of the data holder.

Given the desired weight of performance, privacy, and number of attributes, the best subset that results in highest $E(S)$ among candidate subsets is selected and returned by the privacy-aware feature selection system that employs the proposed evaluation function. We selected *PBI* as the privacy measure associated with each selected subset of attributes. However, the measure of privacy, in such a solution, could be any innate measure since our combinatory evaluation measure considers the rank of the privacy factor (*prv*) among possible candidates independent of the specific privacy measure that is used. The same argument is made about the performance measure. We consider accuracy in our experiments but it could be any other performance measure such as AUC, precision, etc. Once again, $E(S)$ is concerned about is the rank of the performance factor (*prf*) w.r.t. the other candidate subsets.

This evaluation function could be applied even in the cases where the original dataset is anonymized and when feature selection is applied on anonymized dataset.

Recall that $E(S)$ is not tied to a particular utility/privacy measure. It is mainly used for comparison purposes in order to compare candidate subsets and to select the one with

maximum value. We just mentioned that by setting the weights of one or two factors to zero, the measure still could be used to evaluate the subsets based on the remaining factor(s). The flexibility of the proposed measure also implies that, we could easily expand/extend $E(S)$ to include more factors if necessary.

In addition to obtaining the maximum $E(S)$ according to chosen weights, we can also get a ranked $E(S)$ for the same weight ratios. This, yet, gives the data holder more flexibility in trading-off performance, privacy, and dimensionality of the resulting dataset. For example, we the data holder will have the option of obtaining second best subset, third best subset, and so on.

In our current work, we apply $E(S)$ to the candidate attribute subsets in order to evaluate them and to obtain the subset with maximum $E(S)$ in a given setting. The $E(S)$ evaluation measure in its extended (complete) form i.e. $E(S)=w1.prf+w2.prv+w3.num$ could be plugged in into existing feature selection techniques. We will use $E(S)$ (on-the-fly) as the evaluation measure during forward selection or backward elimination in wrappers. In such case, rather than obtaining the complete list of candidate privacy-aware subsets, at each step, we can compare the associated{*prf*, *prv*, and *num*} rank of the updated subset (due to adding/removing of an attribute) with the {*prf*, *prv*, and *num*} rank of the preceding subset and maintain the subset with higher $E(S)$ given the weight. This logic will be applied to both cases of forward selection and backward elimination.

## 8.9. Summary

In this chapter, we proposed a multi-dimensional privacy-aware evaluation function in automatic feature selection. We defined a new attribute subset dependent measure of privacy, namely, Privacy Breach Increase (*PBI*). Using this measure, we proposed a system that generates candidate privacy-aware attribute subsets. The generated privacy-aware candidate subsets provide the data custodian with flexibility of choosing a subset that is most suitable for the existing scenario by adjusting the level of privacy and efficacy. We also proposed an evaluation function $E(S)$. Given the list of candidate privacy-aware subset, $E(S)$ enables the data custodian to determine the best privacy-aware subset according to

efficacy, privacy, and dimensionality preferences. The proposed techniques were experimentally validated using some of the UCI datasets.

# Chapter 9

# Conclusions and Future Work

In this work we aimed at studying a privacy dimension of automatic feature selection, a dimension which, to the best of our knowledge, has not been explored in the past. Combining privacy and automatic feature selection is rewarding specially in this era of abundance of data and rapidly expanding and increasing amount of collected digital information. It follows that, in many cases, such data consists of personal information and sensitive attributes about individuals, businesses, etc, that needs to be protected. Furthermore, the utility and usefulness of data should be retained for future analysis purposes.

We discussed the notion of privacy-by-design in Chapter 1. We particularly focused on two principles of the PbD, namely, "*privacy as the default setting*" and "*privacy embedded into design*". With this notion and its principles in mind, we aimed at turning feature selection into a privacy-aware process by proposing several techniques in the context of

task oriented privacy preserving data publishing. Incorporating privacy into automatic feature selection and (partially) realizing the notion of privacy-by-design is especially important due to the existing and ever increasing amount of data and the potential benefit of data reduction step within the KDD process.

Feature selection as an indispensable dimensionality reduction technique has several advantages associated with it. These advantages include, amongst others: enhancing understandability of data, lower computational cost, reducing negative impact of curse of dimensionality, and improving the predictive performance of learning algorithms. When privacy considerations is incorporated into feature selection process, yet, another important benefit is added to the above list.

Looking at it from a different perspective, experimental results have shown that feature selection, in general, leads to better utility of the resulting dataset specially on future and unseen data. It provides a degree of manoeuvrability which could be well used to the benefit of privacy preservation.

In our works we studied the relation between feature selection and privacy from different dimensions under the title of "Task Oriented Privacy-preserving Data Publishing Technologies Using Feature Selection".

We first considered the implicit role of automatic feature selection as a privacy preserving tool. In Chapter 5 we showed that, even without privacy-aware feature selection, it is possible that some or all of the identified quasi-identifier attributes get excluded from the final selected attribute subset. In other words, feature selection does not select them because they are either irrelevant or weakly relevant and redundant. By excluding some of the attributes in the QI set we implicitly achieve three goals. First, we only retain the attributes that are highly predictive of the target class. Second, by reducing the number of released QI attributes it becomes much more difficult to single out an individual relying only on the remaining QI attributes. Therefore, we reduce the risk of re-identification. Third, since anonymization targets the QI attributes only, more details of the relevant attributes are retained to achieve the same privacy level. This is an obvious benefit of special purpose anonymization vs. general purpose anonymization which results in over-anonymization of the dataset.

By investigating this dimension of feature selection we obtain resulting datasets that, while satisfying a given privacy model, retain high utility for further analysis. In our work we considered the analysis task to be classification. Our results presented in Chapter 5 showed that in both cases of *K*-anonymity and differential privacy (each belonging to different category of privacy models), using feature selection, we were able to satisfy the privacy requirements while maintaining a high level of utility compared with the baseline dataset.

We then attempted to turn automatic feature selection into a privacy-aware process. In other words, rather than passively investigating the possible privacy role of feature selection, we added privacy to feature selection to make it privacy-aware. We studied this dimension in Chapter 6 and Chapter 7 to address privacy-aware filters and wrappers respectively. Recall that filters and wrappers constitutes the two main branches of automatic feature selection. In either case, privacy was added as an extra step following the feature selection step. It was shown in both cases of wrappers and filters that using the very correlation among attributes and the correlation between attributes and the target class, it is possible to further eliminate and/or replace the quasi-identifiers with other non quasi-identifier attributes. In some cases, it was possible to eliminate all of the quasi-identifier attributes and yet to obtain a dataset with a classification accuracy not significantly different than the baseline dataset. We showed that how the utility gains associated with automatic feature selection could be used for privacy purposes.

Finally, we considered incorporating privacy during automatic feature selection. This is different than the second dimension in which privacy is added as an extra step following regular feature selection step . In this dimension, we considered the notion of empirical privacy based on the change in prior and posterior inference of sensitive attribute(s) by the attacker that has access to sophisticated data mining tools. This definition of privacy was modified to include the impact of eliminating of attributes on the increase or decrease of inference ability of the attacker. The main benefit of such definition of privacy is that it is does not require categorization of attributes into quasi-identifiers vs. non quasi-identifiers (though still actively used in the privacy-preserving data publishing/mining community and is required in different syntactic privacy models). We introduced the *PBI* privacy measure which enables the data holder to choose from a collection of candidate privacy-

aware attribute subset trading-off privacy and utility. We defined four regions of privacy vs. utility allowing the data holder to choose appropriate privacy-aware candidate attribute subset according to its privacy and utility preferences. We further expanded this dimension by introducing a multi-dimensional privacy-aware evaluation function in automatic feature selection which enables the data holder to select and release a subset of attributes according to the required preferences of utility (i.e., associated classification accuracy), privacy, and dimensionality of data. With the multi-dimensional privacy-aware evaluation function it is possible to reduce the list of privacy-aware candidate attribute subsets to a single subset.

