uOttawa

L'Université canadienne
Canada's university

FACULTÉ DES ÉTUDES SUPÉRIEURES ET POSTOCTORALES

u Ottawa

L'Université canadienne
Canada's university

FACULTY OF GRADUATE AND POSDOCTORAL STUDIES

Isis Peña

AUTEUR DE LA THÈSE / AUTHOR OF THESIS

M.C.S.

GRADE / DEGREE

School of Information Technology and Engineering

FACULTÉ, ÉCOLE, DÉPARTEMENT / FACULTY, SCHOOL, DEPARTMENT

Utility-based Data Mining: An Anthropometric Case Study

TITRE DE LA THÈSE / TITLE OF THESIS

Herna Viktor

DIRECTEUR (DIRECTRICE) DE LA THÈSE / THESIS SUPERVISOR

Eric Paquet

CO-DIRECTEUR (CO-DIRECTRICE) DE LA THÈSE / THESIS CO-SUPERVISOR

EXAMINATEURS (EXAMINATRICES) DE LA THÈSE / THESIS EXAMINERS

Liam Peyton

Michael Weiss

Gary W. Slater

Le Doyen de la Faculté des études supérieures et postdoctorales / Dean of the Faculty of Graduate and Postdoctoral Studies

Utility-based Data Mining: An Anthropometric Case Study

by

Isis Peña

Thesis submitted to the
Faculty of Graduate and Postdoctoral Studies
In partial fulfillment of the requirements
For the MCS degree in
Ottawa-Carleton Institute for Computer Science

School of Information Technology and Engineering
Faculty of Engineering
University of Ottawa

Your file Votre référence
ISBN: 978-0-494-48500-2
Our file Notre référence
ISBN: 978-0-494-48500-2

Abstract

One of the most important challenges for the apparel industry is to produce garments that fit the population properly. In order to achieve this objective, it is crucial to understand the typical profile of consumer's bodies. In this work, we aim to identify the typical consumer from the virtual tailor's perspective. To this end, we perform clustering analysis on anthropometric and 3-D data to group the population into clothing sizes. Next, we perform multi-view relational classification to analyze the interplay of different body measurements within each size. In this study, we analyze three different populations as contained in the CAESAR^{TM} database, namely, the American, the Italian and the Dutch populations.

Throughout this study, we follow a utility-based data mining approach. The goal of utility-base data mining is to consider all utility aspects of the mining process and to thus maximize the utility of the entire process. In order to address this issue, we engage in dimension reduction techniques to find a smaller set of body measurement that reduces the cost and improves the performance of the mining process. We also apply objective interestingness measures in our analysis of demographic data, to improve the quality of the results and reduce the time and search space of the mining process. The analysis of demographic data allows us to better understand the demographic nature of potential customers, in order to target subgroups of potential customers better.

Acknowledgements

I would like to express my sincere gratitude to my supervisor, Dr. Herna Viktor for her excellent supervision, support and valuable guidance throughout my graduate studies, and during the completion of this thesis. Her inspiring discussion and insightful feedback have made positive contributions to every chapter of this thesis.

I would like to express my deep appreciation to my co-supervisor, Dr. Eric Paquet whose expertise, fruitful discussions and constructive comments were of great assistance in the preparation of this thesis. I appreciate his detailed reviews, feedback and suggestions on every chapter of this thesis.

I would like to thank all members in the IDeAL group at the University of Ottawa, in particular Divine Muhivu, Nadia Azam and Hongyu Guo, also to Pauline Anthonysamy for their constructive discussions and friendship.

I would also like to thank my mother for the support and trust she always provide me on every project I engaged on. Thanks to my sister for being the best sister I could possible have and her contagious optimism. Special thanks to Mauricio for his proof reading assistance, support, encouragement and unconditional love through all this time.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Rapid advances in data collection and storage technology have enabled organizations to accumulate huge amounts of data. However, extracting useful information from these data is a challenging task. Traditional data analysis tools and techniques rely on manual analysis and interpretation. This process is prone to biases and errors, and is extremely time consuming [Han and Kamber, 2006]. Data mining then grows out of this gap between data and information. Data mining is the process of automatic discovery of previously unknown, useful, novel and non-trivial information from large data repositories [Han and Kamber, 2006, Tan et al., 2005, Witten and Frank, 2005].

Data mining has been successfully applied in many fields such as health insurance, to predict health outcomes and risk factors of potential clients [Chae et al., 2001]. In production schedule, to extract scheduling knowledge about due date assignments in shop floor control, and reduce the due date predictions errors [Sha and Liu, 2005]. In biomedical technology, to do classification of biological sequences such as deoxyribonucleic acids (DNA) and proteins [Maddouri and Elloumi, 2002]. Also, in investment risk prediction [Becerra-Fernandez et al., 2002], to classify a country's investing risk based on a variety of factors, amongst others. Nevertheless, little research has been done on the application of data mining techniques for the manufacture of garments. In this work we apply data mining techniques on anthropometric data with the aim of facilitating the design and tailoring of garments. Moreover, we apply utility-based data mining techniques, which account for economic aspects in the mining process, to improve the quality of the results and reduce the cost of the mining process.

In this chapter we present the motivation and goals of our study. We also present a high level overview of the steps performed to achieve the proposed goals.

1.1 Motivation

Apparel manufacturers develop sizing systems with the goal of satisfying consumer's needs for apparel that fits. Sizing is the process used to establish a size chart of key body measurements for a range of apparel sizes. To produce garments that fit the population, the sizes must correspond to real groupings within the population. However, Shofield et al. [Schofield and LaBat, 2005] collected forty size charts for women's clothing from retailers, dressmaking texts, published sizing standards and apparel companies. After analyzing them, they found that three common features are present in all size charts:

1. The different sizes are defined using arbitrary constant intervals between sizes for the primary measurement. For example, Aldrich [Aldrich, 2006] presents standard body measurements for mature males where the small size is defined as having 88 cm of chest circumference, the medium size 92 cm, the large size 96 cm, the x-large size 100 cm, and so on. There is then a constant increment of 4 cm from one size to another on the primary measurement.

2. All vertical measurements increase as the size increases. That is, there is a fixed relationship between length and girth measurements based on the assumption that the larger a person is the taller he or she is, and vice versa. People whose height does not match their size according to these categories, would have to chose between fitting their length or their girth measurements.

3. The differences between the principal girths (hip circumference minus bust circumference and bust circumference minus waist circumference) are constant for all sizes. The size charts are then designed to target a fixed body shape.

Considering the aforesaid situation, it is easily understandable why repeated studies of degree of satisfaction with apparel in the United States found that about 50% of women and 62% of men cannot find satisfactorily fitting clothes [Ashdown and Dunne, 2006]. These numbers increase with segments of the population that traditionally have not been part of the target markets for garment manufacture. For example, 69% of women aged 55 and above report poor or lack of fit in clothes, and 77% mention needing alterations of apparel and patterns [Salusso et al., 2006]. Moreover, billions of dollars of lost revenue are reported by the apparel industry due to returned garments or garments that are never sold [Ashdown and Dunne, 2006].

In the process of manufacturing garments [Schofield and LaBat, 2005], grading rules (applying increases and decreases at points of a clothing pattern to make the pattern larger or smaller) are used to create the garments. The grading rules should be based on body measurements associated with a specific size. In turn, these body measurements in size charts should be based on anthropometric data (a collection of body measurements that describe the human body).

According to Ashdown et al. [Ashdown and Loker, 2005] two main issues have limited the aforementioned process and the ability of the apparel companies to produce garments with quality fit. First, there has been a lack of up-to-date anthropometric data to describe the civilian population. Over the past years, there have been many comprehensive anthropometric surveys of military populations [Gordon et al., 1989]. Military personnel, however, are not representative of the civilian population since they have to meet strict fitness criteria and tend to be younger than the general civilian population. Second, there is a lack of information about the principal aspects to consider when designing garments for a variety of body sizes and shapes [Ashdown and Loker, 2005]. Marketing of apparel, typically consider the age, income and lifestyle choices of the target market, which are not necessarily a predictor of size and body shape.

Recent work has addressed, to some extent, these problems. Many anthropometric surveys such as the *CAESAR™ project* [Robinette et al., 2002], *Size USA* [siz, 2003], *Size China* [siz, 2007] and *Size UK* [siz, 2001] have been carried out on civilian populations. These surveys include several body measurements such as acromial height, waist circumference, hip circumference, height, weight, etc. of thousands of people from different countries and ethnic groups. What is more, these surveys not only include several body measurements, but the 3-D body scans of each participant. That is, a laser scanning device measures and records a detailed geometry of the participant's body surface.

Also, some recent research attempt to find the most important factors or measurements to be taken into account when designing garments. Veitch et al. [Veitch et al., 2007] aim to produce a well fitting bodice (the upper part of the dress). The study considers measures taken on 1,265 Australian women of age range 18-70+ years. First, twelve out of fifty-four measures were selected and ranked by an apparel pattern maker. Next, principal components analysis (PCA) was applied to condense these twelve anthropometric measures into the first two principal components. Based on the results, they defined thirty-six categories: twelve sizes and three body shapes within each size.

Hsu et al. [Hsu et al., 2007] identified three body types and thirty-eight sizes for the female adult Taiwanese population. They selected eleven anthropometric measures

(seven linear and four girth measures) and after PCA, two principal components were kept. They identified the sizes by applying hierarchical and non-hierarchical clustering methods on the data obtained from PCA. That is, they integrated k-means with Ward's minimum variance algorithm in a two-step clustering.

Finally, Viktor et al. [Viktor et al., 2006] attempted to find body size groupings in the American male population as contained in the CAESAR^{TM} database. They identified five groups that correspond to Small - XXLarge sizes. From these results, they derived rules that may be used as constraints when designing garments for the different body sizes.

Although the abovementioned work attempt to address the problem of identifying the main aspects or measurements that should be consider for the design of garments, they focus only either a specific body part or on a reduced population (males or females). Moreover, they do not account for the economic factors of the process.

1.2 Thesis goals

As mentioned previously, in this work we focus on applying data mining techniques with the aim of facilitating the garment design and manufacture. More specifically, focusing on anthropometric and demographic data, we pursue the following goals:

First, we aim to understand the typical consumers' body profile by identifying the natural body size groups and their distinctive characteristics. Second, we attempt to find the most important body measurements that define each size, and how these measurements interrelate to each other. Third, we intend to reduce the cost of the mining process by reducing the number of body measurements used for mining. Fourth, we aim to understand the demographic nature of the individuals within each size to identify potential target markets e.g. for the clothing industry. Fifth, we attempt to improve the quality of the mined results and reduce the search space of the mining process by applying utility-based data mining techniques.

In order to identify the body size groups within each population we apply different clustering techniques. We perform multi-view classification to find the most important relationships between body measurements within each size. To achieve the reduction of the number of body measurements, we apply dimension reduction techniques such as feature selection and feature extraction. In order to determine target populations for the clothing industry, we perform association analysis and apply interestingness measures to reduce the search space and find truly interesting associations.

1.3 Thesis outline

The remaining of this thesis is organized as follows. In Chapter 2 we provide an overview of data mining and knowledge discovery. We also show what utility-based data mining entitles and present the major research trends in this emerging area. Chapter 3 gives an overview of Anthropometry and introduces the CAESAR^{TM} database, the anthropometric survey that we use in this study. In Chapter 4 we present our experimental design as well as our results when aiming to identify the body size groupings and characterize the American, Italian and Dutch populations. Chapter 5 presents the approach we follow in order to identify the relationships between body measurements and the resulted measurements relationships for the different sizes. In Chapter 6, we present the techniques we use, as well as the results, when attempting to reduce the cost of the mining process by reducing the number of body measurements. In Chapter 7 we present the result of our analysis of the demographic data when aiming to better understand the demographic nature of the individuals within each size. Finally, in Chapter 8 we present our conclusions, summarize our contributions and discuss future work.

Chapter 2

Utility-based data mining

Utility-based data mining aims to account for all the economic aspects that impact the mining process, and maximize the utility of the whole process. This chapter provides an introduction to data mining and reviews the purpose of utility-based data mining. *Interestingness measures* are introduced as a way to reduce the time and cost as well as to improve the quality of the results of the data mining process. We also present the basic steps to perform Principal Component Analysis, a statistical technique that may be used as an interestingness measure in order to simplify the working datasets.

2.1 Data mining and knowledge discovery

With the increasing use of computer technology, huge amounts of digital data are being collected, processed, managed and stored every day. There is an urgent need for powerful tools to assist human in extracting useful information from these growing volumes of data. Traditional data analysis techniques rely on manual analysis and interpretation. Unfortunately, this process is prone to biases and errors, and is extremely time-consuming and costly [Han and Kamber, 2006]. Moreover, traditional data analysis techniques often cannot be used because of the massive size of a dataset [Tan et al., 2005]. The field of data mining grows out of the widening gap between data and information.

Data mining blends traditional data analysis methods with sophisticated algorithms for processing large volumes of data. Data mining is then the process of automatically discovering useful information in data repositories. Data mining techniques are deployed to scour large quantities of data in order to find implicit, non-trivial, novel and useful patterns that might otherwise remain unknown [Chen et al., 1996, Han and Kamber, 2006,

Tan et al., 2005, Witten and Frank, 2005]. Although there is no standard definition of what constitutes a *pattern*, Fraweley et al. [Frawley et al., 1992] provide a general and broad definition:

> "Given a set of facts (data) F, a Language L, and some measures of certainty C, a pattern is a statement S in L that describes the relationships among a subset F_S of F with certainty C, such that S is simpler (in some sense) that the enumeration of all facts in F_S"[1]

Data mining is often referred as *Knowledge Discovery in Databases (KDD).* However, several authors [Fayyad et al., 1996, Han and Kamber, 2006, Tan et al., 2005] agree that the discovery of knowledge in databases is the overall process of converting raw data into useful information or *knowledge.* This process consists of an iterative sequence of steps, as depicted in figure 2.1.

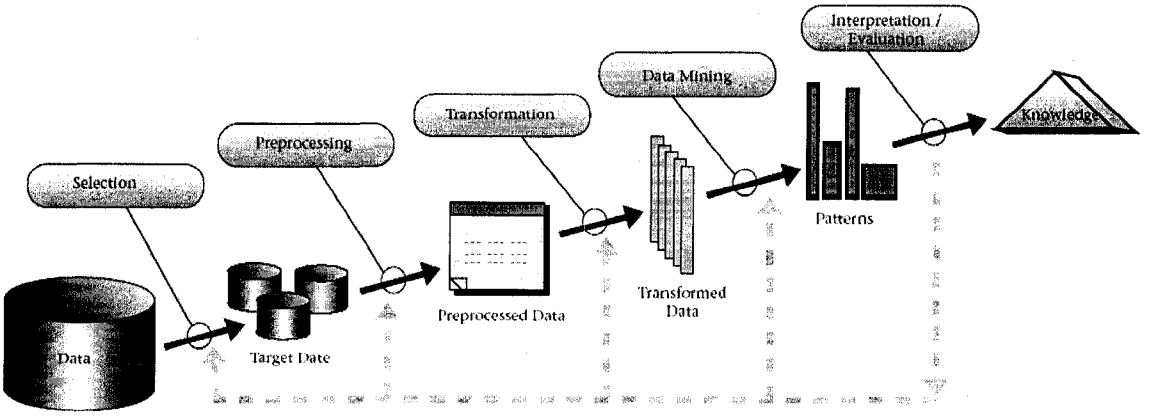


Figure 2.1: The knowledge discovery process as proposed by Fayyad et al. [Fayyad et al., 1996]

At the top level, the knowledge discovery process consists of the following steps [Fayyad et al., 1996, Han and Kamber, 2006]:

Data selection: The first step is creating a target dataset. That is, selecting a dataset, a subset of attributes or a subset of data samples. Data relevant to the analysis

[1] William J. Frawley, Gregory Piatetsky-Shapiro, Christopher J. Matheus, *Knowledge discovery in databases: an overview*, AI Magazine, vol 13, 1992, page 58.

task are then retrieved from the database. Data integration may also be performed if the data are contained in different data sources. Data integration merges data from different sources into a coherent data store.

Data preprocessing: Data preprocessing is an important issue for the knowledge discovery process, as real world data tend to be incomplete, noisy and inconsistent. Data preprocessing includes data cleaning routines. These routines attempt to fill missing values, smooth noise, and correct inconsistent data. Data cleaning is usually performed as an iterative two-step process consisting of discrepancy detection and data transformation.

Data transformation: In this step the data is converted into appropriate forms for mining. As part of the data transformations, data reduction may be performed. Data reduction techniques such as data cube aggregation, attribute subset selection, dimensionality reduction, and discretization may be used to obtain a reduced representation of the data while minimizing the loss of information content.

Data mining: Data mining is the use of automated data analysis techniques to uncover previously unknown relationships among the data. Data mining is therefore the core of knowledge discovery, because it uncovers hidden patterns. Data mining tasks are usually divided into two major categories:

Predictive methods: The goal of these methods is to predict unknown or future values based on patterns determined from know variables. A classical example of predictive methods is *Classification.* Classification is a data mining technique used to predict group membership for data instances. Classification is explained in section 5.1.

Descriptive methods: The objective of descriptive methods is to find patterns or relationships that describe the existing data. Examples of descriptive methods are *Cluster analysis* and *Association analysis.* Cluster analysis is the process of partitioning a set of data into a set of meaningful sub-sets. Clustering techniques are described in section 4.1. Association analysis identify associations or regularities in the data. An overview of association analysis is given in section 7.1.

Pattern evaluation and interpretation: In this step, interesting patterns representing knowledge are identified. A pattern represents knowledge if it is easily under-

stood by humans; valid on test data with some degree of certainty; and potentially useful, novel, or validates an expectation of the user. This step may also involve visualization and/or the employment of knowledge representation techniques to present the extracted patterns or models to the user.

2.2 What is utility-based data mining?

Most of the early research in data mining centered on producing accurate models with little consideration of the complex circumstances in which models were built and applied. It was usually assumed that training data were freely available and only simple objective measures such as predictive accuracy were considered. Over time, it became clear that these assumptions were unrealistic and that *economic utility* had to be considered during the three main phases of data mining: 1) acquiring training data, 2) building the model and 3) applying the model in realistic environments [Weiss et al., 2005, Zadrozny et al., 2006].

Recent work addresses this problem by considering economic factors related to the cost of acquiring data, the computational cost of mining the data, and the benefits of using the mined knowledge. That is, different economic factors to examine the impact of economic utility throughout the data mining process are taken into consideration. This has lead to the emerging field of *utility-based data mining.* The goal of utility-based data mining is to consider all utility aspects in the data mining process, and maximize the utility of the entire process [Weiss et al., 2005].

Research addressing the role of economic utility in data mining has focused on different aspects of the process. *Cost-sensitive learning* [Elkan, 2001], considers the costs and benefits associated with using the learned knowledge, and how these costs and benefits should be factored into the data mining process. *Active information acquisition* [Provost, 2005], focuses on methods for cost-effective acquisition of information for the training and test data. Other work also considers the cost and benefits of cleaning and transforming the training data, and the computational costs associated with building the model [Weiss and Tian, 2006]. Additional utility considerations that are relevant in the data mining context include work that focuses on reducing the time and space cost of the mining process using *interestingness measures.* These measures are used to evaluate and select the most useful patterns according to their potential interest to the user.