## 9.1. Limitations

Some of the techniques presented in this work assume that the QI attribute sets are already identified. One limitation (which is beyond the scope of this work) is the challenge of choosing the QI attributes and was discussed in Section 2.2.4. Furthermore, in both cases of privacy-aware wrappers and filters, the assumption was made that the attributes that constitute the QI set are given same weight. This is an existing assumption in PPDP (and was inherited in the first two dimensions of our work). However, because we are using feature selection to eliminate some of the QI attributes we might give those attributes different privacy weights according to their associated privacy risk.

Another limitation, related to both dimensions of (privacy aware filters and wrappers) and (privacy-aware evaluation measure), is formulated in the following question: What if the same data recipient requires the same dataset for another purpose later and after obtaining the task-oriented dataset, e.g. having different target classes? This results in releasing different sets of attributes. Then, how we can ensure that multiple view publishing of the same dataset does not collectively breach the privacy?

These limitations can be addressed in the future work along other possibilities of extending the current work.

## 9.2. Future Work

To address the above mentioned limitations, we may include the following changes in the future work.

The fact that our approach considers the final analysis goal of the data recipient may be well used in order to prioritize the attributes in the QI set. In such case, we can improve the proposed algorithms and to enable the elimination of the QI attributes in a certain order taking into account both the privacy risk and the utility contribution of each of the QI attributes.

To address the potential privacy breach due to multi view data publishing, in the case of privacy-aware filters and wrapper, a possible future work is to implement a validation step at the DH's end (Section 2.6) in which the system would check if multiple views of the same data are requested by the DR. If that is the case, the union of all published anonymized views should still satisfy a given privacy model (i.e. be *K*-anonymous). To address the multi view data publishing issue corresponding to the privacy-aware evaluation measure, a possible future direction is to introduce and implement a new threshold $\beta$ at the DH's end. This new threshold ensures that the PBI of the union of (the projected datasets based on) selected attributes in multiple releases is bounded by the $\beta$. Additionally, we may consider sequential data release in which the number of attributes will change in future releases and some of those added/removed attributes happen to be quasi-identifiers. Other possibility is to generate synthetic data and to build models using that data. In general, multi-view data publishing and sequential data publishing, both, have been addressed in (Nergiz et al. 2009, Stokes and Torra 2012), (Yao et al. 2005), (Kifer and Gehrke 2006), (Wang and Fung 2006), (Fung et al. 2008), etc. As a future dimension, we can implement feature selection combined with those existing techniques and see how privacy-aware feature selection may improve the existing works in multi-view and sequential data publishing.

In addition to the aforementioned limitations and their possible remedies, our current work can be expanded in different directions. We already identified potential future directions for extending the three major dimensions proposed, discussed, and validated in the previous chapters. At the end of each chapter potential future directions were proposed.

Some other future improvements include the following: The wrapper feature selection approaches result in higher performance compared with filter approaches. This performance is achieved at higher computational cost. By including the privacy cost as another factor, filter approaches may provide very useful results. This is specially the case, when using wrapper approach is not applicable. It is possible to consider the trust level of the DR when publishing TOP anonymized dataset. We could also consider distributed task oriented privacy preserving data publishing where the data is horizontally or vertically distributed among different data holders.

We may consider generating task oriented synthetic data based on the statistical information of the original data.

Another dimension is to consider the stability of feature selection when it is used for privacy preserving purposes. Stability of feature selection refers to the case where when further records are included in the dataset or some records are removed from the dataset, feature selection would produce the same selected features. It is possible to differentiate between black-box vs. white-box classifiers and add another feature to the TOP framework by building and releasing black-box classifiers without scarifying privacy.

Another further extension of this work is studying the minimum amount of data that is sufficient build a data mining model. In other words, determining a threshold in which after that adding more data records would not impact the built models.

In our work we focused on the two commonly used and main approaches of automatic feature selection, namely, filters and wrappers. Another future dimension is to consider a third approach to feature selection known collectively as embedded methods. These methods learn which features contribute best to accuracy of the model during the creation of the model itself. An example of imbedded methods is LASSO (Least Absolute Shrinkage and Selection Operator) (Tibshirani 1996). A possible extension is to consider privacy-aware LASSO incorporating privacy considerations into its very functionality.

Finally, another future direction is to develop a multi-dimensional utility measure that will deliver its results in a multidimensional space (or at least 2-dimensional space, privacy and efficacy (accuracy) being the two dimensions). With such measure, it becomes possible to select the right point in this two dimensional space based on the Pareto efficiency. In other words, we develop a methodology to identify the best point (being the best subset of

attributes) in which no further improvement of privacy (or efficacy) is possible without harming the efficacy (or privacy) respectively. In a similar setting, such right point could be obtained based on Multi Criterion Decision-Making (MCDM) {Zeleny, 1973 #10}. MCDM refers to a class of methods based on the idea that, given a set of decision criteria and alternatives, what would be the best alternative.

# References


Acquisti, A. (2010). "The Economics of Personal Data and the Economics of Privacy-30 Years after the OECD Privacy Guidelines.", OECD Conference Center, Joint WPISP-WPIE Rountable, 1 December 2010.

Adam, N. R. and J. C. Worthmann (1989). "Security-control methods for statistical databases: a comparative study." ACM Comput. Surv. 21(4): 515-556.

Aggarwal, C. C. (2005). On k-anonymity and the curse of dimensionality. Proceedings of the 31st international conference on Very Large Data Bases. Trondheim, Norway, VLDB Endowment: 901-909.

Aggarwal, C. C., J. Pei and B. Zhang (2006). On privacy preservation against adversarial data mining. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA, ACM: 510-516.

Aggarwal, C. C. and P. S. Yu (2005). On Variable Constraints in Privacy Preserving Data Mining. Proceedings of the SIAM international conference on data mining. Newport Beach, CA, USA, SIAM: 115-125.

Aggarwal, C. C. and P. S. Yu (2008). An Introduction to Privacy-Preserving Data Mining. Privacy-Preserving Data Mining. C. Aggarwal and P. Yu, Springer US. 34: 1-9.

Agrawal, R., A. Evfimievski and R. Srikant (2003). Information sharing across private databases. Proceedings of the 2003 ACM SIGMOD international conference on management of data. San Diego, California, ACM: 86-97.

Agrawal, R., J. Kiernan, R. Srikant and Y. Xu (2002). Hippocratic databases. Proceedings of the 28th international conference on Very Large Data Bases. Hong Kong, China, VLDB Endowment: 143-154.

Amaldi, E. and V. Kann (1998). "On the approximability of minimizing nonzero variables or unsatisfied relations in linear systems." Theoretical Computer Science 209(1–2): 237-260.

Atzori, M., F. Bonchi, F. Giannotti and D. Pedreschi (2008). "Anonymity preserving pattern discovery." The VLDB Journal 17(4): 703-727.

Ayala-Rivera, V., P. McDonagh, T. Cerqueus and L. Murphy (2014). "A Systematic Comparison and Evaluation of k-Anonymization Algorithms for Practitioners." Transactions on Data Privacy 7(3): 337-370.

Bayardo, R. J. and R. Agrawal (2005). Data privacy through optimal k-anonymization. Proceedings of the 22nd International Conference on Data Engineering Tokyo, Japan: 217-228.

Bhumiratana, B. and M. Bishop (2009). Privacy aware data sharing: balancing the usability and privacy of datasets. Proceedings of the 2nd International Conference on PErvasive Technologies Related to Assistive Environments. Corfu, Greece, ACM: 1-8.

Bisiani, R. (1992). Beam Search. Encyclopedia of Artificial Intelligence, John Wiley and Sons. 2: 1467-1468.

Bolón-Canedo, V., N. Sánchez-Maroño and A. Alonso-Betanzos (2013). "A review of feature selection methods on synthetic data." Knowledge and Information Systems 34(3): 483-519.

Brand, R. (2002). Microdata Protection through Noise Addition. Inference Control in Statistical Databases, From Theory to Practice, Springer-Verlag: 97-116.

Brands, S. A. (2000). Rethinking public key infrastructures and digital certificates : building in privacy. Cambridge, Mass., MIT Press.

Brickell, J. and V. Shmatikov (2008). The cost of privacy: destruction of data-mining utility in anonymized data publishing. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. Las Vegas, Nevada, USA, ACM: 70-78.

Byun, J.-W., E. Bertino and N. Li (2005). Purpose based access control of complex data for privacy protection. Proceedings of the tenth ACM symposium on Access control models and technologies. Stockholm, Sweden, ACM: 102-110.

Cavoukian, A. (2009). "Privacy by design: The 7 foundational principles." Information and Privacy Commissioner of Ontario, Canada.

Chandrashekar, G. and F. Sahin (2014). "A survey on feature selection methods." Computers & Electrical Engineering 40(1): 16-28.

Ciriani, V., S. D. di Vimercati, S. Foresti and P. Samarati (2008). k-Anonymous Data Mining: A Survey. Privacy-Preserving Data Mining. C. Aggarwal and P. Yu, Springer US. 34: 105-136.

Ciriani, V., S. D. C. Di Vimercati, S. Foresti and P. Samarati (2007). Microdata protection. Secure data management in decentralized systems, Springer: 291-321.

Clarke, R. (2006). from https://www.priv.gc.ca/information/pub/guide_ind_e.asp.

Clifton, C. and T. Tassa (2013). "On syntactic anonymity and differential privacy." Transactions on Data Privacy 6(2): 161-183.

Cormode, G., C. M. Procopiuc, E. Shen, D. Srivastava and T. Yu (2013). Empirical privacy and empirical utility of anonymized data. IEEE 29th International Conference on Data Engineering Workshops (ICDEW), Brisbaane, QLD, 77-82.

Cox, L. H. (1980). "Suppression methodology and statistical disclosure control." Journal of the American Statistical Association 75(370): 377-385.

CRISP-DM (2001). "CRISP-DM." ftp://ftp.software.ibm.com/software/analytics/spss/support/Modeler/Documentation/14/UserManual/CRISP-DM.pdf.

Dalenius, T. (1977). Towards a methodology for statistical disclosure control. Statistik Tidskrift. 15: 429-444.

Dankar, F. K. and K. El Emam (2012). The application of differential privacy to health data. Proceedings of the 2012 Joint EDBT/ICDT Workshops. Berlin, Germany, ACM: 158-166.

DHS Information Sharing and Access Agreement (May 2009). H. Security, Department of Homeland Security.

Domingo-Ferrer, J. (2007). A Three-Dimensional Conceptual Framework for Database Privacy. Secure Data Management. W. Jonker and M. Petković, Springer Berlin Heidelberg. 4721: 193-202.

Domingo-Ferrer, J. (2008). A Survey of Inference Control Methods for Privacy-Preserving Data Mining. Privacy-Preserving Data Mining. C. Aggarwal and P. Yu, Springer US. 34: 53-80.

Dwork, C. (2006). Differential Privacy. Automata, Languages and Programming. M. Bugliesi, B. Preneel, V. Sassone and I. Wegener, Springer Berlin Heidelberg. 4052: 1-12.

Dwork, C. (2008). Differential Privacy: A Survey of Results. Theory and Applications of Models of Computation. M. Agrawal, D. Du, Z. Duan and A. Li, Springer Berlin Heidelberg. 4978: 1-19.

Dwork, C., F. McSherry, K. Nissim and A. Smith (2006). Calibrating Noise to Sensitivity in Private Data Analysis. Theory of Cryptography. S. Halevi and T. Rabin, Springer Berlin Heidelberg. 3876: 265-284.

El Emam, K., E. Jonker, L. Arbuckle and B. Malin (2011a). "A Systematic Review of Re-Identification Attacks on Health Data." PLoS ONE 6(12): e28071.

El Emam, K., J. Mercer, K. Moreau, I. Grava-Gubins, D. Buckeridge and E. Jonker (2011b). "Physician privacy concerns when disclosing patient data for public health purposes during a pandemic influenza outbreak." BMC Public Health 11(1): 1-16.

Fayyad, U., G. Piatetsky-Shapiro and P. Smyth (1996a). "The KDD process for extracting useful knowledge from volumes of data." Commun. ACM 39(11): 27-34.

Fayyad, U. M. and K. Irani (1993). Multi-interval discretization of continuous-valued attributes for classification learning. Proceedings of the 13th international joint conference on artificial intelligence. Chambery, France: 1022-1027.

Fayyad, U. M., G. Piatetsky-Shapiro and P. Smyth (1996b). From data mining to knowledge discovery: an overview. Advances in knowledge discovery and data mining. M. F. Usama, P.-S. Gregory, S. Padhraic and U. Ramasamy, American Association for Artificial Intelligence: 1-34.

Flach, P. (2012). Machine Learning The Art and Science of Algorithms that Make Sense of Data. Cambridge, Cambridge University Press.

Friedman, A., R. Wolff and A. Schuster (2008). "Providing k-anonymity in data mining." The VLDB Journal 17(4): 789-804.

Fung, B. C. M., K. Wang, R. Chen and P. S. Yu (2010a). "Privacy-preserving data publishing: A survey of recent developments." ACM Comput. Surv. 42(4): 1-53.

Fung, B. C. M., K. Wang, A. W. C. Fu and P. S. Yu (2010b). Introduction to Privacy-preserving Dta Publishing : Concepts and Techniques. Boca Raton, Chapman & Hall/CRC.

Fung, B. C. M., K. Wang and P. S. Yu (2005). Top-Down Specialization for Information and Privacy Preservation. Proceedings of the 21st International Conference on Data Engineering, IEEE Computer Society: 205-216.

Fung, B. C. M., K. Wang and P. S. Yu (2007). "Anonymizing Classification Data for Privacy Preservation." Knowledge and Data Engineering, IEEE Transactions on 19(5): 711-725.

Fung, B. C. M., K. Wang, A. W. C. Fu and J. Pei (2008). Anonymity for continuous data publishing. Proceedings of the 11th international conference on Extending database technology: Advances in database technology. Nantes, France, ACM: 264-275.

Gehrke, J. (2005). Models and methods for privacy-preserving data publishing and analysis: invited tutorial. Proceedings of the twenty-fourth ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems. Baltimore, Maryland, ACM: 316-316.

Gehrke, J., M. Hay, E. Lui and R. Pass (2012). Crowd-Blending Privacy. Advances in Cryptology – CRYPTO 2012. R. Safavi-Naini and R. Canetti, Springer Berlin Heidelberg. 7417: 479-496.

Gkoulalas-Divanis, A. and G. Loukides (2013). Anonymization of electronic medical records to support clinical analysis. New York, Springer.