In this work we focus on the last trend, i.e. interestingness measures for data mining. A detailed description of these measures is given in section 2.2.1.

2.2.1 Interestingness measures for data mining

A data mining system has the potential to generate many hundreds and often thousands of patterns. However, only a small fraction of the patterns generated would actually be of interest to any given user. Manual sifting through the patterns to identify the most interesting is an overwhelming task. It is therefore necessary to establish a set of measures for evaluating the quality of the patterns to determine the most useful patterns which are not trivial or already well known. Measuring the interestingness of discovered patterns is an active and important area of utility-based data mining research.

Many interestingness measures have been proposed in the literature [Lenca et al., 2008, Bhatnagar et al., 2005, Hilderman and Hamilton, 2000, Ohsaki et al., 2004, Sahar, 1999, Padmanabhan and Tuzhilin, 1999, Tan et al., 2002]. The objective of the interestingness measures is to prune, evaluate, select and rank patterns according to their potential interest to the user. Thus, using interestingness measures facilitates a general and practical approach to automatically identify interesting patterns. According to Geng et al. [Geng and Hamilton, 2006] interestingness measures may be used during the data mining process in three different ways, which they call the *roles* of interestingness measures. These roles are shown in figure 2.2.

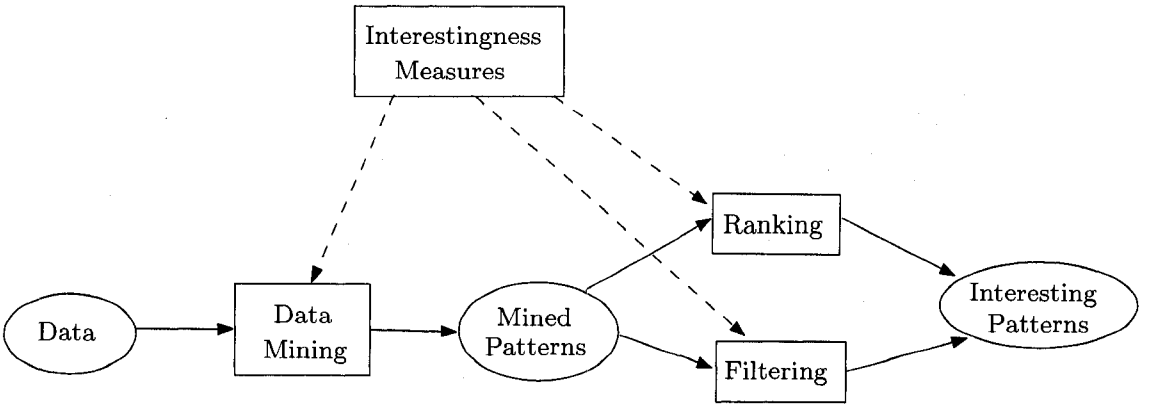


Figure 2.2: Roles of interestingness measures in the data mining process as presented in [Geng and Hamilton, 2006]

As may be seen from the figure, interestingness measures may be used to 1) prune uninteresting patterns during the mining process to narrow the search space, and thus improve the mining efficiency, 2) rank patterns according to a certain interestingness score

and 3) select or filter just the interesting patterns during post-processing. Interestingness measures then play an important role in the data mining process, since they allow the time and space costs of the mining process to be reduced [Geng and Hamilton, 2006].

2.2.2 Interestingness criteria

Although much work has been conducted in this area, so far there is no widespread agreement on a formal definition of interestingness in this context. However, based on the diversity of the definitions presented to-date, the following criteria are proposed to determine whether a pattern is interesting or not [Geng and Hamilton, 2006, McGarry, 2005]:

Generality: A pattern is *general* if it covers a large subset of the dataset. Generality measures the *coverage* of a pattern, that is, the fraction of all records in the dataset that are described by the pattern. According to [Webb and Brain, 2002] a pattern tends to be more interesting if it characterizes more records in the dataset. Many interestingness measures to find general patterns based on probability have been proposed [Agrawal et al., 1993, Brin et al., 1997b, Brin et al., 1997a, Piatetsky-Shapiro, 1991].

Reliability: A pattern is *reliable* if a high percentage of the applicable cases satisfy the relationship described by the pattern. Many measures based on statistics, probability and information retrieval have been used to measure the reliability of a pattern [Agrawal et al., 1993, Ohsaki et al., 2004, Tan et al., 2002].

Conciseness: A pattern is *concise* if contains few attribute-value pairs or conditions. This feature is desirable since concise patterns tend to be more general, easier to understand and remember. Consequentially these patterns are added more easily to the knowledge of the user. Measures to find concise patterns are mainly *form-dependent*, that is, based on the form of the rule or pattern. Some measures following this approach are proposed in [Li and Hamilton, 2004].

Peculiarity: A pattern is *peculiar* if it is significantly different from other discovered patterns. Peculiar patterns are usually derived from *outliers*, which are values that are greatly different from the rest of the data. Some interestingness measures discard the outliers as exceptions or noise. Nevertheless, in some applications such as fraud detection the rare events could be more interesting [Han and Kamber, 2006]. Dong et al. [Dong and Li, 1998] proposed two *neighborhood-based unexpectedness* measures to find peculiar patterns.

Diversity: A pattern is *diverse* if the elements that compound the pattern differ significantly from each other. Diversity is a common measure of interest for summaries. A summary is a set of attribute-value pairs and aggregate counts, where these aggregate values are given at higher level of generality that the original input data [Geng and Hamilton, 2006]. A number of measures for diversity in summaries is given in [Hilderman and Hamilton, 2000, Hilderman and Hamilton, 2001, Zbidi et al., 2006].

Novelty: A pattern is *novel* if the user did not know it before and if it is not possible to infer that pattern from other previously known patterns. Usually, in order to determine if a pattern is novel the user has to explicitly identify it as novel or observe that the pattern is not derivable from previous knowledge. In [Bhatnagar et al., 2005, Chen and Wu, 2006] the novelty criterion is used to define some measures of interest.

Surprisingness: A pattern is *surprising* or *unexpected* if it contradicts the user's expectations or beliefs. Surprising patterns are interesting because they may identify some errors in previous knowledge and could suggest aspects of the data that need to be analyzed in more detail. In [Padmanabhan and Tuzhilin, 1999, Silberschatz and Tuzhilin, 1996] the notion of unexpectedness and actionability are used to define subjective measures of interest.

Utility: A pattern is of *utility* if its application contributes to reaching a goal of the user. A *utility-based* measure considers both the raw data and the utility of the mined patterns. The simplest method to incorporate utility is called *weighted association rule mining*, which assigns to each item a weight representing its importance. This method is called *horizontal weights*. If the weight is instead assigned to transactions is called *vertical weights* [Lu et al., 2001].

Actionability: A pattern is *actionable* or *applicable* if it allows the user to make decisions about some future actions that are to his or her advantage. Different approaches have been proposed to define interestingness measures for actionability [Ling et al., 2002, Wang et al., 2002].

These interestingness criteria are not independent from one to another. For example, generality usually means reduced sensitivity to noise, which is a form of reliability. Concise patterns tend to have high coverage, and are therefore general. Peculiarity may

conflict with generality but may coincide with novelty. The difference between surprisingness and novelty is that a novel pattern is new and do not contradicts previous knowledge, while a surprising patterns contradicts the expectations of the user. Moreover, according to [Silberschatz and Tuzhilin, 1996] surprisingness may be a good approximation to actionability, and vice versa.

These nine criteria are usually divided into two categories: *objective* and *subjective*. Objective measures are based on the statistical strengths or properties of the discovered patterns. No knowledge about the user or application is required in order to assess the degree of interestingness of a pattern. Conciseness, generality, reliability, peculiarity and diversity depend only on the data patterns, consequently these criteria are considered objective. Objective measures are further discussed in section 2.2.3.

A subjective measure considers the data, but also incorporate the user's subjective knowledge into the assessment strategy. Background knowledge about the data is then needed in order to define a subjective measure. The criteria novelty, surprisingness, utility and actionability depend on the user, and therefore are considered subjective. Subjective measures are discussed in section 2.2.4. Figure 2.3 shows the taxonomy of interestingness criteria.

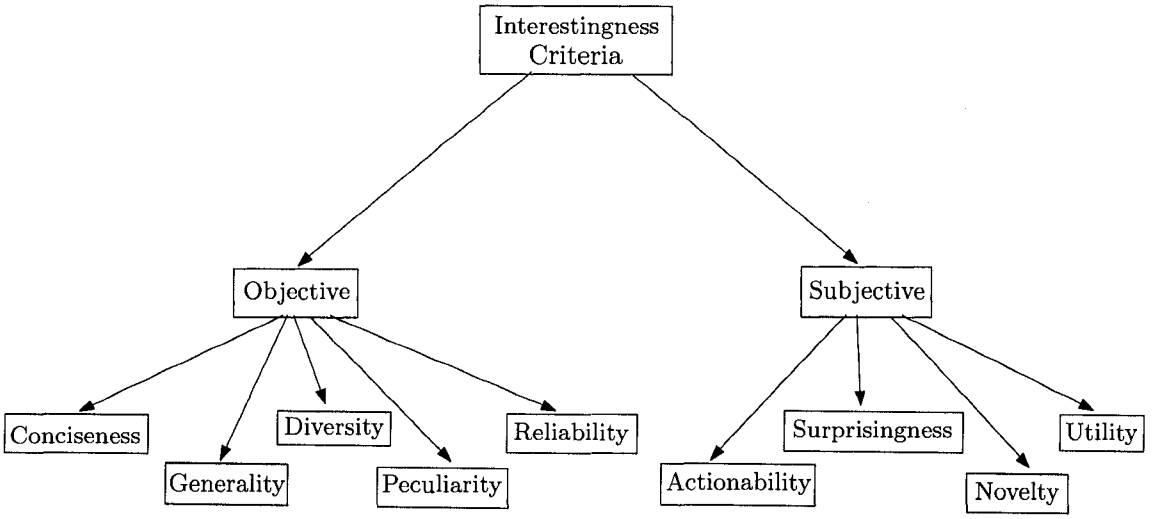


Figure 2.3: Taxonomy of the interestingness criteria

2.2.3 Objective interestingness measures

An objective measure is a data-driven approach for evaluating the quality of a pattern. It is domain-independent and requires minimal input from the user. Several objective measures of pattern interestingness have been proposed. These are based on the structure of the discovered patterns and the statistics properties underlying them. Nevertheless, most objective measures are based on theories in probability, statistics, or information theory [Geng and Hamilton, 2006].

Researchers have proposed interestingness measures for various kinds of patterns. Interestingness measures that use *diversity* as criterion have been proposed only for summaries. As mentioned previously, a summary is a set of attribute-value pairs and aggregate counts, where the aggregate values are given at higher level of generality that the original input data. The most important interestingness measures proposed for summaries in the context of diversity include the Variance, Simpson and Shannon measures [Hilderman and Hamilton, 2001, Zbidi et al., 2006]. However, most interestingness measures have been proposed for evaluating association and classification rules.

An *association rule* is defined in the following way [Agrawal and Srikant, 1994]. Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of items. Let D be a set of transactions, where each transaction T is a set of items such that $T \subseteq I$. An association rule is an implication of the form $A \rightarrow C$, where $A \subset I$, $C \subset I$ and $A \cap C = \emptyset$. For example, $\{milk, eggs\} \rightarrow \{bread\}$ is an association rule that says that when milk and eggs are purchased, bread is likely to be purchased as well.

A *classification rule* is an implication $A_1 \diamond a_1, A_2 \diamond a_2, \ldots, A_n \diamond a_n \rightarrow C = c$, where A_i is a conditional attribute, a_i is a value that belongs to the domain A_i, C is the class attribute, c is a class value, and $\diamond$ is a relational operator such as $=$ or $>$. For instance, $Age = middle\_aged$, $Income = low \rightarrow Loan\_decision = risky$, is a classification rule which says that it is risky to give a loan to a client who is middle aged and has a low income.

Probability-based measures

Probability-based objective measures evaluate the *generality* and *reliability* of the pattern. They are usually computed based on the frequency counts in a *contingency table.* A contingency table stores the frequency counts that satisfy given conditions.

Table 2.1 shows a contingency table for the rule $A \rightarrow C$, where A represents the antecedent and C the consequent. In the table, $n(AC)$ denotes the number of records

satisfying both A and C, and N denotes the total number of records.

	C	$\neg C$	
A	$n(AC)$	$n(A\neg C)$	$n(A)$
$\neg A$	$n(\neg AC)$	$n(\neg A\neg C)$	$n(\neg A)$
	$n(C)$	$n(\neg C)$	N

Table 2.1: 2 x 2 contingency table for the rule $A \rightarrow C$

Support and *confidence* were the original interestingness measures proposed for association rules [Agrawal et al., 1993]. We describe these measures next. In order to make the measures comparable, all measures are defined using probabilities. $P(A) = \frac{n(A)}{N}$ denotes the probability of A.

Support Support represents the fraction of transactions that the given rule satisfies. That is, the fraction of transactions that contain both A and C.

$$support(A \rightarrow C) = P(AC)$$

Support is used as a measure of significance or importance of an itemset. Since it basically uses the count of transactions it is often called a frequency constraint. Support is usually used to eliminate uninteresting rules, because a rule with very low support may occur simply by chance. Support's main feature is that it possesses the down-ward closure property (anti-monotonicity) which means that all nonempty subsets of a frequent itemset are also frequent [Sheikh et al., 2004]. This property is used to prune the search space. Its values are in the range [0; 1]. If the antecedent and consequent are not occurring together in any transaction, it is equal to 0. If they are occurring in all transactions, its value is equal to 1.

Confidence Confidence is defined as the probability of seeing the rule's consequent under the condition that the transaction also contains the antecedent.

$$confidence(A \rightarrow C) = \frac{P(AC)}{P(A)}$$

Confidence measures the reliability of the inference made by the rule. For a given rule $A \rightarrow C$, the higher the confidence, the more likely it is for C to occur together

with A. Confidence is not down-ward closed and it is usually used together with support. Confidence is directed and gives different values for the rules $A \rightarrow C$ and $C \rightarrow A$. Its values are in the range [0; 1]. If the antecedent and the consequent are independent, it is equal to 0. For implications occurring in all the cases, its value is equal to 1.

Using just *support* and *confidence* in assessing the pattern's interestingness may still produce a large number of rules. Moreover, confidence is sensitive to the frequency of the consequent, i.e. consequents with higher support will produce higher confidence values even if there is no association between the items [Han and Kamber, 2006]. Hence, additional interestingness measures are needed to find interesting patterns. Alternative measurements have been proposed to overcome the aforementioned problem. Here, we review *lift*, *conviction* and *leverage*, since these measures proved to be useful in finding interesting patterns [Sheikh et al., 2004].

Lift Lift is a correlation measure introduced by Brin et al. [Brin et al., 1997a] and it is given as follows. The occurrence of A is independent of the occurrence of C if $P(AC) = P(A)P(C)$. Otherwise, A and C are dependent and correlated.

$$lift(A \rightarrow C) = \frac{P(AC)}{P(A)P(C)}$$

Lift measures how many times more often A and C occur together than expected, if they were statistically independent. In other words, it assesses the degree to which the occurrence of one *lifts* the occurrence of the other. This is a measure of the importance of the association that is independent of coverage and support [Sheikh et al., 2004]. Its values are in the range $[0; +\infty)$. If the value is less than 1, the occurrence of A is *negatively correlated* with the occurrence of C. That is, satisfying the condition of the antecedent decreases the probability of occurrence of the consequent. If the value is greater than 1, A and C are *positively correlated.* This means that the occurrence of the antecedent implies the occurrence of the consequent. If the resulting value is 1, A and C are independent and there is no correlation between them.

Conviction Conviction [Brin et al., 1997b] is an interesting measure introduced by Brin et al. that also measures the independence of A and C.

$$conviction(A \rightarrow C) = \frac{P(A)P(\neg C)}{P(A\neg C)}$$

Conviction compares the probability that A appears without C if they were dependent with the actual frequency of the appearance of A without C. Conviction is similar to lift in that both measure the independence of A and C. However, in contrast to lift, conviction is sensitive to rule direction, since it also uses the information of the absence of the consequent. Its values are in the range $[0; +\infty)$, if A and C are independent, the value is equal to 1. If the value is greater than 1, A and C are positively correlated.

Leverage Leverage was introduced by Piatetsky-Shapiro in [Piatetsky-Shapiro, 1991]. Leverage measures the difference of A and C appearing together in the dataset and what would be expected if A and C where statistically dependent.

$$leverage(A \rightarrow C) = P(AC) - P(A)P(C)$$

Leverage is then the proportion of additional records covered by both the antecedent and the consequent above those expected if the antecedent and consequent were independent of each other. If A and C are independent, the value of leverage is 0. If A or C alone occurs more often, the value is less than 0. If the value is greater than 0, both A and C occur more often together.

Many other probability-based measures have been proposed to measure the interestingness of association rules [Geng and Hamilton, 2006, Lenca et al., 2008, McGarry, 2005]. For classification rules, the most important role of probability-based interestingness measures is to act as heuristics to select attribute-value pairs for inclusion in classification rules [Geng and Hamilton, 2006].

Interestingness measures are also used for *feature selection.* The main idea of feature selection is to choose a subset of features or attributes by eliminating features in the data that are statistically uncorrelated with the class labels. That is, removing the attributes with little or no predictive information. This reduces the set of attributes to be used, thus improving both efficiency and accuracy [Han and Kamber, 2006, Kim et al., 2003]. Next, we describe some of the interestingness measures widely used in this way.

The notation used here is as follows. Let D, be the dataset; suppose the class attribute has m distinct values defining m different classes, C_i (for $i = 1, \ldots, m$). Let $C_{i,D}$ be the

set of records of class C_i in D. Let $|D|$ and $|C_{i,D}|$ denote the number of records in D and $C_{i,D}$, respectively.

Information gain Information gain [Quinlan, 1986] is based on the concept of *entropy*. Informally, the entropy of a dataset may be considered to be how disordered the data is. Entropy is related to information, in the sense that the higher the entropy, or uncertainty, of some data, the more information is required in order to completely describe that data. We measure the entropy of a dataset D, with the following calculation:

$$entropy(D) = \sum_{i=1}^{m} p_i log_2(p_i)$$

Where p_i is the probability than a record in D belongs to class C_i and is estimated by $\frac{|C_{i,D}|}{|D|}$. A log function to the base of 2 is used since the information is encoded in bits. To measure the effect of selecting a particular attribute A, we use the Information Gain measure. This calculates the reduction in entropy (and therefore gain in information) that would result on selecting the attribute A. Suppose A has v distinct values, $\{a_1, a_2, \ldots, a_v\}$ and split D into v partitions or subsets $\{D_1, D_2, \ldots, D_v\}$. The information gain is calculated as follows:

$$InfoGain(A) = entropy(D) - \sum_{j=1}^{v} \frac{|D_j|}{|D|} \times entropy(D_j)$$

Where D_j contains those records in D that have outcome a_j. $Gain(A)$ tell us how much we gain by selecting the attribute A. That is, the expected reduction of the information requirement caused by knowing the value of A. The attributes with the highest information gain are then selected to be part of the feature subset.