Guyon, I. and A. Elisseeff (2003). "An introduction to variable and feature selection." J. Mach. Learn. Res. 3: 1157-1182.

Hall, M. A. (1999). Correlation-based feature selection for machine learning, The University of Waikato. PhD Dissertation.

Hall, M. A. and G. Holmes (2003). "Benchmarking Attribute Selection Techniques for Discrete Class Data Mining." IEEE Trans. on Knowl. and Data Eng. 15(6): 1437-1447.

Hall, M. A. and L. A. Smith (1998). "Practical feature subset selection for machine learning."

Han, J., M. Kamber and J. Pei (2012). Data mining concepts and techniques, third edition. The Morgan Kaufmann series in data management systems. Waltham, Mass., Morgan Kaufmann Publishers.

Health, N. I. o. (2003). NOT-OD-03-032: Final NIH Statement on Sharing Research Data, NIH Bethesda, MD, USA.HIPPA.

http://archive.ics.uci.edu/ml/. "UCI repository".

https://www.priv.gc.ca. "Office of the Privacy Commissioner of Canada."

Hu, J. (2011). Privacy-preserving Data Integration in Public Health Surveillance. Computer Science Ottawa, Ontario, Canada, Univeristy of Ottawa. PhD Dissertation.

Hua, M. and J. Pei (2008). A Survey of Utility-based Privacy-Preserving Data Transformation Methods. Privacy-Preserving Data Mining. C. Aggarwal and P. Yu, Springer US. 34: 207-237.

Inan, A., M. Kantarcioglu and E. Bertino (2009). Using Anonymized Data for Classification. Proceedings of the 2009 IEEE International Conference on Data Engineering, IEEE Computer Society: 429-440.

Iyengar, V. S. (2002). Transforming data to satisfy privacy constraints. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Edmonton, Alberta, Canada, ACM: 279-288.

Jafer, Y. (2014). Task Oriented Privacy (TOP) Technologies. Advances in Artificial Intelligence. M. Sokolova and P. van Beek, Springer International Publishing. 8436: 375-380.

Jafer, Y., S. Matwin and M. Sokolova (2014a). Privacy-aware filter-based feature selection. 2014 IEEE International Conference on Big Data (Big Data), Washignto, DC, 1-5.

Jafer, Y., S. Matwin and M. Sokolova (2014b). Task Oriented Privacy Preserving Data Publishing Using Feature Selection. Advances in Artificial Intelligence. M. Sokolova and P. van Beek, Springer International Publishing. 8436: 143-154.

Jafer, Y., S. Matwin and M. Sokolova (2014c). "Using Feature Selection to Improve the Utility of Differentially Private Data Publishing." Procedia Computer Science 37(0): 511-516.

Jafer, Y., S. Matwin and M. Sokolova (2015a). A framework for a privacy-aware feature selection evaluation measure. 13th Annual Conference on Privacy, Security and Trust (PST), Izmir, Turkey, 62-69.

Jafer, Y., S. Matwin and M. Sokolova (2015b). "A Multi-dimensional Privacy-aware Evaluation Function in Automatic Feature Selection." Submitted to the Transactions on Data Privacy (TDP) journal No. 20150915.

Jafer, Y., S. Matwin and M. Sokolova (2015c). Privacy-aware Wrappers. Advances in Artificial Intelligence. D. Barbosa and E. Milios, Springer International Publishing. 9091: 130-138.

Janecek, A., W. N. Gansterer, M. Demel and G. Ecker (2008). On the Relationship Between Feature Selection and Classification Accuracy. JMLR: Workshop and Conference Proceedings 4: 90-195.

Wang, J., L. Yongcheng, Z. Yan and L. Jiajin (2009). A Survey on Privacy Preserving Data Mining. First International Workshop on Database Technology and Applications, Whuan, Hebei, 111-114.

John, G., R. Kohavi and K. Pfleger (1994). Irrelevant Features and the Subset Selection Problem. Proceedings of the Eleventh International Conference on Machine Learning, Morgan Kaufmann Publishers, San Francisco, CA., 121-129.

Wang, K., Y. Xu, A. W. C. Fu and R. C. W. Wong (2009). FF-Anonymity: When Quasi-identifiers Are Missing. IEEE 25th International Conference on Data Engineering, 2009, Shanghai, 1136-1139.

Keng-Pei, L. and C. Ming-Syan (2011). "On the Design and Analysis of the Privacy-Preserving SVM Classifier." Knowledge and Data Engineering, IEEE Transactions on 23(11): 1704-1717.

Kifer, D. (2009). Attacks on privacy and deFinetti's theorem. Proceedings of the 2009 ACM SIGMOD International Conference on Management of data. Providence, Rhode Island, USA, ACM: 127-138.

Kifer, D. and J. Gehrke (2006). Injecting utility into anonymized datasets. Proceedings of the 2006 ACM SIGMOD international conference on Management of data. Chicago, IL, USA, ACM: 217-228.

Kira, K. and L. A. Rendell (1992). A practical approach to feature selection. Proceedings of the ninth international workshop on Machine learning. Aberdeen, Scotland, United Kingdom, Morgan Kaufmann Publishers Inc.: 249-256.

Kohavi, R. and G. H. John (1997). "Wrappers for feature subset selection." Artificial Intelligence 97(1–2): 273-324.

Kononenko, I. (1994). Estimating attributes: Analysis and extensions of RELIEF. Machine Learning: ECML-94. F. Bergadano and L. De Raedt, Springer Berlin Heidelberg. 784: 171-182.

Law, M. H. C., M. A. Figueiredo and A. K. Jain (2004). "Simultaneous feature selection and clustering using mixture models." Pattern Analysis and Machine Intelligence, IEEE Transactions on 26(9): 1154-1166.

LeFevre, K., D. DeWitt and R. Ramakrishnan (2006a). Workload-aware anonymization. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA, ACM: 277-286.

LeFevre, K., D. J. DeWitt and R. Ramakrishnan (2006b). Mondrian Multidimensional K-Anonymity. Proceedings of the 22nd International Conference on Data Engineering. Atlanta, GA, US: 25-25.

Li, N., T. Li and S. Venkatasubramanian (2007). t-Closeness: Privacy Beyond k-Anonymity and l-Diversity. proceedings of the twenty third international conference on data engineering Istanbul, Turkey: 106-115.

Li, N., W. H. Qardaji and D. Su (2011). "Provably private data anonymization: Or, k-anonymity meets differential privacy." CoRR, abs/1101.2604 49: 55.

Li, T. and N. Li (2009). On the tradeoff between privacy and utility in data publishing. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. Paris, France, ACM: 517-526.

Lin, P., I. S. O. Jun Zhang, H. Wang and J. Wang (2011a). A comparative study on data perturbation with feature selection. Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS), Hong Kong.

Lin, P., N. Thapa, I. S. Omer and L. Liu (2011b). "Feature Selection: A Preprocess for Data Perturbation." IAENG International Journal of Computer Science.

Loukides, G., A. Gkoulalas-Divanis and B. Malin (2011). "COAT: COnstraint-based anonymization of transactions." Knowledge and Information Systems 28(2): 251-282.

Machanavajjhala, A., J. Gehrke, D. Kifer and M. Venkitasubramaniam (2006). L-diversity: privacy beyond k-anonymity. Proceedings of the 22nd International Conference on Data Engineering. Atlanta, Georgia, US: 24-24.

Martin, B. (1995). Instance-based learning: nearest neighbour with generalisation, University of Waikato. Masters Thesis.

Maslow, A. H. (1943). "A Theory of Human Motivation." Psychological Review 50(4): 370-396.