Gain ratio Information gain tends to select attributes having large numbers of values. An extension to information gain known as gain ratio [Quinlan, 1993] overcomes this bias. In order to achieve this, gain ratio takes the number of values of an attribute into account. The gain ratio for attribute A over dataset D is defined as:

$$GainRatio(A) = \frac{InfoGain(A)}{SplitInfo(A)}$$

Where $SplitInfo(A)$ is defined as:

$$SplitInfo_A(D) = -\sum_{j=1}^{v} \frac{|D_j|}{|D|} \times log_2(\frac{|D_j|}{|D|})$$

$SplitInfo_A(D)$ represents the potential information generated by selecting the attribute A. The attributes with the highest gain ratio are selected for the feature subset.

Chi-square Another popular feature selection method is χ^2 [Mills, 1955]. This measure is applied to measure the correlation between two attributes, A and C. In feature selection, A corresponds to a conditional attribute and C is the class attribute. Suppose A has n distinct values, namely $a_1, a_2, \ldots, a_n$. C has m distinct values, namely $c_1, c_2, \ldots, c_m$. The data records described by A and C may be shown as a contingency table. Let (A_i, C_j) denote the event that attribute A takes the value a_i, and attribute C takes the value c_j. That is, where $(A = a_i, C = c_j)$. The χ^2 is calculated as:

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \frac{(o_{ij} - e_{ij})^2}{e_{ij}}$$

Where o_{ij} is the *observed frequency* (i.e., actual number) of the joint event (A_i, C_j) and e_{ij} is the *expected frequency* of (A_i, C_j). The expected frequency is computed as follows:

$$e_{ij} = \frac{count(A = a_i) \times count(C = c_j)}{|D|}$$

Where $count(A = a_i)$ is the number of records having value a_i for A, and $count(C = c_j)$ is the number of records having value c_j for C.

χ^2 is a measure of how much expected counts e_{ij} and observed counts o_{ij} deviate from each other. A high value of χ^2 indicates that the hypothesis of independence, which implies that expected and observed counts are similar, is incorrect, and therefore A and C are correlated. The rationale of χ^2 in feature selection is if the two attributes are dependent, then the occurrence of the conditional attribute makes the occurrence of the attribute class more probable. The attribute is then considered to be helpful as a feature, and included in the feature subset.

Form-dependent measures

A form-dependent measure is an objective measure that is based on the form of the rule. Form-dependent measures have been proposed based on *peculiarity* and *conciseness.*

The *neighborhood-based unexpectedness* [Dong and Li, 1998] measure proposed by Dong et al. attempts to find peculiar association rules. The general idea of this method is that if a rule has different consequent from neighboring rules, it is an interesting rule. The distance $Dist(R_1, R_2)$ between two rules, $R_1 : A_1 \rightarrow C_1$ and $R_2 : A_2 \rightarrow C_2$, is defined as $Dist(R_1, R_2) = \delta_1|A_1C_1 - A_2C_2| + \delta_2|A_1 - A_2| + \delta_3|C_1 - C_2|$, where $A - C$ denoted the symmetric distance between A and C, $|A|$ denotes the cardinality of A, and δ_1, δ_2 and δ_3 are weights determined by the user. Based on this distance, the r-neighborhood of rule R_0, denoted as $N(R_0, r)$, is defined as $\{R : Dist(R, R_0) \leq r,\ R$ is a potential rule$\}$, where $r > 0$ is the radius of the neighborhood. Based on this, two interestingness measures may be defined. The first is called *unexpected confidence*: if the confidence of a rule R_0 is far from the average confidence of the rules in its neighborhood, R_0 is interesting. The second measure is based in the sparsity of the neighborhood, that is, if the number of mined rules in the neighborhood is far less than of all potential rules in the neighborhood, it is considered interesting.

Form-dependent measures for *conciseness* are often used for rulesets rather than for single rules. Two main approaches have been proposed for evaluating the conciseness of rules. The fist method is based on logical redundancy [Li and Hamilton, 2004]. In this method no measure is defined for conciseness, instead algorithms are designed to find nonredundant rules. For example, Li et al. [Li and Hamilton, 2004] proposed an algorithm to find a minimum ruleset and an inference system. The ruleset is minimum in the sense that no redundant rules are present. All other rules may be derived from this ruleset using the inference system.

The second method to evaluate the conciseness of a ruleset is the *minimum description-length (MDL) principle* proposed by Rissanen [Rissanen, 1978]. The MDL takes into account both the complexity and the accuracy of the theory or ruleset. The first part of the MDL measure, $L(H)$ where H is a theory, is called the theory cost which measures the theory complexity. The second part, $L(D|H)$ where D denotes the data, measures the degree to which the theory fails to account for the data. For a group of theories or rulesets, a more complex theory tends to fit better than a simpler one. The former has then higher $L(H)$ value and smaller $L(D|H)$ value. The theory with the shortest description-length has the best balance between these two factors and is therefore pre-

ferred. The MDL principle has been applied to both classification and association rules [Geng and Hamilton, 2006].

2.2.4 Subjective interestingness measures

Subjective interestingness measures are user-driven in the sense that the user's subjective knowledge is taken into account in the assessment strategy. Subjective measures are therefore domain-dependent. Usually, the domain knowledge may be obtained either during the data mining process by interacting with the user or explicitly representing the user's knowledge or expectations in the data mining system. As mentioned previously, subjective interestingness measures are based on the *surprisingness* or *unexpectedness*, *novelty* and *actionability* criteria.

In order to find unexpected, novel or actionable patterns, three approaches have been proposed [Geng and Hamilton, 2006] based on the roles of the measures in the data mining process: 1) A formal specification of the user's knowledge is provided, and after obtaining the mining results, the system chooses which patterns to present to the user, 2) According to the user's interactive feedback, the system removes uninteresting patterns, and 3) The system applies the user's specifications as constraints during the mining process to narrow the search space and provide fewer results. We discuss each of these approaches next.

Filtering interesting patterns from the mined results

Silberschatz et al. [Silberschatz and Tuzhilin, 1996] presented a general framework for defining unexpected and actionable patterns. They related unexpectedness and actionability to a belief system. To define the belief system they used predicate formulae in first-order logic and assigned a confidence factor to each belief. They classified beliefs as *hard* or *soft* beliefs. Hard beliefs are the constraints that cannot change with new evidence or patterns. If a pattern contradicts these beliefs, a mistake is assumed to have been made in acquiring this new evidence. Soft beliefs are the ones that the user is willing to change with new evidence. They used a Bayesian approach and assumed that the degree of belief is measured with conditional probability.

Liu et al. [Liu et al., 1999] proposed a technique to rank classification rules according to the user's background knowledge using fuzzy rules. Based on the user's existing knowledge, three kinds of interesting rules may be mined: *unexpected*, *confirming* and *actionable*. An unexpected pattern is one that is unexpected or previously unknown to

the user. A confirming pattern is a rule that partially or completely matches the user's knowledge. An actionable pattern is one that allows decision making to do something to the user's advantage. To allow actionable patterns to be identified, the user has to describe the situations in which he or she may take actions. For the three categories, the user has to provide some patterns in the form of fuzzy rules that reflect his or her knowledge. The system matches each discovered pattern with these fuzzy rules. The discovered patterns are then ranked according to the degree to which they match.

Eliminating uninteresting patterns

In order to reduce the amount of computation and avoid the filtering of patterns after the mining process, Sahar [Sahar, 1999] proposed a method for removing uninteresting patterns, rather than selecting the interesting ones. No interestingness measure is defined. Rather, the interestingness of a rule is determined by the user through an interactive process. This process consists of three steps. First, the rule with the largest *cover list* is selected as the *best candidate rule.* The cover list of a rule R is all the mined rules that contain the antecedent and consequent of R. Second, the best candidate rule is presented to the user for classification into one of four categories: not-true-not-interesting, not-true-interesting, true-not-interesting, and true-and-interesting. Rules classified as not-true-not-interesting and true-not-interesting are removed together with its cover list. If the rule is classified as not-true-interesting the system removes this rule and all the rules in its cover list that have the same antecedent and keeps the rules in its cover list that have more specific antecedents. If the rule is true-and-interesting, the system keeps it. This process iterates until the ruleset is empty or the user halts the process.

Constraining the search space

Padmanabhan et al. [Padmanabhan and Tuzhilin, 1999] proposed a method to narrow down the mining space based on the user's expectations. The user's beliefs are represented in the same format as the mined rules. No interestingness measure is defined, instead surprising rules, that is, rules that contradict existing beliefs, are mined. The algorithm consists of two parts: ZoominUR and ZoomoutUR. For a given belief $X \rightarrow Y$, ZoominUR finds all the rules of the form $A \rightarrow \neg Y$ that have sufficient support and confidence. These are more specific rules that have the contradictory consequent to the given belief. Next, ZoomoutUR generalized the rules found by ZoominUR. Using this approach the mining space is reduced since the system only has to find the rules that

conflict with the user's knowledge.

2.3 Principal components analysis

Principal Components Analysis (PCA) is an statistical technique used for data compression that may be used for dataset reduction, by finding a set of attributes that preserves most of the features of the original data. PCA is an orthogonal linear transformation that converts a number of correlated attributes into a smaller number of uncorrelated attributes. The objective of PCA is to reduce the dimensionality or number of attributes of the dataset but retain those characteristics that contribute most to its variance [Tan et al., 2005].

Suppose we have a dataset that consists of tuples or data vectors described by n attributes or dimensions. PCA searches for the k n-dimensional orthogonal vectors that may best be used to represent the data, where $k \leq n$. The original data are then projected onto a much smaller space, resulting in dimensionality reduction. Unlike feature selection, which reduces the attribute set size by retaining a subset of the original set of attributes, PCA combines the original attributes and creates an alternative, smaller set of new attributes called *principal components.* The first principal component accounts for as much of the variability in the data as possible, and each succeeding component accounts for as much of the remaining variability as possible.

In two dimensions, principal components may be seen as a rotation of axes. In figure 2.4 the scatter plot shows a collection of (x, y) pairs. However, if the original $X - Y$ axes are rotated to form axes $U - V$, it is clear that most of the variation occurs along the U axis. The first principal component is therefore U and, it is a linear combination of X and Y. The second principal component is V. It is orthogonal to U and is also a linear combination of X and Y. The number of dimensions is then reduced by eliminating the weaker components, V in this case.

The basic steps to perform Principal Components Analysis on a set of data are as follows [Han and Kamber, 2006, Smith, 2002]:

1. The input data is normalized. The mean of each attribute is calculated and subtracted from each element belonging to that attribute, that is, $A_i - \overline{A}$, where A_i is the value of the attribute A in tuple i, and $\overline{A}$ is the mean of the values of attribute A. This produces a dataset whose mean is zero. This step helps ensure that the attributes with large domains will not dominate attributes with smaller domains.

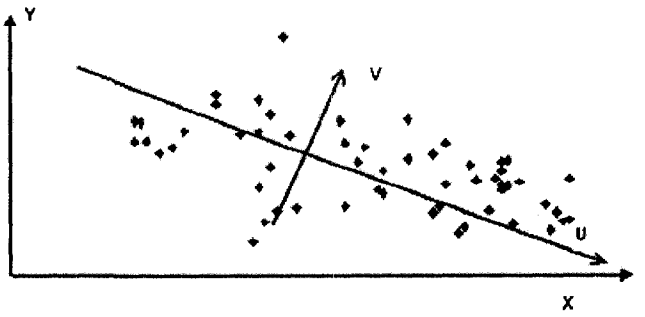


Figure 2.4: Principal components analysis in two dimensional data. U and V are the principal components of the data [Infometrics, 2004].

2. The covariance matrix of the data is calculated. Next, the eigenvectors and eigenvalues of the covariance matrix are computed. This allows finding k orthonormal vectors that provide a basis for the normalized input data. These are unit vectors such that each vector points in direction perpendicular to the others. These vectors are referred as the principal components.

3. Once the eigenvectors are found from the covariance matrix, the next step is to sort them by eigenvalue, from the highest to the lowest. This gives the order of decreasing significance or strength of the principal components. The eigenvector with the highest eigenvalue is the first principal component of the dataset. The principal components serve as a new set of axes for the data, providing important information about the variance. That is, the first axis shows the most variance among the data, the second axis shows the next highest variance, and so on.

4. Since the component are sorted according to decreasing order of significance, the size of the data may be reduced by eliminating the weaker components, that is, the ones with low variance. Figure 2.5 shows a plot of the percentage of the overall variance accounted for by principal component of the covariance matrix. This type of plot is know as *scree plot* [Tan et al., 2005] and is useful for determining how many principal components should be keep to capture most of the variability of the data. In the figure the resulted number of principal components is ten. We may want to keep just the first three principal components, because subsequent components contribute in a small proportion to the data variability.

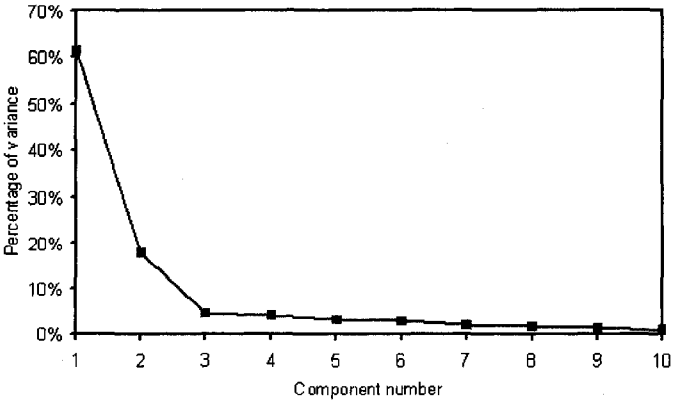


Figure 2.5: Percentage of variance accounted for by each principal component. The first three components capture the highest variance of the data.

5. The last step consists of deriving the new dataset. Once we have chosen the principal components (eigenvectors) that we wish to keep, we build a matrix where each row corresponds to each eigenvector. This matrix is called the *feature vector*. To obtain the data in terms of the chosen components, we build a matrix with the original normalized data, where each column corresponds to each data attribute. This matrix is called the *adjusted data* matrix. The final dataset is obtained by multiplying the row feature vector by the adjusted data matrix. This gives us the original data solely in terms of the vectors we have chosen.

Here we presented PCA in the context of utility-based data mining, since we use PCA for feature extraction. In this context the eigen values may be seen as the measure of interest of a given eigenvector or feature. This is further discussed in section 6.1.2.

2.4 Chapter summary

In this chapter we provided an overview of the knowledge discovery process. This process consists of a series of steps, whit data mining as the core, to uncover novel, useful and previously unknown patterns on the data. There are several economic factors that impact the overall data mining process. This leads to the emerging field of utility-based data mining. The goal of utility-based data mining is to consider the economic factors that impact the data mining process and maximize the overall utility of the entire process.

Different trends may be identified in utility-based data mining research. One of them is to incorporate interestingness measures into the data mining process, in order to reduce the time and complexity of the process, and therefore the costs. Interestingness measures are usually categorized as objective and subjective. Objective measures are data-driven in the sense that consider the statistical properties and strengths of the mined patterns. Subjective measures incorporate the user's knowledge into the assessment strategy, and are therefore user-driven.

A brief introduction to principal components analysis was given. We considered this technique in the context of utility-based data mining, as an interestingness measure to reduce the complexity of the datasets.

In the next chapter we provide a description of the data we use in this study: the CAESAR^{TM} database. The CAESAR^{TM} database is an anthropometric survey that includes, along with anthropometric and demographic data, 3-D body scans of each participant.

Chapter 3

Anthropometric data

In this chapter, a general introduction to anthropometry is provided. We describe the dataset used in this work: the CAESAR^{TM} database, a survey of anthropometric measurements and 3-D body scans of three different populations. The different components (Anthropometric, Demographic and 3-D data) of the database as well as the populations considered in this work are described.

3.1 Anthropometry overview

Anthropometry is the study of the measurement of the human body dimensions, with the aim of establishing the physical geometry, mass properties and strength capabilities thereof [Roebuck, 1993]. These measures may be lengths (e.g. the length of the arm from shoulder to wrist), breadths (e.g. the width across the head, the head breadth), girths (e.g. chest circumference) as well as more common measures such as stature (height) and mass (weight). It also includes the measurement of skinfold thickness at various parts of the body (e.g. overlying the shoulder blade, the subscapular skinfold). The purpose of anthropometry is then to describe and characterize the human body through a set of measurements [Paquet et al., 2000].

Anthropometry plays an important role in many applications ranging from design and evaluation of vehicles, work sites and equipment, ergonomic furniture and architecture to clothing design. In all these applications, statistical data about the distribution of body dimensions in the population are used to improve the quality of the products.

Many anthropometric surveys have been carried out [Gordon et al., 1989] in which thousands of subjects were measured. However, in most of them only traditional mea-

sures (e.g. taken with an anthropometer, caliper and tape) were considered. Furthermore, most of these surveys were conducted to obtain measurement data of military personnel. Recent surveys, such as the CAESAR^{TM} project, provide more complete and detailed information about the measurements of adult civilian population. This is achieved by considering also the whole 3-D surface, allowing the statistical analysis of not only a set of scalar measurements, but the whole shape as well [Azouz et al., 2006].

3.2 The CAESAR^{TM} database

The Civilian American and European Surface Anthropometric Resource (CAESAR) project is a survey of European and North American civilian populations. It was a joint effort that involved collaboration of the U.S. Air Force, the Society of Automotive Engineers (SAE), representing the automotive and aerospace industries, and the American Society of Testing and Material (ASTM) constituted from representatives of the apparel industry.

The purpose of this project was to create an anthropometric database containing up-to-date information about the civilian adult population of the North Atlantic Treaty Organization (NATO) countries. This survey involved a large number of individuals from Italy, the Netherlands and North America. These countries were chosen because Italians represent the shortest population of NATO countries, while the Dutch represent the tallest population in the western world, and Americans represent the most heterogeneous population, due to its broad ethnic diversity [Robinette et al., 2002].[1]

The CAESAR^{TM} database [Blackwell et al., 2002] includes forty-five traditional measurements such as height, weight, acromial height, waist circumference, foot length, etc. that were recorded by domain experts, as well as three-dimensional body measurements for the whole body. A laser scanning device measured and recorded detailed geometry of the subjects' body surface. Each subject was scanned in three different postures: standing, seated in a comfortable working position, and seated in coverage position. In addition to the traditional measurements and the 3-D scans, demographic data were also recorded. This data include family income, age, occupation, number of children, etc. of the participants.