Matwin, S. and T. Szapiro (2010). Data privacy: from technology to economics. Advances in Machine Learning II, Springer: 43-74.

Meyer, P. E., C. Schretter and G. Bontempi (2008). "Information-Theoretic Feature Selection in Microarray Data Using Variable Complementarity." Selected Topics in Signal Processing, IEEE Journal of 2(3): 261-274.

Microsoft (2012). Differential privacy for everyone. Microsoft.

Miller, A. J. (2002). Subset selection in regression. Boca Raton, Chapman & Hall/CRC.

Mohammed, N., R. Chen, B. C. M. Fung and P. S. Yu (2011). Differentially private data release for data mining. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. San Diego, California, USA, ACM: 493-501.

Mohammed, N., B. C. M. Fung, P. C. K. Hung and C.-k. Lee (2009). Anonymizing healthcare data: a case study on the blood transfusion service. Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining. Paris, France, ACM: 1285-1294.

Monreale, A. (2011). Privacy by design in data mining, Universita Delgli Studi Di Pisa. PhD Dissertation.

Motwani, R. and Y. Xu (2008). Efficient algorithms for masking and finding quasi-identifiers. Proceedings of SIAM international workshop on practical privacy-preserving data mining. Atlanta, US: 83-93.

Narayanan, A. and V. Shmatikov (2008). Robust De-anonymization of Large Sparse Datasets. Proceedings of the 2008 IEEE Symposium on Security and Privacy, IEEE Computer Society: 111-125.

Navarro-Arribas, G., D. Abril and V. Torra (2014). Dynamic Anonymous Index for Confidential Data. Data Privacy Management and Autonomous Spontaneous Security. J.

Garcia-Alfaro, G. Lioudakis, N. Cuppens-Boulahia, S. Foley and W. M. Fitzgerald, Springer Berlin Heidelberg. 8247: 362-368.

Navarro-Arribas, G., V. Torra (2015) Advanced Research in Data Privacy. Studies in Computational Intelligence,: IX, 463 p. 472 illus.

Nergiz, M. E., M. Atzori and C. Clifton (2007a). Hiding the presence of individuals from shared databases. Proceedings of the 2007 ACM SIGMOD international conference on Management of data. Beijing, China, ACM: 665-676.

Nergiz, M. E. and C. Clifton (2007). "Thoughts on k-anonymization." Data & Knowledge Engineering 63(3): 622-645.

Nergiz, M. E., C. Clifton and A. E. Nergiz (2009). "Multirelational k-Anonymity." Knowledge and Data Engineering, IEEE Transactions on 21(8): 1104-1117.

Nguyen, H. H., J. Kim and Y. Kim (2013). "Differential Privacy in Practice." Journal of Computing Science and Engineering 7(3): 177-186.

Office for Civil Rights, H. (2002). "Standards for privacy of individually identifiable health information. Final rule." Federal Register 67(157): 53181.

Pattuk, E., M. Kantarcioglu, H. Ulusoy and B. Malin (2015). Privacy-aware dynamic feature selection., 2015 IEEE 31st International Conference on Data Engineering (ICDE), Seoul, 78-88.

Pedreshi, D., S. Ruggieri and F. Turini (2008). Discrimination-aware data mining. Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining. Las Vegas, Nevada, USA, ACM: 560-568.

Powell, W. B. (2007). Approximate Dynamic Programming: Solving the curses of dimensionality, John Wiley & Sons.

Press, W. H. (1988). Numerical recipes in C : the art of scientific computing. Cambridge Cambridgeshire ; New York, Cambridge University Press.

Quinlan, R. J. (1993). C4.5: programs for machine learning, Morgan Kaufmann Publishers Inc.

Rich, E. and K. Knight (1991). Artificial intelligence. New York, McGraw-Hill.

Russell, S. J., P. Norvig and E. Davis (2010). Artificial intelligence : a modern approach. Upper Saddle River, NJ, Prentice Hall.

Samarati, P. (2001). "Protecting Respondents' Identities in Microdata Release." IEEE Trans. on Knowl. and Data Eng. 13(6): 1010-1027.

Samarati, P. and L. Sweeney (1998a). Generalizing data to provide anonymity when disclosing information (abstract). Proceedings of the seventeenth ACM SIGACT-SIGMOD-SIGART symposium on Principles of database systems. Seattle, Washington, USA, ACM: 188.

Samarati, P. and L. Sweeney (1998b). Protecting privacy when disclosing information: k-anonymity and its enforcement through generalization and suppression, Technical report, SRI International.

Sankar, L., S. R. Rajagopalan and H. V. Poor (2013). "Utility-privacy tradeoffs in databases: An information-theoretic approach." Information Forensics and Security, IEEE Transactions on 8(6): 838-852.

Schaffer, C. (1993). "Selecting a classification method by cross-validation." Machine Learning 13(1): 135-143.

Singla, A., E. Horvitz, E. Kamar and R. White (2014). "Stochastic Privacy." Association for the Advancement of Artififical Intelligence (AAAI).

Soria-Comas, J., J. Domingo-Ferrer, D. Sanchez and S. Martinez (2013). Improving the Utility of Differentially Private Data Releases via k-Anonymity. C. abs/1307.0966.

Sramka, M. (2010). "Data mining as a tool in privacy-preserving data publishing." Tatra Mountains Mathematical Publications 45(1): 151-159.

Sramka, M., R. Safavi-Naini, J. Denzinger and M. Askari (2010). A practice-oriented framework for measuring privacy and utility in data sanitization systems. Proceedings of the 2010 EDBT/ICDT Workshops. Lausanne, Switzerland, ACM: 1-10.

Stokes, K. and V. Torra (2012). "Multiple releases of k-anonymous data sets and k-anonymous relational databases." International Journal of Uncertainty, Fuzziness and Knowledge-Based Systems 20(06): 839-853.

Sun, X., H. Wang and J. Li (2009). Injecting purpose and trust into data anonymisation. Proceedings of the 18th ACM conference on Information and knowledge management. Hong Kong, China, ACM: 1541-1544.

Sweeney, L. (1998). Datafly: A System for Providing Anonymity in Medical Data. Proceedings of the IFIP TC11 WG11.3 Eleventh International Conference on Database Securty XI: Status and Prospects, Chapman & Hall, Ltd.: 356-381.

Sweeney, L. (2002a). "Achieving k-anonymity privacy protection using generalization and suppression." Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5): 571-588.

Sweeney, L. (2002b). "k-anonymity: a model for protecting privacy." Int. J. Uncertain. Fuzziness Knowl.-Based Syst. 10(5): 557-570.

Tan, P.-N., M. Steinbach and V. Kumar (2006). Introduction to data mining, Pearson Addison Wesley Boston.

Tibshirani, R. (1996). "Regression shrinkage and selection via the lasso." Journal of the Royal Statistical Society. Series B (Methodological): 267-288.

Tsai, J. Y., S. Egelman, L. Cranor and A. Acquisti (2011). "The effect of online privacy information on purchasing behavior: An experimental study." Information Systems Research 22(2): 254-268.

Vergara, J. and P. Estévez (2014). "A review of feature selection methods based on mutual information." Neural Computing and Applications 24(1): 175-186.

Verykios, V., K. Elmagarmid, E. Bertino, Y. Saygin and E. Dasseni (2004a). "Association rule hiding." Knowledge and Data Engineering, IEEE Transactions on 16(4): 434-447.

Verykios, V. S., E. Bertino, I. N. Fovino, L. P. Provenza, Y. Saygin and Y. Theodoridis (2004b). "State-of-the-art in privacy preserving data mining." SIGMOD Rec. 33(1): 50-57.

Wang, J. and J. Zhang (2012). Customizing Privacy Protection in Data Publishing. Proceedings of the 2nd second international conference on Business Computing and Global INformation (BCGIN), Shanghai, China, 596-602.

Wang, K. and B. C. M. Fung (2006). Anonymizing sequential releases. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA, ACM: 414-423.

Wang, K., B. C. M. Fung and P. S. Yu (2005). Template-Based Privacy Preservation in Classification Problems. Proceedings of the Fifth IEEE International Conference on Data Mining, IEEE Computer Society: 466-473.

Wang, K., P. S. Yu and S. Chakraborty (2004). Bottom-up generalization: a data mining solution to privacy protection. Proceedings of the fourth IEEE International Conference on Data Mining. Bringhton, UK: 249-256.

Watanabe, S. (1960). "Information Theoretical Analysis of Multivariate Correlation." IBM Journal of Research and Development 4(1): 66-82.

Wikipedia. from https://en.wikipedia.org/wiki/Maslow's_hierarchy_of_needs.

Willenborg, L. C. R. J. and T. d. Waal (2001). Elements of statistical disclosure control. New York, Springer.

Wilt, C. M., J. T. Thayer and W. Ruml (2010). A comparison of greedy search algorithms. Third Annual Symposium on Combinatorial Search.

Wolpert, D. H. (1992). "On the connection between in-sample testing and generalization error." Complex Systems 6(1): 47.

Xiao, X. and Y. Tao (2006). Personalized privacy preservation. Proceedings of the 2006 ACM SIGMOD international conference on Management of data. Chicago, IL, USA, ACM: 229-240.

Xiong, L. and K. Rangachari (2008). Towards Application-Oriented Data Anonymization. proceedings of the first SIAM international workshiop on practical privacy-presering data mining. Atlanta, US: 1-10.

Xu, J., W. Wang, J. Pei, X. Wang, B. Shi and A. W. Fu (2006). Utility-based anonymization using local recoding. Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining. Philadelphia, PA, USA, ACM: 785-790.

Yao, C., S. X. Wang and S. Jajodia (2005). Checking for k-anonymity violation by views. Proceedings of the 31st international conference on Very large data bases. Trondheim, Norway, VLDB Endowment: 910-921.

Yu, L. and H. Liu (2004). "Efficient Feature Selection via Analysis of Relevance and Redundancy." J. Mach. Learn. Res. 5: 1205-1224.

Zang, H. and J. Bolot (2011). Anonymization of location data does not work: a large-scale measurement study. Proceedings of the 17th annual international conference on Mobile computing and networking. Las Vegas, Nevada, USA, ACM: 145-156.

Zhang, J. and V. Honavar (2003). Learning decision tree classifiers from attribute value taxonomies and partially specified data. Proceedings of the twentieth international conference on machine learning. Washington DC. US: 880-887.

Zhang, J., D.-K. Kang, A. Silvescu and V. Honavar (2006). "Learning accurate and concise na\&\#239;ve Bayes classifiers from attribute value taxonomies and data." Knowl. Inf. Syst. 9(2): 157-179.

Zhang, N. and W. Zhao (2007). "Privacy-preserving data mining systems." Computer 40(4): 52-58.