Other surveys have been performed; among them there is a study performed on young physically active adults in USA [Heinz et al., 2003]. This study collected nine skeletal

[1]Note that, in this study, subjects from Canada were included in the United States of America (USA) category.

measurements and twelve girth measurements of 507 participants. In addition to the skeletal and girth measurements, the age, gender, weight and height were also recorded for each participant. Another survey is *Size USA* [siz, 2003], where 10,000 subjects were scanned in standing position only, using a 3-D whole body scanner. No anthropometric measurements were considered, but 200 measurements were extracted from the 3-D scans. Other surveys have been performed on other populations, such as *Size UK* [siz, 2001], where 11,000 subjects whose ages range between 16 to 90 years old, were scanned in two poses: standing and seated; only 8 anthropometric measurements were taken using traditional instruments, and 130 measurements were computed from the 3-D scan. *Size China* [siz, 2007] is another anthropometric survey that aims to collect the head and face sizes of the Chinese population through 3-D digital scanning. Both anthropometric measurements and 3-D data were considered in this survey. Other surveys are *Size Mexico* and *Size Thailand* [wik,], but these surveys have not been completed at the date of this work.

In this work, we chose the CAESAR^{TM} database due to three main reasons. First, the CAESAR^{TM} database includes both 3-D scans and forty-five anthropometric measurements taken by domain experts, which is a better approximation of the kind of measurements that are considered for industrial purposes. Second, the CAESAR^{TM} database provides homogeneous information about three populations that have diverse body constitutions. Third, although the number of individuals considered in CAESAR^{TM} is smaller than the number of subjects in e.g. Size USA, the information about 3-D scans is richer, since each individual is scanned in three different poses.

A detailed description of the anthropometric measures considered in the CAESAR^{TM} database and how these were acquired is given in section 3.3. The acquisition of 3-D body scans and a description of the three distinct postures is shown in section 3.4. The variables considered in the demographic data are presented in section 3.5. Section 3.6 describes the populations used in our study.

3.3 Anthropometric measurements

In the CAESAR^{TM} project, traditional anthropometric measurements taken with tape, anthropometer, caliper, etc. along with the 3-D scans were considered. The reason for this choice is that software for automatic location and extraction of anthropometric measurements from 3-D scans has not yet reached a maturity level to be used it as a completely reliable tool. In many cases, the measurement extraction soft-

ware identifies the waist, bust, hip and abdomen, either at a higher or lower location than they actually were. The measurements are then measured at the wrong location [Devarajan and Istook, 2004]. Therefore, the need of traditional anthropometric measurements.

A total of forty-five traditional anthropometric measurements are included in this survey. By convention, in the CAESAR^{TM} project, the names of the measurements have up to four components: the body segment (e.g. *acromion*, *arm*, *foot*, etc.), followed by the measurement type (which are *length*, *height*, *breadth*, *circumference*, *reach* and *skinfold*), and followed, if applicable, by the pose (*standing* or *sitting*) and the side (*left* or *right*). These measurements are shown in table 3.1, and a complete graphic depiction showing how these measurements are taken is given in Appendix A. A more detailed description of these measurements is presented in [Blackwell et al., 2002].

To ensure the precision and accuracy of the anthropometric measurements, domain experts took them using instruments with a precision up to 1 mm. The obtained values were double-checked [Paquet et al., 2007] by another anthropometric expert to avoid mistakes. Moreover, in the data collection phase, the information was recorded in paper form and entered into the computer. When entering the data, the system performed an automatic check for detecting eventual outliers and prompted the user, in order to minimize errors during this process. The range of the outliers was determined by the minimum and maximum values from past anthropometric surveys. During the data analysis phase, the paper recorded data were compared with the electronic data, and corrections were made whenever discrepancies were found. A final verification of all measurements was performed using a regression outlier analysis which is a several step process. First, measurements correlations were used and the regression models with the highest correlations were selected. Next, an examination of the predicted values against the actual ones was done. Any value higher than 4.5 standard errors was checked. This checking involves examining the paper forms, other related measurements and viewing the 3-D scans of the subject. In some cases, an alternated measurement taken form the 3-D scan was used to verify the accuracy of a measurement and correct or delete clear errors [Robinette et al., 2002].

3.4 3-D scan data

In the acquisition of the three-dimensional body surfaces, two different 3-D scanning technologies were used but they followed the same principles: a low intensity laser moved

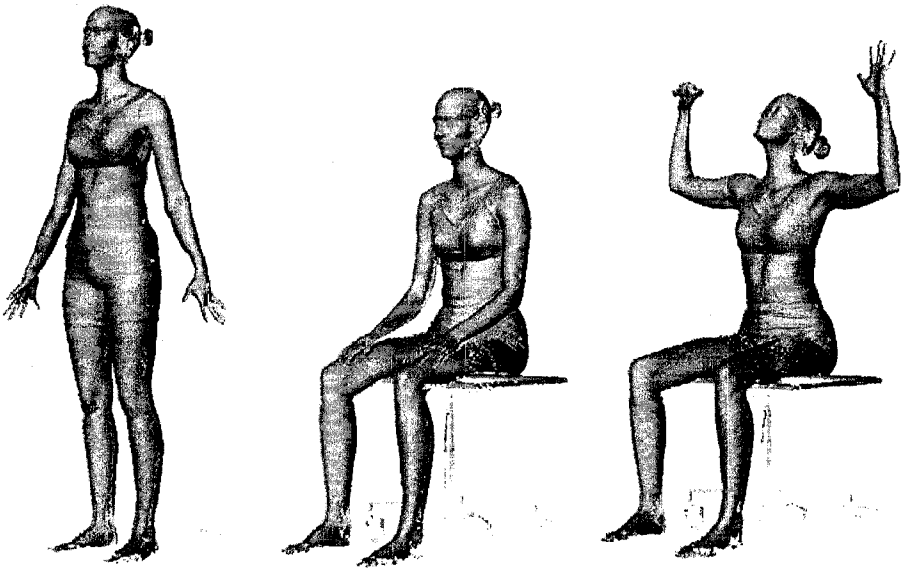
Anthropometric Measurements	
1. Acromial Height, Sitting	24. Hip Circumference, Maximum
2. Ankle Circumference	Height
3. Arm Length (Shoulder-Elbow)	25. Knee Height, Sitting, Right
4. Arm Length (Shoulder-Wrist)	26. Neck Base Circumference
5. Arm Length (Spine-Wrist)	27. Shoulder Breadth
6. Armscye Circumference (Scye	28. Sitting Height
Circumference over Acromion)	29. Spine to Shoulder
7. Bizygomatic Breadth	30. Spine to Elbow
8. Bust/Chest Circumference	31. Stature
9. Bust/Chest Circumference under	32. Subscapular Skinfold, Right
Bust*	33. Thigh Circumference, Maximum,
10. Buttock-Knee Length, Right	Right
11. Chest Girth (Chest Circumference	34. Thigh Circumference, Maximum,
at Scye)	Sitting, Right
12. Crotch Height	35. Thumb Tip Reach, Right
13. Elbow Height, Sitting, Right	36. TTR1†
14. Eye Height, Sitting, Right	37. TTR2†
15. Face Length (Menton-Sellion Length)	38. TTR3†
16. Foot Length, Right	39. Total Crotch Length
17. Hand Circumference, Right	40. Triceps Skinfold
18. Hand Length, Right	41. Vertical Trunk Circumference,
19. Head Breadth	Right
20. Head Circumference	42. Waist Circumference, Preferred
21. Head Length	43. Waist Front Length
22. Hip Breadth, Sitting	44. Waist Height, Preferred
23. Hip Circumference, Maximum	45. Weight (Mass)

* This measure is considered only for females. † These are the same measurement (Thumb Tip Reach), it was measured three times because this measurement is difficult to take.

Table 3.1: Traditional anthropometric measurements considered in CAESAR™.

around the subject capturing the shape and color of the entire human body in a few seconds. In the United States and Italy, a WB4 Cyberware scanner was used, and a scanner built by Vitronic was used in the Netherlands [Robinette et al., 2002].

Each subject was scanned in three different postures: Pose A is a standing posture, Pose B is a seated posture where the subject takes a comfortable working position, and in Pose C the subject is seated with his or her arms raised and the head slightly backward to provide a better coverage. Figure 3.1 shows the three scanning postures [Blackwell et al., 2002, Robinette et al., 2002].



(a) (b) (c)

Figure 3.1: The three 3-D scanning postures: (a) standing posture, (b) seated in comfortable working posture and (c) seated in coverage posture.

Pose A: Standing Posture The subject place his or her feet on foot outlines which are separated around 10 cm. The subject is standing up and looking straight ahead. Using a 20 cm dowel, the subject's arm position is adjusted. Thus the arms are 20 cm away from the lateral-most point of the hip or thigh area as viewed from the front. The arms and the wrist are straight, and the palms with the fingers spread are facing the body.

Pose B: Seated comfortable working posture This posture is different from the standard traditional measurement pose, where the subject is seated in a rigid po-

sition. The purpose of this posture is to capture the natural and comfortable working position. The subject sits on a modified stool, and the height is adjusted so both subject's feet are flat on the platform. The subject sits up straight and look straight. The hands are placed on the thighs to avoid the medial and lateral femoral epicondyles from being blocked. Finally, the subject is asked to relax the rigidity of the position and assume a comfortable working position.

Pose C: Seated coverage posture The purpose of this posture is to expose and capture hard-to-see areas of the body that usually remain blocked in the previous postures. These areas usually are underneath the arms and chin, and between the legs. In this posture the subject sits up on the stool and looks straight ahead with his or her feet on foot outlines. The legs are slightly spread to allow coverage in this area. The hands are placed over the head forming a right angle with the shoulders and elbows. The right hand is closed around a dowel and the fingers of the left hand are spread. The head is tilted backward so the chin/neck is greater than 90°.

As part of the quality control of the 3-D data, the scanned image was previewed (option available only in the Cyberware scanner) for about one minute to ensure it was of good quality and all the landmarks were visible. If this was not the case the scan was re-taken. Additionally, an automatic heuristic check was used to identify landmarks out of place or with the wrong name. If an error was found during this check, an operator made the necessary corrections [Robinette et al., 2002].

3.5 Demographic data

The demographic data [Blackwell et al., 2002] were collected through a demographic questionnaire that was filled out by the participants. Most of this information corresponds to a typical anthropometric survey questions, but data about the subject's car were also recorded in the CAESAR™ project. This data were then entered into the computer and, in case of apparent errors or inconsistencies, the information was clarified with the participant. A total of 25 questions were included in the questionnaire. The list of all demographic variables is shown in table 3.2 and a description of each variable is given below:

Country of Data Collection This represents the country where the data were collected. The possible values for this variable are *Italy*, *Netherlands* or *United States*. The data collected in Canada were included under the United States category.

Demographic Variables	
1. Country of Data Collection	15. Car Model
2. Site of Data Collection	16. Gender
3. Date of Data Collection	17. Race
4. Time of Data Collection	18. Reported Height
5. Civilian or Military	19. Reported Weight
6. Date of Birth	20. Marital Status
7. Age in years	21. Family Income
8. Birth State	22. Shoe Size
9. Occupation	23. Jacket Size
10. Education Level	24. Pants Size Waist
11. Number of Children	25. Pants Size Inseam
12. Fitness Level	26. Blouse Size
13. Car Make	27. Pants Size Woman
14. Car Year	28. Bra Size

Table 3.2: Demographic variables considered in CAESAR^{TM}.

Site of Data Collection This refers to the city where the data collection was carried out. In the US a total of twelve different locations were included (including the Site located in Ottawa, Ontario), two in Italy and one for the Netherlands.

Date of Data Collection This is the date when the data were collected for a given subject. This was computer-generated.

Time of Data Collection This was computer-generated and represents the arrival time of the subject to the data collection site.

Civilian or Military This refers to whether the subject was part of the armed forces or not. The value *No* is given if the subject was part of the armed forces and the value *Yes* is recorded otherwise.

Date of Birth This represents the date of birth of the subject as he or she entered in the questionnaire.

Age in years This was computer-generated subtracting the subject's date of birth from the date of data collection.

Birth State In the United States this represents the state, in the Netherlands the providence, and in Italy the Italian region where the subject was born.

Occupation This represents the occupation of the subject. This value was chosen by the subject from a list of occupations. The values *Retired*, *Student*, *Unemployed* and *Other* were included.

Education Level This represents the highest level of education completed by the subject. The possible values range from elementary school to post-doctoral studies.

Number of Children This is the number of children the subject has; the possible values are from *0* to *7 or more.*

Fitness Level This is the number of hours per week of structured exercise the subject exercise; the range of values are from *0* to *More than 10.*

Car Make This is the manufacturer of the primary vehicle of the subject. This was selected from a list of manufacturers.

Car Year This is the year of manufacture of the primary vehicle of the subject.

Car Model This is the model of the primary vehicle of the subject.

Gender This represents the gender of the subject.

Race This is the subject's race. The values included in the US were *African-American*, *Caucasian*, *Native American*, *Spanish/Hispanic*, *Asian/Pacific Islander* and *Mixed Race.* The values *Not listed above* and *No response* were also included. For Italy, the possible values were *Italian* and *Other.* For the Netherlands the values were *Dutch* and *Other.*

Reported Height This the subject height as perceived by the subject himself or herself.

Reported Weight This the subject weight as perceived by the subject himself or herself.

Marital Status This is the subject's marital status. The subject selected the status from a list.

Family Income This is the subject's family net income. For the Americans family income was divided into nine categories ranging from *less than $10,000 US* to *more than $100,000 US* annually. In the case of Italians the income was given in Euros divided in ten categories. For the Dutch the income was given in thousands of guilders, the income was divided into eight categories from *less than 20k* to *160k-200k*. For comparison purposes we convert the income of the Italians and Dutch into US dollars.

Shoe Size This represents the subject's shoe size. The values are given accordingly to each country shoe sizes.

Jacket Size This is the jacket size of the subject, this variable if for men only. For the Americans the possible values are from *30 or smaller* to *48 or larger*. For the Italians the possible values are from *46 or smaller* to *58 or larger*. In the case of the Dutch, the values range from 34 to 62.

Pants Size Waist This is the subject's pants size waist. The values for the Americans and Dutch range from *28 or smaller* to *46 or larger*. For the Italians the subject entered a value in cm.

Pants Size Inseam This is the subject's pants length inseam. The possible values for the Americans and the Dutch range from 28 to 40. In the case of the Italians it is common to buy unfinished pants and later a tailor adjusted them accordingly, then the subject entered a value in cm.

Blouse Size In the United States and the Netherlands this represents the numeric value of women blouse sizes. For the Americans these values are from 4 to 22. For the Dutch the possible values are from 34 to 62. In Italy the blouse sizes for women are small, medium, large, x-large and xx-large.

Pants Size Woman This is the pants size for women. The ranges of values for the Americans are from 2 to 20. The values for the Dutch range from 34 to 60. The possible values for the Italians are from 36 to 60.

Bra Size In the United States and Italy this represents the bra size. In the Netherlands this represents the chest circumference under the bust in cm.

3.6 Composition of the population

Since the goal of the CAESAR^{TM} project was to characterize the population of the NATO countries, a stratified sampling strategy was followed to ensure all these groups were equally represented. Thus, the population was sampled considering the age, gender and race. Additionally to these strata, aspects such as height, weight, education and region were considered to ensure the sample was representative of the civilian population.

North America was chosen by its population diversity, then, three ethnic groups strata were considered: *Caucasian*, *African-American* and *Other*. In the *Other* group, Asian, Native American and Hispanic populations were included. The strata considered in North America is shown in table 3.3.

Age Strata	**Gender Strata**	**Ethnic Group Strata**
18-29	Male	Caucasian
30-44	Female	African-American
45-65		Other

Table 3.3: Distribution of the strata for North America.

In Italy and the Netherlands the purpose was to gather data of the shortest and tallest populations of the western world. Then, only two ethnic groups were considered: *Caucasian* and *Other*, where the *Caucasian* group was conformed by the subjects for whom both parents were born in the country, i.e. Italy and the Netherlands. All the remaining subjects were considered in the *Other* group [Robinette et al., 2002]. The strata considered in both Italy and the Netherlands is presented in table 3.4.

Age Strata	**Gender Strata**	**Ethnic Group Strata**
18-29	Male	Caucasian
30-44	Female	Other
45-65		

Table 3.4: Distribution of the strata for Italy and the Netherlands.

The goal of the stratified sampling was to have equal number of subjects on each stratum to ensure that the different populations are well represented in the sample.

However, achieving a high number of subjects on the minority groups was a difficult task. The actual number of subjects per strata on each population is given in tables 3.5-3.8.

Age	Caucasian	African-American	Other	Total
18-29	57	8	16	**81**
30-44	87	5	14	**106**
45-65	61	4	4	**69**
Total	**205**	**17**	**34**	**256**

Table 3.5: Number of subjects per strata for females in North America.

Age	Caucasian	African-American	Other	Total*
18-29	102	7	18	**127**
30-44	156	12	21	**189**
45-65	91	1	6	**98**
Total*	**349**	**20**	**45**	**414**

* Totals include subjects who have missing data or are outside the age ranges.

Table 3.6: Number of subjects per strata for males in North America.

Females				Males			
Age	**Italian**	**Other**	**Total***	**Age**	**Italian**	**Other**	**Total***
18-29	254	5	**259**	18-29	238	14	**252**
30-44	67	4	**71**	30-44	103	7	**110**
45-65	57	1	**58**	45-65	50	1	**51**
Total*	**378**	**10**	**388**	**Total***	**391**	**22**	**413**

* Totals include subjects who have missing data or are outside the age ranges.

Table 3.7: Number of subjects per strata in Italy.

Note that the American population considered in this work is a subset of the whole population in the CAESAR^{TM} project, since only this portion corresponds to civilian

Females				Males			
Age	**Dutch**	**Other**	**Total***	**Age**	**Dutch**	**Other**	**Total**
18-29	158	44	**202**	18-29	144	28	**172**
30-44	199	50	**249**	30-44	160	26	**186**
45-65	186	63	**249**	45-65	176	33	**209**
Total*	**543**	**157**	**700**	**Total**	**480**	**87**	**567**

* Totals include subjects who have missing data or are outside the age ranges.

Table 3.8: Number of subjects per strata in the Netherlands.

population. The remaining data in the sample of the American population corresponds to military personnel. In the case of Italy and the Netherlands we are using the total sample as released by the CAESAR^{TM} project. Thus, the total number of participants we use in this study is as follows: in North America 670 subjects, in Italy 801 subjects, and in the Netherlands 1,267 subjects.

From table 3.7 we may observe the Italian sample is biased toward young people. Actually around 64% of the sample (females and males) is in the age range 18-29. In section 7.3 we analize the impact in the results of this bias in a demographic profile. We notice that in North America and Italy, the number of subjects in the minority groups are very small. This sample cannot be considered then as representative of the whole population belonging to these groups in those countries. The best sampling was obtained in the Netherlands, where the number of the subjects per age strata is more balanced as well as the amount of subjects in the *Other* group.

3.7 Chapter summary

In this chapter we provided an overview of what anthropometry entails and described the CAESAR^{TM} database, the measurements it provides, the demographic data and the 3-D scans that will be further used in this work. Our choice of this survey is based in the fact that, in this database, both the anthropometric measurements and the 3-D data are considered. Moreover, the survey includes three populations with different typical body constitutions.

In the next chapter we present the approach we follow in order to characterize the American, Italian and Dutch populations considering both, the anthropometric measure-

ments and the 3-D body scans. We describe the data mining reduction technique that will allow us to find representative individuals and their distinctive characteristics. Then, using these selected individuals we analyze the three populations. We also compare the results obtained from the anthropometric measurements with the results obtained from 3-D data.