# Appendix A

## A1. The PBI(DS) vs. Perf(DS) Plot Results

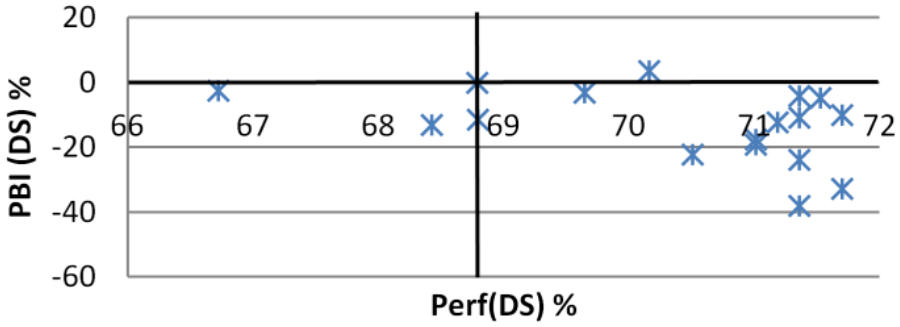



**Figure A.1: Liver Patients dataset - C4.5. - PBI(DS) vs. Perf(DS).**

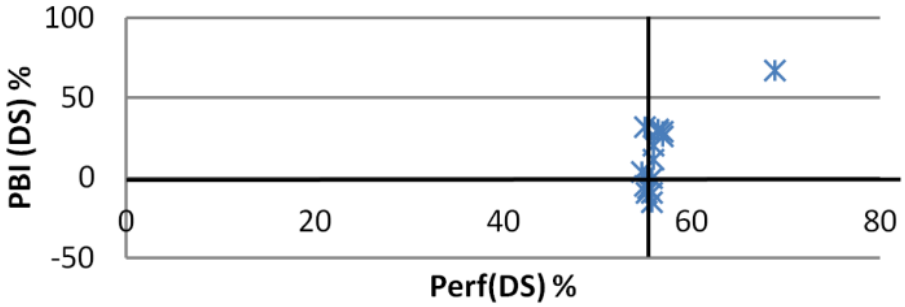



**Figure A.2: Liver Patients dataset - N.B. - PBI(DS) vs. Perf(DS).**

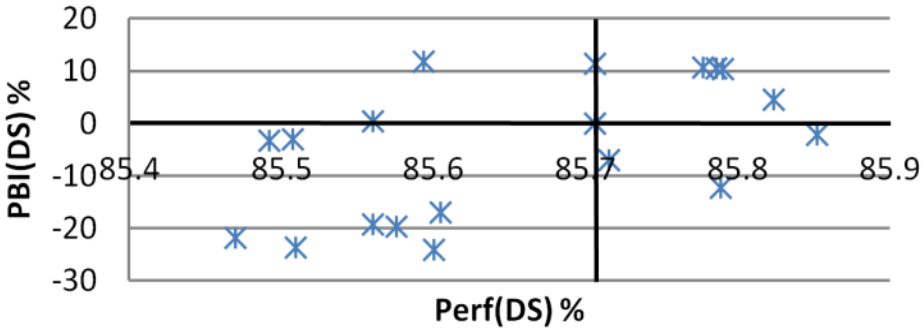



**Figure A.3: Adult dataset – C4.5. - PBI(DS) vs. Perf(DS).**

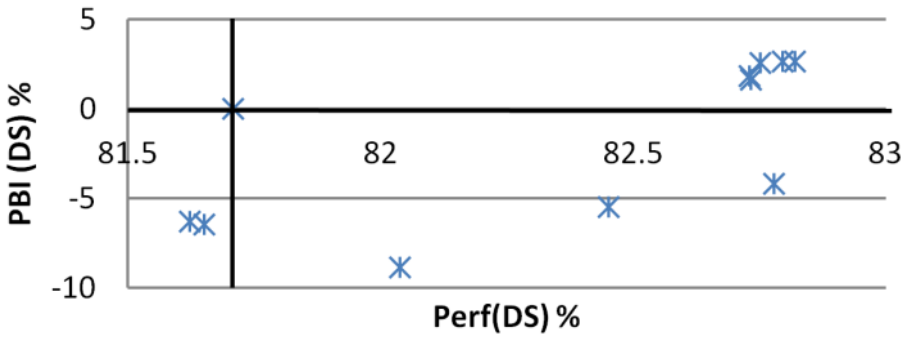



**Figure A.4: Adult dataset – N.B. - PBI(DS) vs. Perf(DS).**

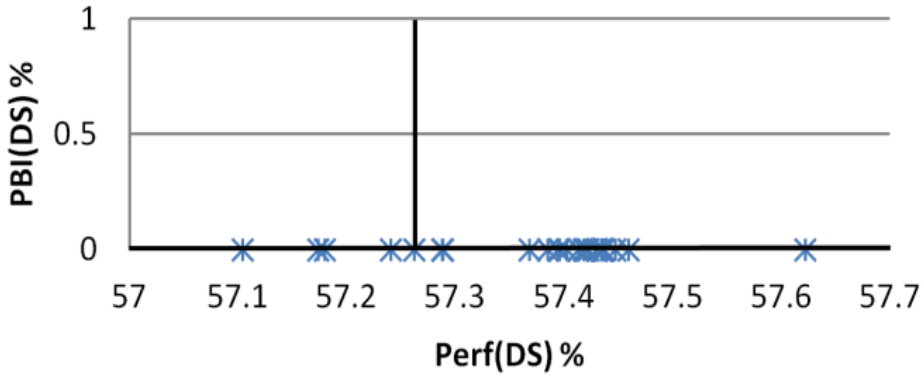



**Figure A.5: Diabetes dataset - C4.5 - PBI(DS) vs. Perf(DS).**

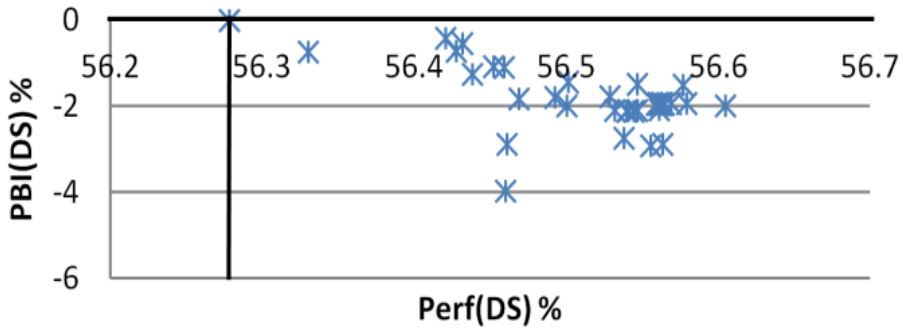



**Figure A.6: Diabetes dataset - N.B. - PBI(DS) vs. Perf(DS).**

# Appendix B

## B1. The Evaluation Function *E*(S) Results

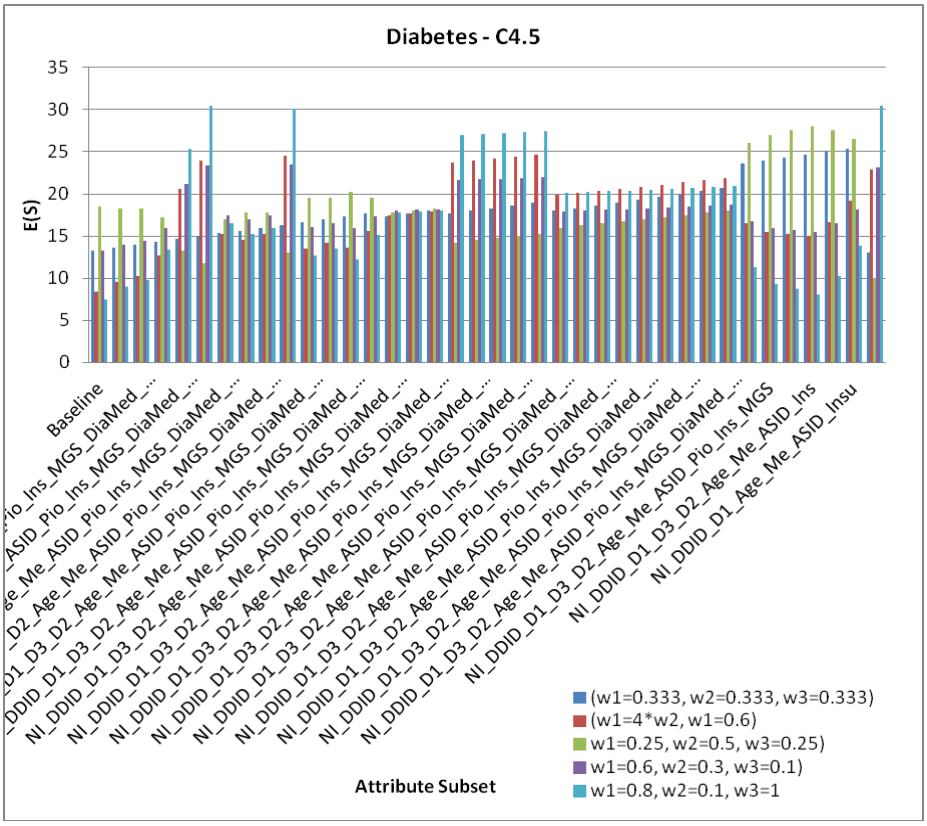



**Figure B.1: The E(S) corresponding to different weight ratios w.r.t. candidate subsets (Diabetes - C4.5).**

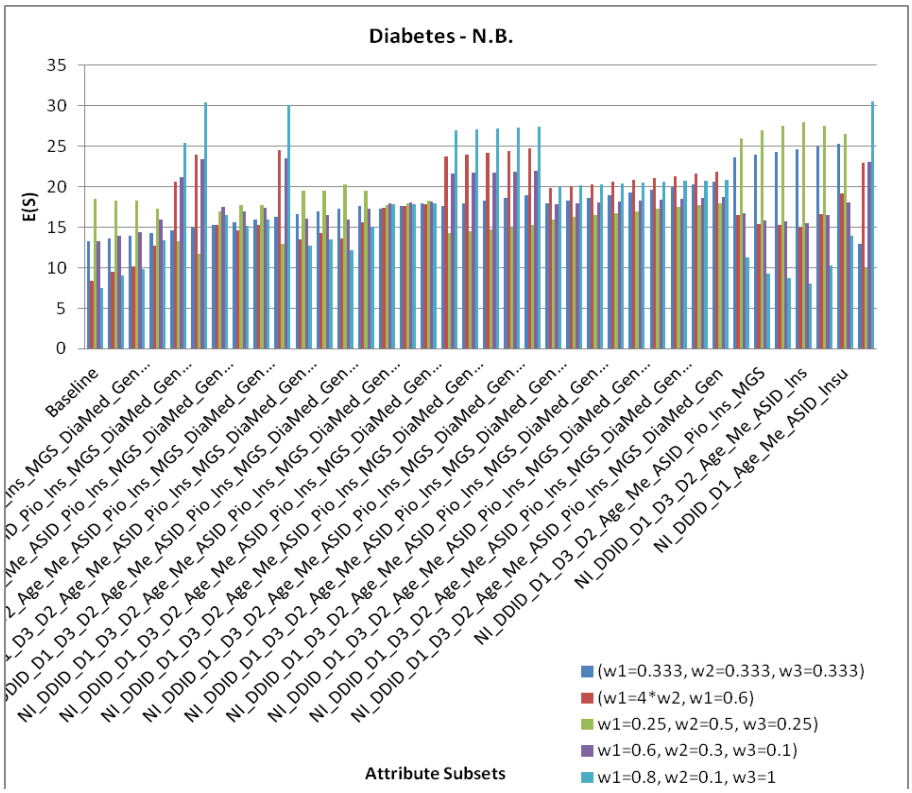



**Figure B.2:The E(S) corresponding to different weight ratios w.r.t. candidate subsets (Diabetes - N.B.).**

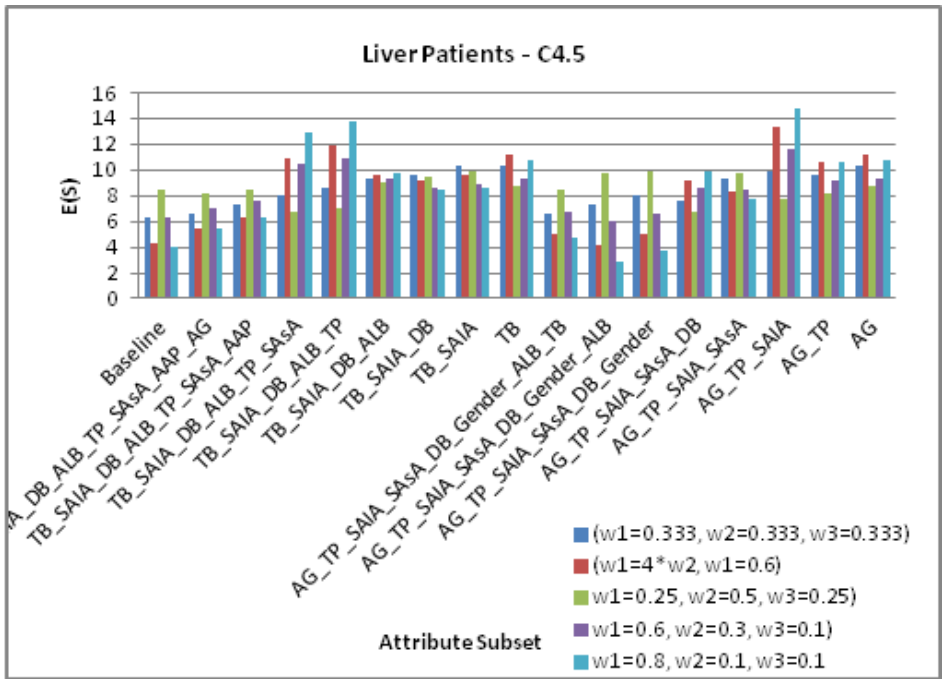



**Figure B.3: The E(S) corresponding to different weight ratios w.r.t. candidate subsets (Liver Patients - C4.5).**

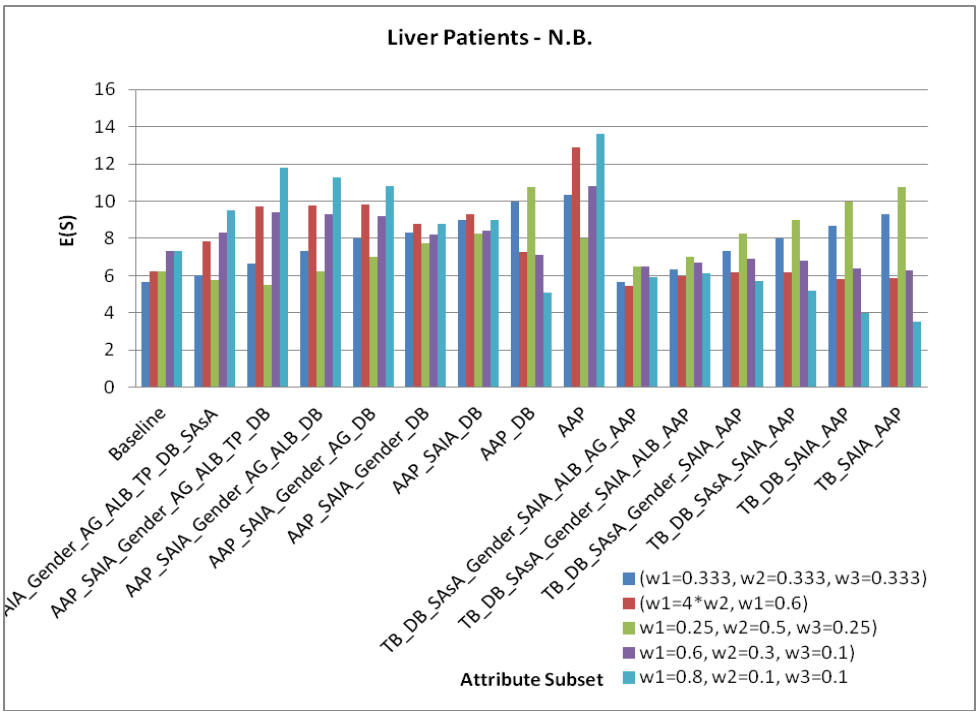



**Figure B.4: The E(S) corresponding to different weight ratios w.r.t. candidate subsets (Liver Patients - N.B.).**

# Appendix C

**Table C.1: Description of the datasets.**

| Dataset | Number of instances | Number of attributes | List of attributes |
|---|---|---|---|
| heart stat logs | 270 | 14 | **age, sex, chest**, **resting_blood_pressure**, **serum_cholestoral**, **fasting_blood_sugar**, **resting_electrocardiographic_results**, **maximum_heart_rate_achieved**, **exercise_induced_angina**, **oldpeak**, **slope**, **number_of_major_vessels**, **thal**, **class** { absent, present} |
| pima diabetes | 768 | 9 | **preg**, **plas**, **pres, skin**, **insu**, **mass**, **pedi**, **age class** {tested_negative, tested_positive} |
| german credit | 1000 | 21 | **checking_status, duration, credit_history, purpose, credit_amount, saving_status, employment, installment_commitment, personal_status, other_parties, residence_since, property_magnitude, age, other_payment_plans, housing, existing_credits, job, num_dependents, own_telepone, foreign_worker, class** {good, bad} |
| liver_patients | 583 | 11 | **age, gender, TB, DB, AAP, SAIA, SAsA, TP, ALB, A/G, class** {A, B} |
| CRX | 653 | 16 | **A1, A2, A3, A4, A5, A6, A7, A8, A9, A10, A11, A12, A13, A14, A15, class** {+,-} |
| CMC | 1,473 | 10 | **wife_age, wife_education, husband_education, number_of_children_ever_born, wife_religion, wife_now_working, husband_occupation, standard-of-living-index, media_exposure, contraceptive_method_used** {1,2,3} |
| Winconsin Breast Cancer | 683 | 10 | **clump_thickness, cell_size_uniformity, cell_shape_uniformity, marginal_adhesion, single_epi_cell_size, bare_nuclei, bland_chromatin, normal_nucleoli, mitoses, class** {benign, malignant} |
| Adult | 45,222 | 15 | **age, workclass, fnlwgt, education, education_num, marital_status, occupation, relationship, race, sex, capital_gain, capital_loss, hours_per_week, native_country, class{**>50k, <=50k**}** |
| Diabetes (130-US hospitals for years 1999-2008) | 101,766 | 55 | **encounter_id, patient_nbr,race, gender, age, weight, admission_type_id , discharge_disposition_id, admission_source_id, time_in_hospital, payer_code, medical_specialty, num_lab_procedures, num_procedures, num_medications, number_outpatient, number_emergency, number_inpatient, diag_1, diag_2, diag_3, number_diagnoses, max_glu_serum, A1Cresult, metformin, repaglinide, nateglinide, chlorpropamide, glimepiride, acetohexamide, glipizide, glyburide, tolbutamide, pioglitazone, rosiglitazone, acarbose, miglitol, troglitazone, tolazamide, examide, citoglipton, insulin, glyburide-metformin, glipizide-metformin, glimepiride-pioglitazone, metformin-rosiglitazone, metformin-pioglitazone, change, diabetesMed, readmitted** {Yes, No} |