Chapter 4

Characterization of the three populations

One of the greatest challenges for the apparel industry is to provide garments that fit the customers well, are esthetically pleasing and comfortable. This is of crucial importance not only for the expensive designer labels but also for the mass market. Poor fitting garments may never be sold or customers may return them. Moreover, poor fitting garments are a major reason of lack of satisfaction for customers and the cause of billions of dollars of lost revenue for the apparel industry [Ashdown and Dunne, 2006]. In order to produce better fitting garments, accurate and up-to-date measurements of the human body, and a better characterization thereof, are needed. This implies the sizes (e.g. small, medium, large) must correspond to real body shapes, in the sense that one or more *archetypes* should represent any individual belonging to the same size [Viktor et al., 2006].

To address the aforementioned problem, we aim to find the natural body size groupings within the CAESAR^{TM} anthropometric database. From these groups we intend to identify size *archetypes* and their most important characteristics. We also want to analyze the relationships between the anthropometric measurements of each size.

In order to perform our analysis, we follow the approach proposed by Viktor et al. [Viktor et al., 2006]. They perform *cluster analysis* to partition the American male population in the CAESAR^{TM} database, into five groups or *clusters*. Each of these clusters represents a clothing size. In the cluster analysis, they consider a number of clustering algorithms. By inspection of the obtained clusters and through verification of the results against the 3-D body scans, they found that the k-means algorithm produced the best results. From each of these clusters they obtain *archetypes* and their main

features. The archetype corresponds to the closest individual to the Centroid of the cluster. Next, using the obtained cluster sizes as class labels, they engage in *multi-view classification* in order to find the interplay of the body measurements for each size. *Multi-view classification* is a technique that allows mining multi-relational data using single-table learning algorithms. Throughout this and the next chapters we present a more detailed description of the whole process.

In this chapter we provide a short overview of cluster analysis as well as the main clustering techniques. We introduce the methodology we follow in order to perform the cluster analysis on the CAESAR^{TM} populations. We first perform clustering on the anthropometric data. From these clusters, we identify the cluster archetypes. Based on the obtained archetypes, we analyze the three populations, and highlight some differences and similarities among the three populations. We also present the results of applying clustering techniques to the 3-D data. Finally, we compare the clustering results based on anthropometric data with the results of clustering the 3-D data.

4.1 Cluster analysis overview

Cluster analysis or *clustering* is an unsupervised learning data mining technique used to group data records into unlabeled classes. More formally, clustering is the process of partitioning a set of physical or abstract objects into subsets or *clusters* based on data similarity [Tan et al., 2005]. A cluster is a collection of objects that are similar to one another and are significantly different from the objects in other clusters. The similarity or dissimilarity between two data objects is determined based on the attribute values describing the object, and some distance measure. Euclidean, Manhattan and Minkowski distances are well known methods for distance measurement [Abdali et al., 2004, Han and Kamber, 2006].

Clustering techniques are grouped into *partitioning, hierarchical, density-based* and *model-based* methods [Abdali et al., 2004, Han and Kamber, 2006, Tan et al., 2005].

Partitioning methods construct the clusters by dividing the dataset into k partitions or *clusters*, where k is the desired number of clusters provided by the user. Partitioning algorithms create an initial partitioning and, by an iterative reallocation process, attempt to improve the partitioning according to some criterion function. The criterion function to evaluate the quality of a partitioning could be, for example, the square-error function.

Hierarchical methods create a hierarchical decomposition of the dataset, i.e. a tree of clusters also known as *dendogram*. Hierarchical methods are categorized into *agglomerative* (bottom-up) and *divisive* (top-down). The agglomerative approach starts with one-object clusters and recursively merges clusters that are close to one another. On the other hand, the divisive approach starts with all the objects in the same cluster and recursively splits into smaller non-overlapping clusters. The process continues until a termination criterion (for example the requested number k of clusters) is achieved.

Density-based methods consider the clusters as dense regions of objects in the data space that are separated by regions of low density. The general idea employed by these methods is to continue growing a cluster as long as the density (number of objects or data points) in the neighborhood exceeds some threshold. Density-based methods thus discover clusters of arbitrary shape.

Model-based methods attempt to optimize the fit between the dataset and some mathematical model. These methods are based on the assumption that the data is a mixture of independent samples from a series of heterogeneous group populations. In these methods, it is assumed that the data are generated by a model, the clustering algorithm then tries to recover the original model from the data. Since these methods attempt to find heterogeneous groups in the data, this leads to an automatic way of determining the number of clusters.

Other clustering methods include *grid-based*, *constraint-based* and *fuzzy* clustering [Abdali et al., 2004, Han and Kamber, 2006, Tan et al., 2005]. Clustering is considered a data reduction technique, in the sense that through clustering we may characterize the objects in each cluster in terms of a cluster *archetype*. An archetype is defined as the centrally located subject in each cluster under the assumption that the cluster has a spherical or quasi spherical symmetry. If it is not the case, more than one archetype may be necessary to fully characterize the cluster. Archetypes are then representatives of all other subjects that belong to the same cluster [Abdali et al., 2004]. Archetypes may be used for designing common consumer products, e.g. clothes that fit the population better.

4.2 Clustering of anthropometric data

As introduced in this chapter, we aim to characterize the CAESAR^{TM} population. In order to identify the natural body size groupings within the CAESAR^{TM} populations, we first use the anthropometric data. In this study we consider the American female population as well as the Italian and Dutch, both male and female populations. The data was first separated based on the gender of the subject. The resulting sets consist of 256 American females, 413 Italian males, 388 Italian females, 567 Dutch males and 700 Dutch females.

All our experiments are implemented in WEKA, a collection of machine learning algorithms for data mining tasks written in Java, developed at the University of Waikato [Witten and Frank, 2005]. Additionally, in order to verify the quality of our results, we use the Cleopatra system, a 3-D information retrieval system developed at the NRC [Paquet et al., 2000].

For economical and practical reasons, in the context of tailoring, the ideal scenario is to cover the greatest number of people with the fewest number of sizes [Hsu et al., 2007, Hsu and Wang, 2005]. Therefore, we aim to find the minimum number of clusters that fully characterize the population.

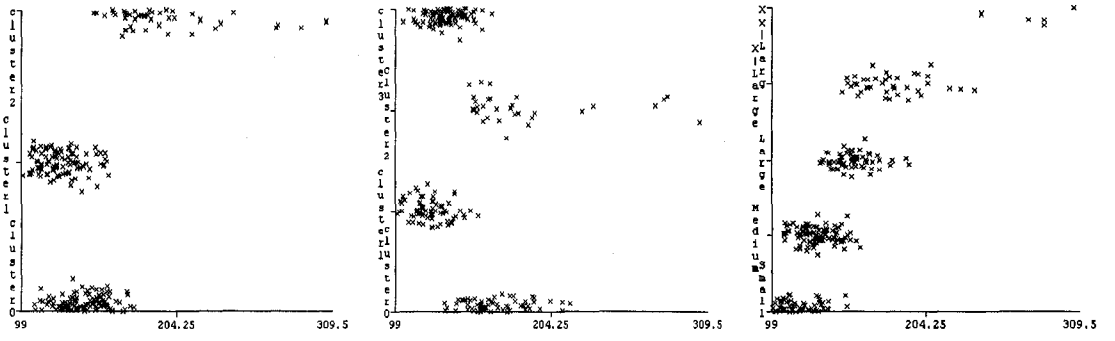
4.2.1 Determining the number of clusters

So far, there is no standard method to determine the correct or *natural* number of cluster in a given dataset [Witten and Frank, 2005]. In order to determine the number of clusters that best describe the populations of our study, we first consider the algorithms *EM*, an algorithm that performs expectation-maximization analysis based on statistical modeling [Dempster et al., 1977], and *Classit*, a popular and simple method of incremental conceptual clustering that creates a classification tree [Gennari et al., 1989]. EM and Classit are model-based algorithms that, as discussed in section 4.1, may be used to automatically determine the number of clusters.

The number of clusters obtained with EM and Classit were significantly too large for our purposes. The number of clusters generated by Classit goes from 243 to 647 in all populations. EM produced 11 to 14 clusters for the Italian and Dutch populations, both males and females. For the American females, EM produced 6 clusters. As may be seen, EM produces a smaller number of clusters than Classit, but still large from a tailoring point of view.

Thus, we follow the approach proposed by Witten et al. in [Witten and Frank, 2005]

to determine the optimal number of clusters. They propose an iterative process where the dataset is first divided into two clusters. Next, by inspection of the cluster distribution, they decide whether is worth splitting the clusters. A cluster that is highly spread indicates that further splitting is needed. This process is repeated until no more splitting is needed. This process allows finding the minimum number of clusters that characterize properly the data, since we start from the minimum number of clusters and we increase (split) a cluster only when a cluster is not well defined. We illustrate this process on the American female population in figure 4.1.



(a) (b) (c)

Figure 4.1: Determining the number of clusters for the American females. Figure shows the cluster distribution with (a) three clusters, (b) four clusters and (c) five clusters. The y axis represent the clusters, the x axis is the weight range.

Since three clusters is the minimum number that makes sense from a tailoring point of view, i.e. *small*, *medium* and *large*, we start partitioning the data into tree clusters, as presented in figure 4.1(a). The figure shows that the cluster in the top spreads almost over the entire graph, suggesting further splitting. Next, we consider four clusters as depicted in figure 4.1(b), again the second cluster from top to bottom need to be split. We then consider the five clusters in figure 4.1(c). We observe the clusters look like compact clouds of points and they are well defined. Further splitting may produce a partition with a cluster containing a very small number of subjects. Therefore, by inspection of the cluster distribution and through the analysis of the results using Cleopatra, the number of clusters for the American female population was set to five. These five clusters correspond to the number of clothing sizes for the American females. Following the aforesaid process, we determine the best number of clusters for the Italian and Dutch populations. The results indicate that, for the Italian males and Dutch males, the best number of clusters is five. For the Italian females and Dutch females, the best number of clusters is six.

4.2.2 American female population

For our clustering experimentation, we consider four different cluster algorithms. *EM*, a model-based algorithm that performs expectation-maximization analysis using the finite Gaussian mixture model [Dempster et al., 1977]; *K-means*, a centroid-based partitioning algorithm in which each cluster is represented by the mean value of the objects in the cluster [MacQueen, 1967]; *Farthest First*, a hierarchical divisive algorithm that implements the farthest-first traversal of a set of points [Dasgupta and Long, 2005]; and an algorithm using a variation of the *density-based* clustering algorithm with k-means components. This algorithm performs a density redistribution of the k-means partition [Witten and Frank, 2005]. Here, we do not consider the *Classit* algorithm since, as opposed to *EM*, *Classit* does not allow manual setting of the number of clusters.

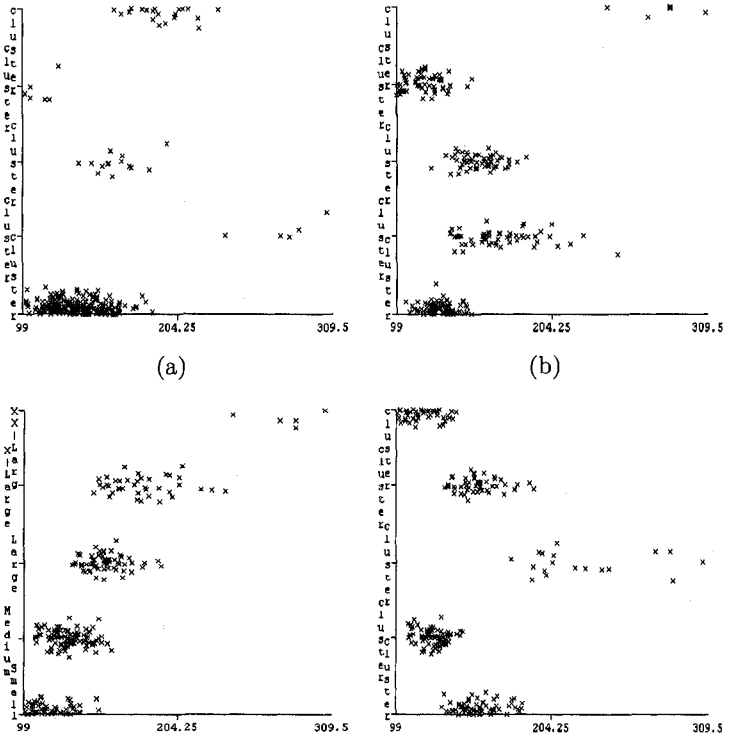


Figure 4.2: Cluster visualization of the different algorithms for the American females. (a) Farthest first, (b) EM, (c) K-means and (d) Density-based algorithm with k-means components. The y axis represent the clusters, the x axis is the weight range.

The results of applying the mentioned clustering algorithms to the American female

population are discussed next. Figure 4.2 shows the cluster visualization of the different algorithms applied to the data. Table 4.1 presents the number of subjects per cluster according to each clustering algorithm.

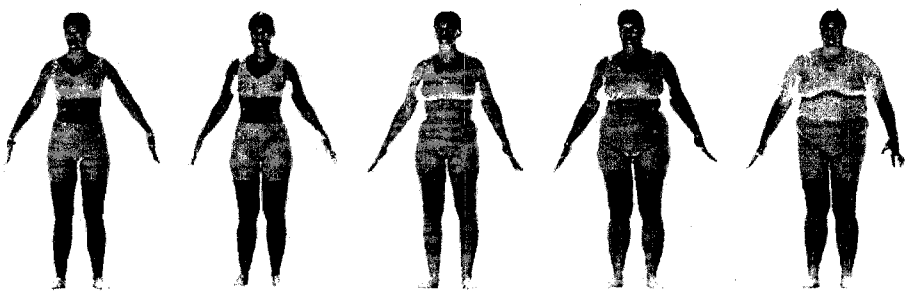
	Small	**Medium**	**Large**	**X-Large**	**XX-Large**
Farthest First	6 (2%)	211 (82%)	14 (5%)	20 (8%)	5 (2%)
EM	56 (22%)	82 (32%)	58 (23%)	55 (21%)	5 (2%)
K-means	53 (21%)	99 (39%)	60 (23%)	39 (15%)	5 (2%)
Density-based	43 (17%)	81 (32%)	65 (25%)	49 (19%)	18 (7%)

Table 4.1: Number of subjects per cluster for the American females.

We observe that the Farthest First algorithm produces very imbalanced clusters. From table 4.1 it may be seen that one cluster (Medium) includes more than 82% of the subjects, and the remaining data is divided into four clusters. The clusters produced by the Density-based algorithm with k-means components are balanced, yet the cluster in the middle is not well defined and is highly spread. The same situation is observed in the clusters defined by the EM algorithm. The clusters generated by k-means are compact and well defined, making k-means a good choice.

The k-means is a Centroid-based partitioning technique in which each cluster is represented by the mean value of the objects in the cluster. The Centroid is viewed as the cluster's center of gravity [Han and Kamber, 2006]. This algorithm is suitable for the discovery of convex clusters that are well separated from one another, in continuous n-dimensional data. The k-means algorithm is sensitive to noise or outliers, since such data may significantly influence the mean value. Recall that, the anthropometric measures were double checked by two anthropometric domain experts and an automatic detection of possible outliers was done during the data collection. This made the level of noise and the occurrence of outliers rare.

In order to validate the quality of the clusters produced by k-means, we verify the cluster membership through querying the 3-D body scans using the Cleopatra system. Table 4.2 shows some of the characteristics of the Centroids of the American female population. Figure 4.3 shows the 3-D body scans of the human subjects that correspond to these measurements, highlighting the difference in body types of the five clusters. Thus, by inspecting the mean values in table 4.2, the cluster distribution, and through the analysis of the results using Cleopatra, we observe that the anthropometric clusters discriminate between *Small*, *Medium*, *Large*, *X-Large* and *XX-Large* body sizes.



(a) (b) (c) (d) (e)

Figure 4.3: Cluster centroids for the American female population. (a) Small, (b) Medium, (c) Large, (d) X-Large and (e) XX-Large.

	Small	**Medium**	**Large**	**X-Large**	**XX-Large**
BustCircumference	87.7	88.8	94.4	105.4	127.1
	(5.8)	(4.6)	(4.5)	(6.9)	(9.6)
WaistCircumference	69.1	70.8	77.4	88.4	110.5
	(6.1)	(5.0)	(6.3)	(9.1)	(5.6)
HipCircumference	96.0	99.4	105.5	113.9	133.2
	(5.0)	(5.1)	(5.3)	(7.7)	(9.8)
ArmShouldertoWrist	53.4	57.1	60.5	56.9	63.6
	(2.0)	(1.7)	(2.3)	(1.7)	(2.7)
Stature	155.9	163.7	171.9	162.0	180.6
	(4.9)	(3.8)	(5.0)	(4.0)	(5.2)
ShoulderBreadth	41.2	42.4	44.5	46.8	55.4
	(1.7)	(1.8)	(1.8)	(2.2)	(4.1)
Weight (lbs)	119.0	130.4	154.7	178.4	278.9
	(12.5)	(10.8)	(11.9)	(22.3)	(23.7)
Num. of Subjects	53 (21%)	99 (39%)	60 (23%)	39 (15%)	5 (2%)

Table 4.2: Body measurements of American female cluster centroids as presented in [Paquet et al., 2007]. Shown are the means (in cm), the standard deviation in parenthesis, and the number of subjects on each cluster.

4.2.3 Italian and Dutch populations

Following the aforementioned methodology, we also analyze the Italian and Dutch populations. The cluster visualization of the different algorithms for these populations, and the number of subjects per cluster, are presented in the Appendix B. The results indicate that for the Italian and Dutch, both males and females, the best clustering is obtained using the *density-based* algorithm with k-means components. The clusters obtained with this algorithm are show in figure 4.4.

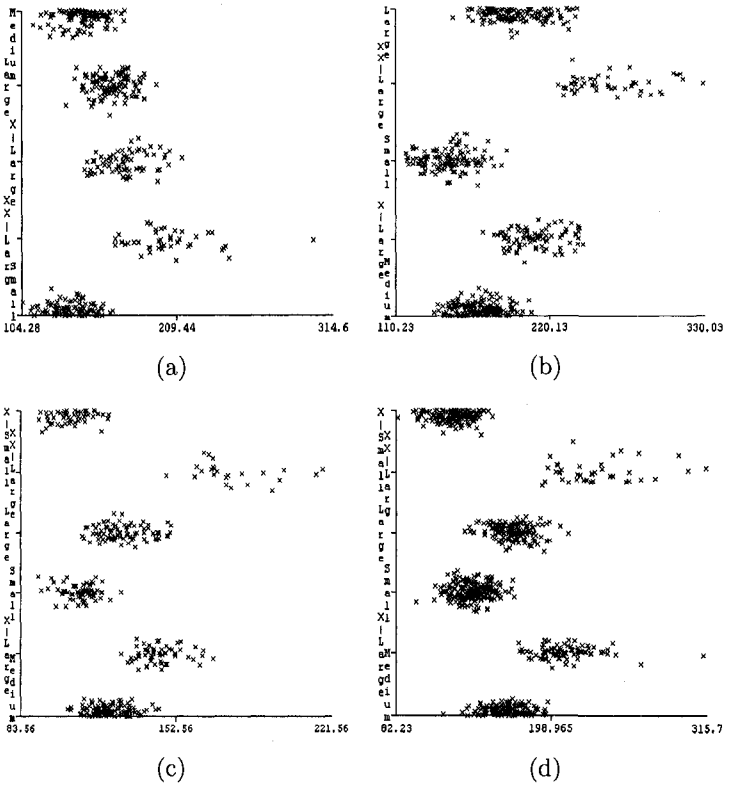


Figure 4.4: Cluster visualization for the Italian and Dutch populations. (a) Italian male population, (b) Dutch male population, (c) Italian female population, and (d) Dutch female population. The y axis represent the clusters, the x axis is the weight range.

As we discuss in section 4.2.1, for the Italian females and Dutch females, the population is best defined by six clusters. This introduces the need of determining the clothing/body sizes that correspond to these six clusters. To achieve this, we compare the results of the Italian and Dutch females with the results obtained for the American females [Paquet et al., 2007]. Table 4.3 shows the mean values of some measurements

that correspond to the smallest and largest centroids for the Italian and Dutch females, and the *Small* and *XX-Large* sizes for the American females.

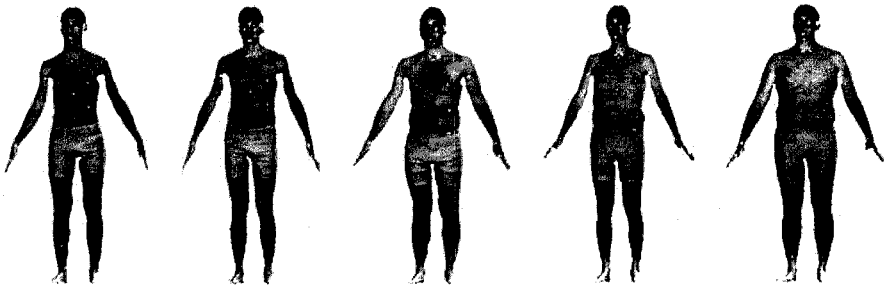
	American		Italian		Dutch	
	Small size	**XX-Large size**	**Smallest centroid**	**Largest centroid**	**Smallest centroid**	**Largest centroid**
BustCircumference	87.7	127.1	83.8	108.0	90.7	120.6
WaistCircumference	69.1	110.5	68.5	92.3	74.1	107.9
HipCircumference	96.0	133.2	91.4	114.6	97.2	121.6
Stature	155.9	180.6	152.7	160.9	160.0	176.9
Weight (lbs)	119.0	278.9	106.7	177.5	126.5	234.3

Table 4.3: Comparison of body measurements for females.

When considering the mean values for the Italian females in table 4.3, the next observations are noteworthy. First, the smallest centroid of the Italian is thinner and shorter than the *Small* American size centroid. Second, the *XX-Large* American centroid is considerably taller and more robust than the largest centroid of the Italian population. Furthermore, the range of the mean measurements for the Italian female population (except for the smallest centroid) is contained in the range defined by the mean measurements for the *Small* and *XX-Large* sizes of the American females. Therefore, we name the cluster of the smallest Italian female population as *X-Small.*

In the case of the Dutch female population, we observe that the Dutch smallest centroid is slightly more robust and taller than the centroid that corresponds to the American *Small* size. Nevertheless, the centroid of the *XX-Large* American size is significantly more robust and taller than the largest Dutch centroid. Thus, we decide to adopt the convention name introduced for the Italian female population and define the clusters for the Dutch female population from *X-Small* to *XX-Large.*

Again, in order to validate the quality of the clusters produced by the density-based algorithm with k-means components, we verify the cluster membership through querying the 3-D body scans using the Cleopatra system. The centroid characteristics for the Italian and Dutch are presented in tables 4.4 to 4.7. Figures 4.5 to 4.8 show the 3-D body scans of the human subjects that correspond to these anthropometric measurements, highlighting the difference in body types of the clusters. By inspecting the tables 4.4–4.7, the cluster distribution, and through the analysis of the results using the Cleopatra system, we observe that the clusters discriminate between the different body sizes.

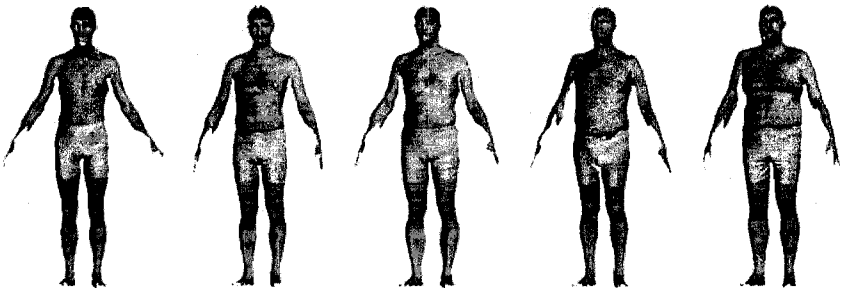


(a) (b) (c) (d) (e)

Figure 4.5: Cluster centroids for Italian male population. (a) Small, (b) Medium, (c) Large, (d) X-Large and (e) XX-Large.

	Small	**Medium**	**Large**	**X-Large**	**XX-Large**
ChestCircumference	90.3	91.2	98.8	98.9	108.4
	(5.1)	(4.6)	(5.6)	(5.9)	(7.1)
WaistCircumference	78.2	79.7	87.5	86.9	97.8
	(5.2)	(3.8)	(6.2)	(5.6)	(9.6)
HipCircumference	93.0	93.9	100.0	100.5	108.7
	(4.1)	(4.0)	(3.9)	(3.8)	(7.8)
NeckBaseCircumference	45.8	46.6	48.2	48.6	50.2
	(1.7)	(1.4)	(1.6)	(1.4)	(1.7)
ArmShoulderto Wrist	60.7	64.2	61.7	67.2	64.9
	(1.9)	(1.8)	(1.8)	(2.1)	(1.9)
Stature	166.8	175.3	170.3	182.6	176.8
	(4.2)	(3.8)	(4.1)	(5.0)	(4.3)
ShoulderBreadth	43.7	45.0	46.2	47.5	49.5
	(2.1)	(1.9)	(1.9)	(2.3)	(2.0)
Weight (lbs)	136.6	147.3	166.7	174.6	203.8
	(12.5)	(11.6)	(11.8)	(14.2)	(24.8)
Num. of Subjects	88 (21%)	107 (26%)	107 (26%)	69 (17%)	42 (10%)

Table 4.4: Body measurements of Italian male centroids. Shown are the means (in cm), the standard deviation in parenthesis, and the number of subjects on each cluster.

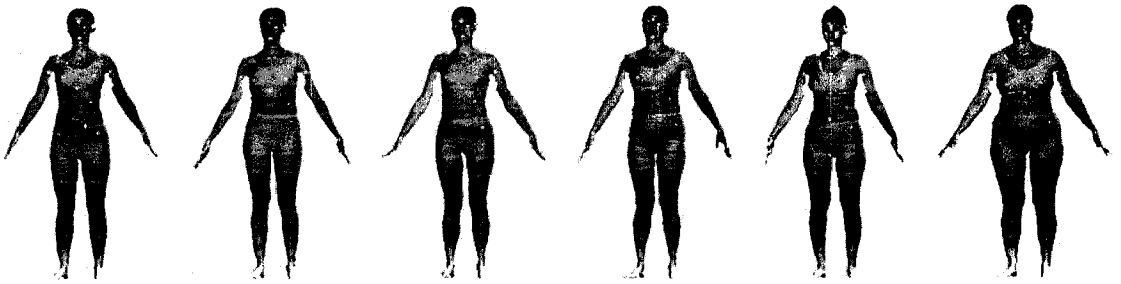


(a) (b) (c) (d) (e)

Figure 4.6: Cluster centroids for Dutch male population. (a) Small, (b) Medium, (c) Large, (d) X-Large and (e) XX-Large.

	Small	**Medium**	**Large**	**X-Large**	**XX-Large**
ChestCircumference	93.1	96.5	107.8	104.9	121.2
	(6.0)	(4.8)	(6.7)	(6.2)	(7.0)
WaistCircumference	82.5	86.5	96.3	97.4	112.5
	(6.9)	(5.7)	(7.4)	(5.7)	(9.1)
HipCircumference	95.3	98.5	103.1	108.1	116.3
	(4.6)	(3.6)	(4.5)	(4.0)	(6.2)
NeckBaseCircumference	45.4	47.4	50.3	51.0	55.2
	(2.4)	(2.5)	(2.7)	(2.3)	(3.4)
ArmShouldertoWrist	60.3	65.4	62.4	67.2	65.6
	(3.3)	(2.6)	(3.0)	(3.2)	(3.6)
Stature	173.0	184.8	176.1	193.2	188.1
	(6.6)	(4.6)	(5.6)	(6.2)	(8.3)
ShoulderBreadth	44.0	46.6	47.8	48.8	52.0
	(1.6)	(1.8)	(2.0)	(2.0)	(3.4)
Weight (lbs)	149.9	171.5	198.9	214.4	264.1
	(14.8)	(12.2)	(17.5)	(16.9)	(29.2)
Num. of Subjects	126 (22%)	173 (31%)	139 (25%)	82 (14%)	47 (8%)

Table 4.5: Body measurements of Dutch male centroids. Shown are the means (in cm), the standard deviation in parenthesis, and the number of subjects on each cluster.

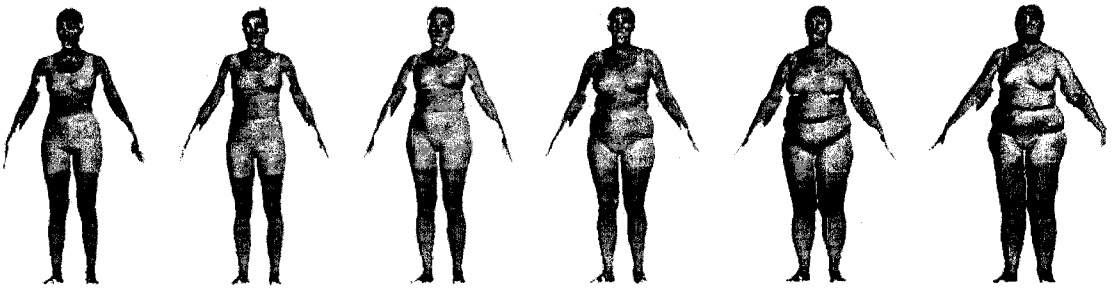


(a) (b) (c) (d) (e) (f)

Figure 4.7: Cluster centroids for Italian female population. (a) X-Small, (b) Small, (c) Medium, (d) Large, (e) X-Large and (f) XX-Large.

	X-Small	**Small**	**Medium**	**Large**	**X-Large**	**XX-Large**
BustCircumference	83.8	82.9	86.9	91.9	92.8	108.0
	(4.6)	(3.4)	(4.2)	(5.7)	(4.8)	(10.6)
WaistCircumference	68.5	70.8	74.2	76.8	78.8	92.3
	(5.5)	(4.8)	(4.7)	(5.8)	(5.4)	(9.9)
HipCircumference	91.4	92.4	96.4	99.1	102.5	114.6
	(3.9)	(3.6)	(3.7)	(4.0)	(4.1)	(5.0)
ArmShouldertoWrist	53.8	57.7	59.3	56.3	61.3	58.4
	(1.4)	(1.2)	(1.8)	(1.5)	(2.1)	(2.1)
Stature	152.7	159.2	164.8	157.1	169.3	160.9
	(3.3)	(2.6)	(2.8)	(3.2)	(4.3)	(4.7)
ShoulderBreadth	38.2	39.0	40.3	41.0	42.2	45.0
	(1.6)	(1.6)	(1.5)	(1.8)	(1.6)	(2.1)
Weight (lbs)	106.7	110.2	124.2	129.5	143.9	177.5
	(8.2)	(6.1)	(8.5)	(9.8)	(9.7)	(18.5)
Num. of Subjects	51 (13%)	65 (17%)	110 (28%)	82 (21%)	56 (14%)	24 (6%)

Table 4.6: Body measurements of Italian female centroids. Shown are the means (in cm), the standard deviation in parenthesis, and the number of subjects on each cluster.



(a) (b) (c) (d) (e) (f)

Figure 4.8: Cluster centroids for Dutch female population. (a) X-Small, (b) Small, (c) Medium, (d) Large, (e) X-Large and (f) XX-Large.

	X-Small	**Small**	**Medium**	**Large**	**X-Large**	**XX-Large**
BustCircumference	90.7	92.1	98.4	106.1	117.5	120.6
	(5.9)	(5.0)	(6.2)	(6.7)	(8.3)	(9.2)
WaistCircumference	74.1	76.4	83.5	90.7	103.3	107.9
	(6.4)	(5.9)	(7.1)	(8.0)	(9.4)	(11.0)
HipCircumference	97.2	101.2	106.4	110.1	117.5	121.6
	(5.1)	(5.4)	(5.1)	(5.8)	(7.7)	(10.4)
ArmShouldertoWrist	55.1	58.5	62.0	56.8	58.6	63.0
	(2.4)	(1.8)	(2.2)	(1.9)	(2.2)	(2.0)
Stature	160.0	169.7	176.9	162.4	167.1	176.9
	(4.7)	(3.7)	(4.8)	(5.1)	(4.3)	(5.6)
ShoulderBreadth	40.3	41.7	44.0	43.6	47.0	47.6
	(1.7)	(1.7)	(2.1)	(2.4)	(2.9)	(2.9)
Weight (lbs)	126.5	141.1	165.3	171.1	210.0	234.3
	(13.1)	(12.4)	(14.5)	(13.3)	(21.8)	(28.9)
Num. of Subjects	130(19%)	198(28%)	125(18%)	125(18%)	83(12%)	39(6%)

Table 4.7: Body measurements of Dutch female centroids. Shown are the means (in cm), the standard deviation in parenthesis, and the number of subjects on each cluster.

4.3 Analysis of the three populations

Using the anthropometric clustering results, we aim to analyze the three populations of our study, and discuss some similarities and differences for both male and female populations. In our analysis, we consider the centroids or archetypes, since these are representatives of the other subjects that belongs to the same cluster.

In the analysis of the male population we also consider the results for the American male population as presented in [Viktor et al., 2006]. Figure 4.9 shows some body measurements and indicates how the three male populations are positioned in the measurement range.

When considering the data in tables 4.4 and 4.5, and figure 4.9, the next observations are worth mentioning. Even though, the three populations are described by five sizes, these are not comparable, per se, to one another. For example, if we consider the *Small* size from American, Italian and Dutch, we observe the Italians are considerably thinner, while both the American and Dutch tend to be more robust. In general we observe the Italians are thinner than the American and Dutch. It may be seen also that the tallest population is the Dutch, while the shortest population is the Italian. Moreover, in the three populations, the *XX-Large* subjects are shorter than the *X-Large* subjects. This is an important feature to consider when designing, for example, pants; the legs should have short lengths for the *XX-Large* size. Furthermore, if we examine the measurements and height of the American and Dutch archetypes of the corresponding sizes, we may observe the range of measurements are similar but in general the Dutch are taller, making the Americans the most robust population.

Regarding the female population, even more interesting aspects are observed. Figure 4.10 shows how the three populations are distributed when considering some body measurements. Since the number of sizes for the Americans is five, while for the Italians and Dutch are six, again we cannot compare them directly. However, from tables 4.2, 4.6 and 4.7, and figure 4.10, it may be observed that the Italians are the thinnest and shortest population. We also observe that the Dutch *X-Small* are more robust and taller than the American *Small*; the Dutch *Small* have wider chests and hips, and are taller than the American *Medium*. If we continue in this way, it follows that the Dutch *XX-Large* are taller and more robust than the American *XX-Large*, but surprisingly, the American *XX-Large* are taller and more robust than the Dutch *XX-Large*. Moreover, we notice some relationships among the different sizes of the three populations. For instance, we notice the body measurements that correspond to the Dutch *Large*, American *X-Large*

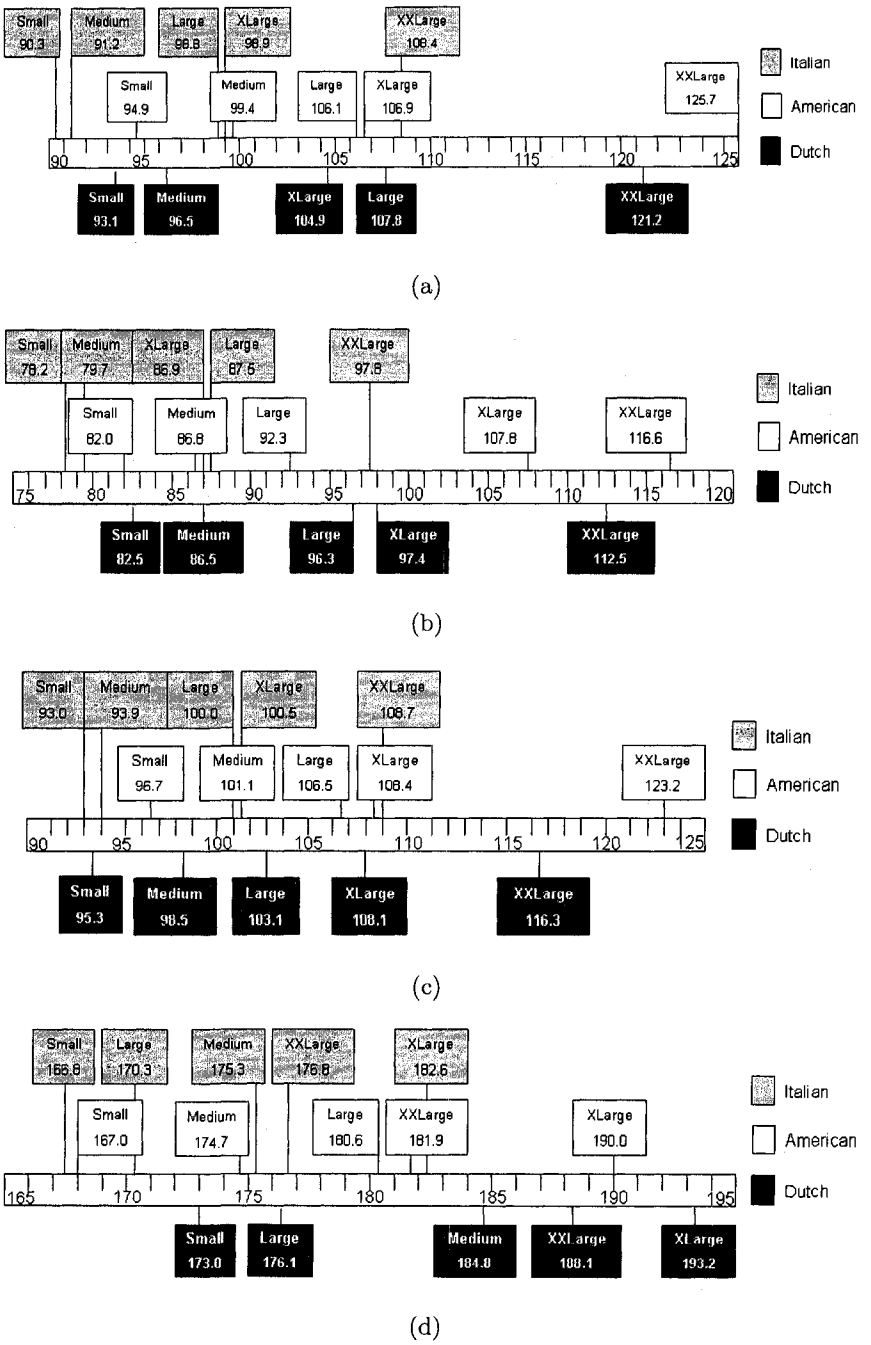
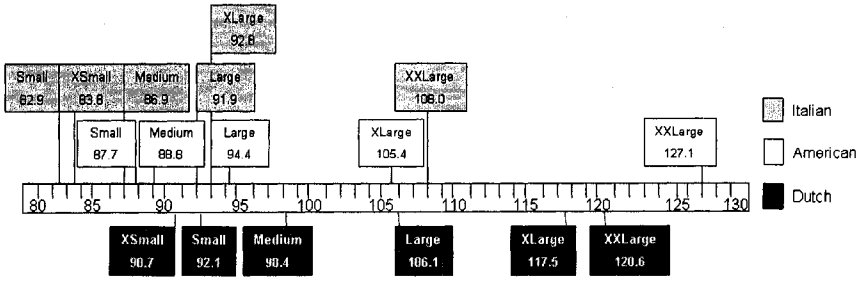
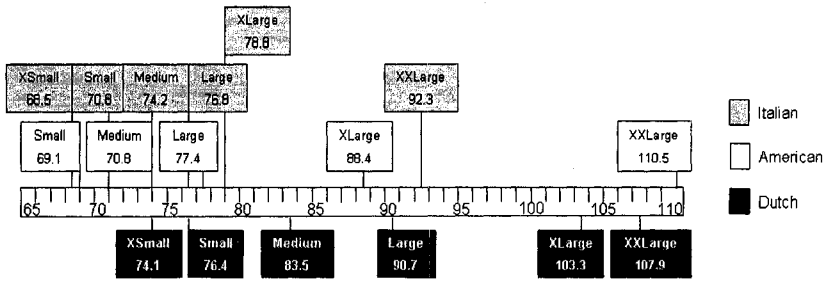


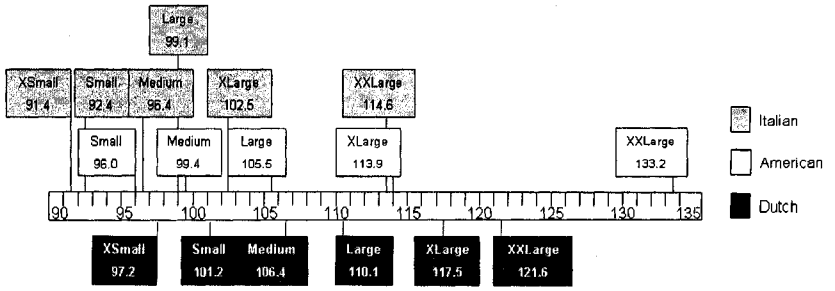
Figure 4.9: Analysis of the male population. Figure shows the mean values corresponding to each centroid or archetype according to: (a) Chest Circumference, (b) Waist Circumference, (c) Hip Circumference and (d) Stature.



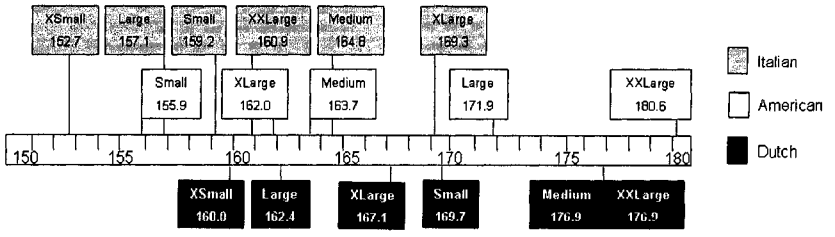
(a)



(b)



(c)



(d)

Figure 4.10: Analysis of the female population. Figure shows the mean values corresponding to each centroid or archetype according to: (a) Bust Circumference, (b) Waist Circumference, (c) Hip Circumference and (d) Stature.

and Italian *XX-Large* are very similar. Furthermore, the size *Small* of the Dutch is comparable to the *X-Large* size of the Italian. The Dutch seems to be larger than the Americans for the smaller sizes, but the American *XX-Large* resulted to be the tallest and most robust of all populations, as may be observed from figures 4.3 to 4.8.

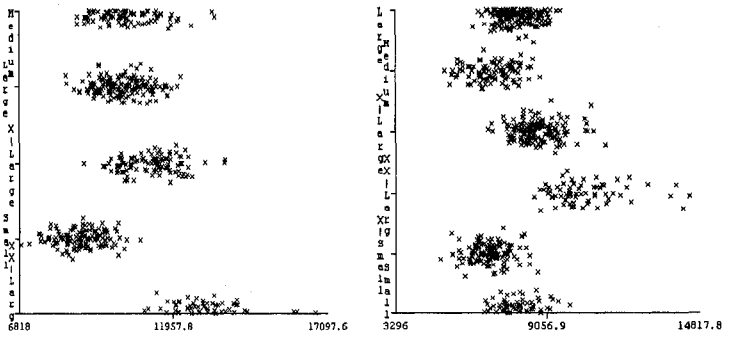
4.4 Clustering of 3-D data

The analysis performed so far is based on anthropometric measurements contained in the CAESAR^{TM} database. However, this database also contains detailed 3-D shape information of each individual. This allows the analysis of the features of the individuals based on 3-D shape information. Following the same approach as with the anthropometric data, we perform cluster analysis on the 3-D data. In this analysis, we consider only the Dutch dataset since, as shown in Chapter 3, the best sampling was achieved for this population. Moreover, for the Italian population there is a substantial percentage of missing 3-D body scans.

3-D scans are represented by a set of three histograms that contain a 3-D shape index for the human body [Brunsman et al., 1997]. Since the 3-D data is normalized and we want the height of the subjects to be considered in the clustering, we first de-normalize the data. Again, the data is separated into two sets based on the gender of the subject. The resulting sets consist of 542 males and 663 females.

Again, we determine the minimum number of clusters that best describe the population following the process described in section 4.2.1. The results indicate that the male population is best described using five clusters. For the female population the best results are obtained with six clusters. Then, the number of clusters was set to five for the males and six for the females. We use the same algorithm that was found to produce the best results for the anthropometric data so the results are consistent and we may be able to compare them. Thus, the density-based algorithm with k-means components is used for both, males and females. The visualization of the resulting clusters for the male and female population are shown in figure 4.11. The number of subjects per cluster is given in tables 4.8 and 4.9.

Again, we verify the cluster membership through querying the 3-D scans using the Cleopatra system. The 3-D body scans of the centroids of the male and female populations are shown in figures 4.12 and 4.13 respectively. From the figures, it may be observed that the centroids highlight the difference in body types of the different clusters. Thus, the results indicate that the 3-D clusters distinguish between the different body sizes.



(a) (b)

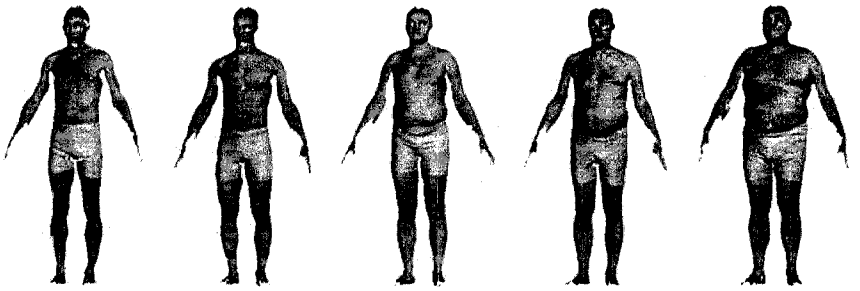
Figure 4.11: Cluster visualization of 3-D data for the Dutch population. (a) Clustering of the male population and (b) clustering of the females population. The y axis represent the clusters, the x axis is an attribute describing the 3-D shape.

Size	**Num. of Subjects**
Small	135 (25%)
Medium	79 (15%)
Large	166 (31%)
X-Large	106 (20%)
XX-Large	56 (10%)

Table 4.8: Number of subjects per cluster for Dutch males.

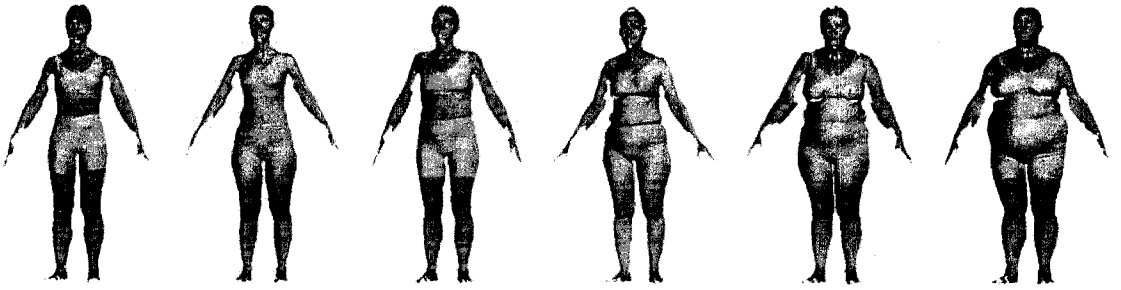
Size	**Num. of Subjects**
X-Small	132 (20%)
Small	63 (10%)
Medium	103 (16%)
Large	149 (22%)
X-Large	147 (22%)
XX-Large	69 (10%)

Table 4.9: Number of subjects per cluster for Dutch females.



(a) (b) (c) (d) (e)

Figure 4.12: Cluster centroids from 3-D data for Dutch males. (a) Small, (b) Medium, (c) Large, (d) X-Large and (e) XX-Large.



(a) (b) (c) (d) (e) (f)

Figure 4.13: Cluster centroids from 3-D data for Dutch females. (a) X-Small, (b) Small, (c) Medium, (d) Large, (e) X-Large and (f) XX-Large.

4.5 Anthropometric versus 3-D data clustering

In previous sections we present the results of clustering the anthropometric data and 3-D data. In this section we present a comparative analysis of the clustering results obtained from the anthropometric measurements and the results obtained from the 3-D data.

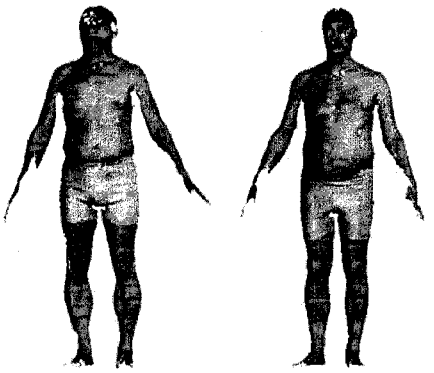
By inspecting the number of subjects per cluster in tables 4.5, 4.7, 4.8 and 4.9, we notice the number of subjects is not the same when the clustering is performed using the anthropometric measurements and the 3-D data.

Actually, we compare, for each subject, the subject's size from the anthropometric data against the subject's size from the 3-D data. We found that 342 subjects (63.1% of the male population) and 523 subjects (78.9% of the female population) change in size. Some subjects increase or decrease one, two, three or even four sizes in the case of the female population. Here, we proceed to analyze the anthropometric measurements and 3-D scans information of a set of representative individuals in our attempt to find the reason for this situation.

In our analysis we consider the anthropometric clustering results as reference to determine if a subject increased or decreased in size. We say a subject increase two sizes, if for instance according to the clustering of anthropometric data, the subject's size is *Small*, but according to the clustering results of the 3-D data, the subject's size is *Large*. We mainly focus on the subjects that have a significant change in size, that is, the subjects that increase or decrease three and four sizes. Next we present some examples of this situation.

Figure 4.14 shows the 3-D body scan of a subject whose size from the anthropometric measurements is *Small*. However, according to the 3-D data, his size is *X-Large*. That is, the subject increases three sizes. Together with the subject's 3-D body scan we show the centroid of the *X-Large* size as obtained from the 3-D data. We also present the subject's anthropometric measurements in table 4.10.

We observe that the body shape similarity between the subject and the *X-Large* centroid is high. This explains why, according to the 3-D data, the subject is considered as *X-Large*. However, by inspecting the body measurements and comparing them with the anthropometric centroids in table 4.5, we notice the body measurements correspond to the *Small* size. This makes sense from the clustering point of view. The clustering of 3-D data is based on body shape similarity while the anthropometric clustering is based on the similarity of the body measurements.



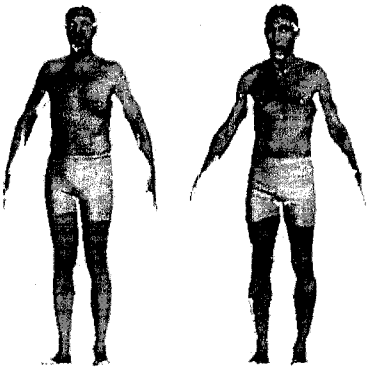
(a) (b)

Figure 4.14: Shape comparison with the X-Large centroid. (a) the subject that increases three sizes and (b) the X-Large centroid from 3-D data.

Measurement	**Value (cm)**
ChestCircumference	96.6
WaistCircumference	88.5
HipCircumference	97.5
NeckBaseCircumference	49.0
ArmShouldertoWrist	61.3
Stature	169.2
ShoulderBreadth	43.6
Weight (lbs)	161.4

Table 4.10: Anthropometric measurements of the subject in figure 4.14(a).

The next example shows the opposite situation, where a subject decreases three sizes. According to the anthropometric clustering the subject's size is *X-Large*. Nevertheless, according to the results of the 3-D clustering the subject is *Small* size. Figure 4.15 show the subject's 3-D body scan and the *Small* centroid from 3-D data clustering. The subject's anthropometric measurements are provided in table 4.11.



(a) (b)

Figure 4.15: Shape comparison with the Small centroid. (a) the subject that decreases three sizes and (b) the Small centroid from 3-D data.

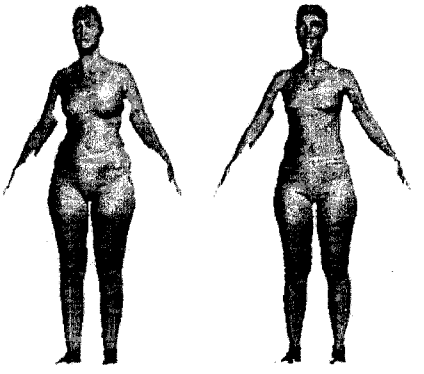
Measurement	**Value (cm)**
ChestCircumference	103.0
WaistCircumference	94.1
HipCircumference	101.7
NeckBaseCircumference	49.5
ArmShouldertoWrist	66.9
Stature	195.6
ShoulderBreadth	47.7
Weight (lbs)	198.4

Table 4.11: Anthropometric measurements of the subject in figure 4.15(a).

From figure 4.15, it may be seen that the subject's body shape is very similar to the body shape of the *Small* centroid, and therefore the subject is grouped into the *Small* cluster. When we inspect the subject's anthropometric measurements, we observe these

correspond to the *X-Large* anthropometric size. Again, this is explained by the fact that the clustering of the 3-D data groups together similar body shapes.

The next two examples show female subjects that increase or decrease four sizes. In the following example we show a subject that decreases four sizes. That is, when considering the anthropometric measurements the subject size is *XX-Large*, while from the 3-D data the subject's size is *Small*. Again, we compare the subject's body shape with the *Small* centroid in figure 4.16. The subject's anthropometric body measurements are shown in table 4.12.



(a) (b)

Figure 4.16: Shape comparison with the Small centroid. (a) the subject that decreases four sizes and (b) the Small centroid from 3-D data.

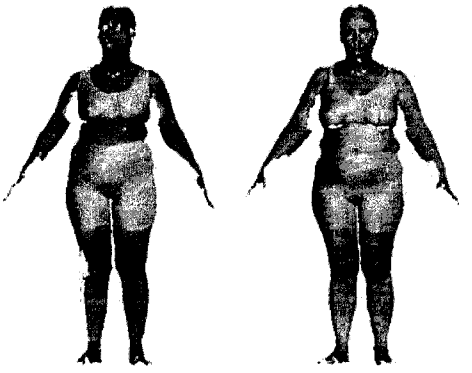
Measurement	**Value (cm)**
BustCircumference	104.6
BustCircumferenceUnderBust	87.8
WaistCircumference	98.4
HipCircumference	122.8
ArmShouldertoWrist	65.4
Stature	194.8
ShoulderBreadth	45.1
Weight (lbs)	228.0

Table 4.12: Anthropometric measurements of the subject in figure 4.16(a).

Again, it may be seen that the body shapes are very similar, but the body measure-

ments correspond to the *XX-Large* anthropometric size.

Finally, in the next example, we show the case where a subject increased four sizes. From anthropometric measurements the subject's size is *X-Small*, while according to the 3-D data the size of the subject is *X-Large*. The subject's anthropometric measurements are shown in table 4.13.



(a) (b)

Figure 4.17: Shape comparison with the X-Large centroid. (a) the subject that increases four sizes and (b) the X-Large centroid from 3-D data.

Measurement	**Value (cm)**
BustCircumference	94.2
BustCircumferenceUnderBust	72.5
WaistCircumference	75.4
HipCircumference	102.7
ArmShouldertoWrist	52.7
Stature	152.0
ShoulderBreadth	42.9
Weight (lbs)	137.1

Table 4.13: Anthropometric measurements of the subject in figure 4.17(a).

From table 4.13 it may be observed that the subject's anthropometric measurements satisfy the anthropometric *X-Small* size, but for the 3-D data clustering, the shape is very similar to the *X-Large* centroid.

By analyzing the previous examples, the following observations are worth mentioning. The clustering of 3-D body scans groups together subjects with similar body shape. On the other hand, when the clustering is performed on the anthropometric data, the algorithm groups in the same cluster the subjects with similar height, weight, bust circumference and so on. The change in size is due to the fact that even though two subjects may have similar body measurements, their body shape may be different. For example, two female subjects may have the same bust circumference. However, one of the subjects has a wide back and shoulders, while the other has a narrow back and large bust, making their body shapes different. Thus, the clustering of the 3-D data and the clustering of anthropometric data provide a different perspective on the subject characterization. As a result, we conclude the clustering results are complementary to one another.

In the clothing industry, however, the design and manufacture of garments is based on body measurements. Therefore, from the application point of view, we adopt the anthropometric measurements clustering, due to the direct application to tailoring and garments design. The clustering of 3-D data may be useful when designing e.g. masks for protection against hazardous materials, custom made products such as artificial legs or helmets, where the goal is to produce good fitting for different shapes.

4.6 Chapter summary

This chapter showed our results when aiming to characterize the three different population of our study. To achieve this objective, we performed cluster analysis using first, the anthropometric body measurements. Next, we identified the cluster archetypes, i.e. the subjects that represent all other subjects belonging to the same cluster. Using our experimental results, we analyzed the three populations. In the analysis, we observed that the thinnest and shortest population are the Italians, the Dutch are the tallest population, and the Americans are the most robust population. We also noticed that for the three male populations, the *XX-Large* subjects are shorter than the *X-Large* subjects.

Following the same approach we applied to the anthropometric data, we performed clustering of the 3-D data, and again identified the archetypes. When comparing the clustering result of the anthropometric data with the results of 3-D data, we observed the results are complementary, in the sense that each presents a different perspective of the subjects.

In the next chapter, we use the clusters of body sizes as class labels, and perform *Classification*. The CAESAR^{TM} database resides in an IBM DB2 relational database

and we intend to use this structure to *directly* mine the data, without flattening the database. To this end, we use a *multi-view relational classification* approach. This allow us to identify classification rules that may be used as constraints when designing garments.

Chapter 5

Describing the relationships between body measurements

The cluster analysis of anthropometric measurements presented in Chapter 4 allows us to define clusters of body sizes. In this chapter we use the clusters of body sizes as class labels, and perform *Multi-View Classification*. *Multi-View Classification* allows the *direct* and *transparent* mining of multi-relational data using single-table learning algorithms. Next, we aim to find general rules that describe the relationships between body measurements for the different sizes. We show these rules may be applied in the industry as constraints for the design and manufacture of clothes.

5.1 Multi-view relational classification

Classification is a supervised learning data mining technique used both to understand the existing data and to predict future data trends. The input data for a classification task is a collection of tuples or examples. Each example is characterized by a set of attributes and a special attribute designated as the *class label.* Each example belongs to a predefined class, as determined by the class label. The goal of classification is to build a model that best fits the relationship between the attribute set and a class label. The model is constructed by analyzing the examples described by the attribute set. Typically, the learned model is represented in the form of classification rules, decision trees, or a mathematical formula. The accuracy of the model or *classifier* is the percentage of the test examples that are correctly classified by the model [Han and Kamber, 2006, Tan et al., 2005].

Most classification methods utilize *flat* data representations, that is, a universal relation containing a number of attributes and tuples. However, in many scenarios the data resides in relational databases. The CAESAR^{TM} data is contained in an IBM DB2 relational database. Figure 5.1 shows the CAESAR^{TM} database consisting of six relational tables, with the *BodySize* class label as an attribute in the *Top_measurements* table.

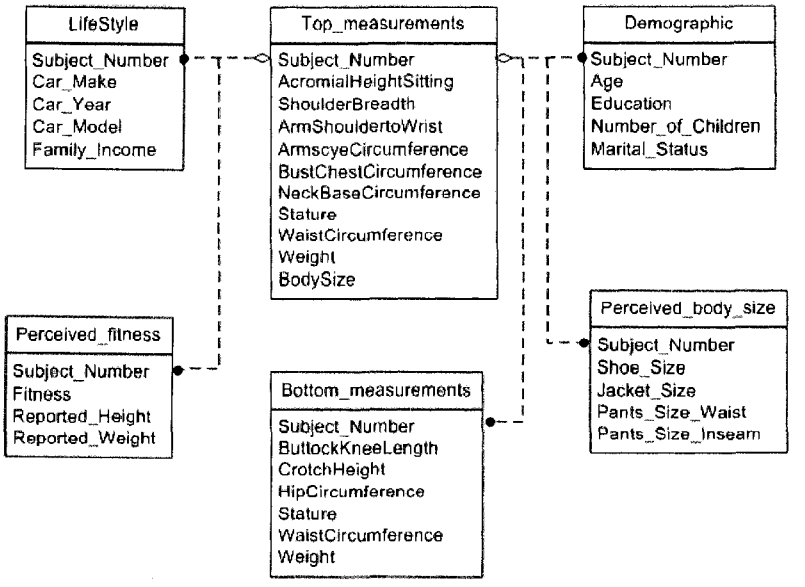


Figure 5.1: The Entity Relationship diagram for the CAESAR^{TM} database as presented in [Viktor et al., 2006].

Transforming relational data into the required flat format has several disadvantages: converting the data into a flat format increases the time and complexity of the mining process. Another drawback is that, not only the data lose their compact representation, but also that this process may lead to loss of semantic information [Dzeroski and Lavrac, 2001] and may create a large amount of redundant data. Moreover, a universal relation with a large number of attributes and tuples usually implies reduction of the efficiency and scalability of the mining process.

Our goal is to provide an easy and intuitive way to analyze the CAESAR^{TM} data for clothing designers (who are not necessarily data miners). Therefore, we aim to *directly* and *transparently* mine the CAESAR^{TM} relational database utilizing this structure. That is, without flattening or propositionalizing [Krogel and Wrobel, 2003] the database into a universal relation. To this end, we follow a *Multi-view relational classification* approach.

Multi-view relational classification [Guo and Viktor, 2005] is based on a multi-view learning framework. The multi-view setting [Blum and Mitchell, 1998] applies to learning problems that have a natural way to divide their features into subsets or *views*. Each view of the data may be expressed in terms of a set of disjoint features that are sufficient to learn the target concept. The target concept is learned independently on each different view using the features present. The results from the views are then combined, each contributing to the learning task. For example, we may classify segments of televised broadcast based either on the video or on the audio information independently.

A multi-view learning problem with n *views* may be seen as n inter-dependent relations, and are therefore applicable to multi-relational learning [Guo and Viktor, 2005]. Mining multi-relational databases involves learning patterns from multiple relations. Each relation or table is then considered as a view. One of these tables is considered the target and the remaining act as background relations. Usually, each of these relations has a natural disjoint feature set that contribute to the target concept to be learned.

In the case of the CAESAR^{TM} database each relation has a natural disjoint feature set. That is, each relation described different aspects of a specific person. For instance, the *LifeStyle* relation contains subject's income and automotive details while the *Top_measurements* and *Demographics* relations include the subject's top measurements, and subject's demographics such as age, education, etc. respectively.

The multi-view classification algorithm [Paquet et al., 2007, Guo and Viktor, 2005] consists of two main stages: (1) the *target concept propagation*, (2) *multi-view learning.*

Input: Relational database $\Re = R_t \times R_{b_1} \times \cdots \times R_{b_n}$,
Multi-view learner $\mathcal{L}$, Meta-learner $\mathcal{M}$,
Maximum length of joint path *MaxJ.*

Output: Classification model $\mathcal{F}$.

1: Convert database schema $\Re$ into graph;
2: Extract join path set $\{\bowtie_n\}$ from graph;
3: Construct relational feature set $\varphi(t)$ for each join path in $\{\bowtie_n\}$; forming candidate view set $\{V_d^1, \cdots, V_d^n\}$;
4: Select a set $\{\mathcal{V}^i\}_1^{n'}$ from $\{V_d^1, \cdots, V_d^n\}$;
5: Train $\mathcal{L}$ with $\{\mathcal{V}^i\}_1^{n'}$, forming hypothesis set $\{\mathcal{H}^i\}_1^{n'}$;
6: Form final model $\mathcal{F}$ by combining $\{\mathcal{H}^i\}_1^{n'}$, using $\mathcal{M}$;
7: return $\mathcal{F}$.

During the target concept propagation stage (steps 1-4), the training datasets of the multiple views are constructed. Each background relation $R_{b_1}, \cdots, R_{b_n}$ obtains the classes from the target relation R_t. If there is any aggregation information it is also propagated, but note the CAESAR^{TM} database is a special case where the tables contain only one-to-one relationships. Next, during the multi-view learning stage (steps 5-7) each of the views is used to construct various hypotheses on the target concept, this is accomplished using traditional single-table learning algorithms. Finally, the model is constructed by combining the different hypotheses.

To illustrate this process let us consider the construction of the view for the *LifeStyle* table. After the target concept propagation stage, both *Top_measurements* and *LifeStyle* relations contain the class label to be learned, namely, *BodySize*. As a result, from each individual relation it is possible to construct hypotheses for the target concept. However, the *Top_measurements* relation only contains information about the subject's top body measurements. Then, hypotheses produced by this relation only reflect how measurements such as waist circumference, shoulder breadth, stature, etc. are related to the class attribute *BodySize*. On the other hand, the *LifeStyle* relation contains income and automotive details of the subject. Therefore, the classification model built using this relation contains the relationships between the *BodySize* and subject's income and car information. In the last step, these hypotheses are combined to form the final model.

5.2 Most significant rules

For our multi-view classification experimentation, we consider three different classification learners: *RIPPER*, *C4.5* and *PART*. *RIPPER* (Repeated Incremental Pruning to Produce Error Reduction) is a widely used inductive rule learner proposed by William W. Cohen [Cohen, 1995]. This algorithm scales almost linearly with the number of the training tuples, and it is suitable for building models from datasets with imbalanced class distributions. RIPPER works well with noisy datasets because it uses a validation set to prevent model overfitting. Rules produced by this algorithm have two desirable features: good generalization accuracy and concise conditions.

C4.5 is an algorithm introduced by Quinlan [Quinlan, 1993] for inducing decision trees from a set of training data. This algorithm builds decision trees using an extension of information gain, known as gain ratio. C4.5 is an extension of ID3 [Quinlan, 1986] to handle both continuous and discrete attributes, data with missing values and pruning of the trees. Decision trees may easily be expressed as if-then rules. Each path from the

root of the tree to a leaf gives the conditions that must be satisfied if an instance is to be classified by that leaf.

PART is a simple but efficient rule learner that constructs a decision list by repeatedly generating partial decision trees. In contrast to RIPPER and C4.5, this algorithm learns one rule at a time, removes the instances covered by that rule, and iteratively induces further rules for the remaining instances. To produce a single rule, a pruned decision tree is built for the current set of instances, the leaf with the largest coverage is made into a rule, and the tree is discarded [Frank and Witten, 1998].

In the experiments we used 10 fold cross validation to test the accuracy of the classification models. We executed the algorithms using the *Top_ and Bottom_measurements* as target relations. The accuracies and the number of rules produced by PART, RIPPER and C4.5 for the female and male populations are shown in tables 5.1 and 5.2 respectively.

Learning task		**PART**	**RIPPER**	**C4.5**
Top and Bottom measurements	American*	75.8% (10)	76.6% (9)	76.2% (17)
	Italian	78.1% (11)	75.8% (13)	76.8% (29)
	Dutch	81.1% (17)	77.9% (18)	77.4% (44)

* Results as presented in [Paquet et al., 2007].

Table 5.1: Accuracy and number of rules obtained from the multi-view classification for females.

Learning task		**PART**	**RIPPER**	**C4.5**
Top and Bottom measurements	American*	78.7% (21)	79.2% (12)	80.1% (30)
	Italian	73.6% (14)	74.3% (13)	73.9% (29)
	Dutch	80.3% (17)	80.6% (14)	78.4% (37)

* Results as presented in [Viktor et al., 2006].

Table 5.2: Accuracy and number of rules obtained from the multi-view classification for males.

The tables show the results are consistent in terms of accuracy and number of rules. Notice the number of rules is specially small in the rulesets produced by RIPPER and PART. Since we aim to obtain general and reliable rules for application in the industry, we then identify the rules with high positive coverage and low negative coverage. Coverage

refers to the number of tuples or examples that satisfy the condition described by the rule. For example, let us consider the rule ($Stature \leq 170.9$) AND ($ChestCircumference \leq 100.2$): Small, with coverage (54/3). This indicates that in the dataset, 54 of the Small subjects satisfy the rule, while there are 3 non-Small sized subjects that also satisfy the rule. Therefore, this rule has high positive coverage and low negative coverage. The percentage of positive coverage is the number of subjects that satisfy the rule, divided by the total number of subjects that belong to the same class.

Tables 5.3 to 5.7 show the most significant rules for each body size in terms of both, the coverage and the information provided by the rule.

Learning Top and Bottom measurements	**Class**	**Coverage**	**Positive coverage(%)**	**Learner**
($ArmLengthSpinetoWrist \leq 74.5$) AND ($VerticalTrunkCircumference \leq 152.5$)	Small	(52/6)	98.1	RIPPER
($Weight \leq 146.0$) AND ($ArmLengthSpinetoWrist > 74.2$) AND ($Stature \leq 172.4$) AND ($SpinetoElbow \leq 52.4$) AND ($ShoulderBreadth > 38.5$) AND ($NeckBaseCircumference > 37.9$)	Medium	(75/1)	75.8	C4.5
($KneeHeightSitting > 52.6$) AND ($NeckBaseCircumference \leq 44.3$) AND ($ChestGirthatScye > 83.5$)	Large	(49/1)	81.7	PART
($SubscapularSkinfold > 2.7$) AND ($SpinetoElbow \leq 54.5$) AND ($ArmLengthSpinetoWrist \leq 80.3$) AND ($Weight > 145.0$)	X Large	(30/0)	77.0	PART

Note: Small number of XX-Large subjects lead to low coverage rules.

Table 5.3: High coverage rules for each clothing/body size for American females as presented in [Paquet et al., 2007].

Learning Top and Bottom measurements	**Class**	**Coverage**	**Positive coverage(%)**	**Learner**
($Stature \leq 162.0$) AND ($Weight \leq 119.5$) AND ($ArmLengthShouldertoWrist \leq 55.8$) AND ($TTR2 \leq 71.0$)	X Small	(48/1)	94.1	C4.5
($Weight \leq 119.5$) AND ($Stature \leq 164.4$) AND ($TTR2 \leq 74.8$) AND ($HipCircumference \leq 98.6$) AND ($ChestCircumferenceunderBust \leq 74.0$) AND ($HipCircMaxHeight > 71.5$)	Small	(63/1)	96.9	PART
($114.0 < Weight \leq 135.4$) AND ($Stature > 162.0$) AND ($ArmscyeCircumference > 33.4$) AND ($WaistHeight \leq 104.0$) AND ($ArmLengthShouldertoWrist > 56.3$)	Medium	(80/1)	72.7	C4.5
($CrotchHeight \leq 76.5$) AND ($ArmLengthSpinetoWrist \leq 76.7$) AND ($TotalCrotchLength > 49.2$) AND ($Stature > 148.5$) AND ($HeadBreadth > 13.8$) AND ($ThighCircumference \leq 62.4$)	Large	(74/1)	90.2	PART
($CrotchHeight > 72.8$) AND ($TTR3 > 72.6$) AND ($VerticalTrunkCircumference > 148.8$) AND ($FootLength > 23.4$) AND ($ThighCircumference > 53.5$) AND ($ChestCircumferenceunderBust > 73.6$)	X Large	(53/0)	94.6	PART
($Weight \geq 161.4$)	XX Large	(23/2)	95.8	RIPPER

Table 5.4: High coverage rules for each clothing/body size for Italian females.

Learning Top and Bottom measurements	**Class**	**Coverage**	**Positive coverage(%)**	**Learner**
($Weight \leq 145.1$) AND ($Stature \leq 166.1$) AND ($TTR1 \leq 72.3$)	X Small	(95/1)	73.1	C4.5
($Weight \leq 157.4$) AND ($Stature > 166.0$) AND ($TTR2 \leq 78.3$) AND ($TTR1 > 70.3$) AND ($CrotchHeight > 75.3$) AND ($SpinetoElbow \leq 55.1$)	Small	(140/2)	70.7	PART
($150.0 < Weight \leq 194.0$) AND ($Stature > 169.7$) AND ($TTR2 > 73.8$) AND ($ArmLengthSpinetoWrist > 78.8$) AND ($SubscapularSkinfold \leq 3.8$)	Medium	(87/1)	69.6	PART
($BustChestCircumference > 100.1$) AND ($ArmLengthShoulderto Wrist \leq 61.0$) AND ($AnkleCircumference \leq 26.5$) AND ($NeckBaseCircumference > 41.7$) AND ($Stature \leq 170.7$) AND ($TTR2 > 66.4$) AND ($ArmLengthSpinetoWrist \leq 78.9$)	Large	(88/0)	70.4	PART
($Weight > 189.4$) AND ($SpinetoElbow \leq 56.3$) AND ($93 < WaistHeight \leq 110.6$) AND ($ThumbTipReach > 72.3$)	X Large	(73/1)	88.0	PART
($Weight > 189.4$) AND ($ThumbTipReach > 79.6$) AND ($ArmLengthSpinetoWrist > 79.8$)	XX Large	(34/1)	87.2	C4.5

Table 5.5: High coverage rules for each clothing/body size for Dutch females.

Learning Top and Bottom measurements	**Class**	**Coverage**	**Positive coverage(%)**	**Learner**
($Weight \leq 157.0$) AND ($TTR3 \leq 77.5$) AND ($ArmLengthSpinetoWrist \leq 84.2$) AND ($BizygomaticBreadth \leq 14.5$) AND ($FootLength \leq 26.1$)	Small	(68/0)	77.3	PART
($Weight \leq 157.0$) AND ($TTR3 > 77.5$) AND ($KneeHeight > 53.4$) AND ($ArmLengthSpinetoWrist \leq 88.5$) AND ($TTR1 \leq 85.4$) AND ($HipCircumference \leq 99.3$)	Medium	(72/0)	67.3	C4.5
($ArmLengthShouldertoWrist \leq 64.2$) AND ($ChestCircumference > 93.6$) AND ($Weight \leq 181.4$) AND ($FootLength \leq 27.9$) AND ($TTR2 > 73.8$) AND ($HipBreadthSitting > 34.5$)	Large	(75/1)	70.1	PART
($Weight > 157.0$) AND ($TTR2 > 79.8$) AND ($ArmLengthSpinetoWrist > 84.1$) AND ($Stature > 172.2$) AND ($ArmLengthShouldertoWrist > 64.2$) AND ($FaceLength > 11.4$) AND ($HipCircumference \leq 103.9$)	X Large	(50/0)	72.5	C4.5
($Weight \geq 184.8$) AND ($CrotchHeight \leq 84.2$)	XX Large	(39/5)	92.9	RIPPER

Table 5.6: High coverage rules for each clothing/body size for Italian males.

Learning Top and Bottom measurements	**Class**	**Coverage**	**Positive coverage(%)**	**Learner**
($Weight \leq 163.4$) AND ($ArmLengthSpinetoWrist \leq 84.1$) AND ($Stature \leq 179.5$) AND ($ShoulderBreadth \leq 47.6$)	Small	(93/1)	73.8	C4.5
($Weight \leq 176.2$) AND ($Stature > 179.5$) AND ($ArmLengthSpinetoWrist > 81.5$) AND ($WaistCircumference > 75.7$)	Medium	(103/2)	59.5	PART
($176 < Weight \leq 225.3$) AND ($Stature \leq 183.6$) AND ($TTR2 \leq 88.4$) AND ($SubscapularSkinfold > 1.4$) AND ($WaistCircumference > 87.4$)	Large	(96/1)	69.1	C4.5
($193 < Weight \leq 225.3$) AND ($Stature > 183.6$) AND ($CrotchHeight > 85.0$) AND ($SittingHeight > 93.5$)	X Large	(52/2)	63.4	C4.5
($Weight > 225.3$) AND ($SpinetoElbow > 56.7$) AND ($SubscapularSkinfold > 2.2$) AND ($WaistCircumference > 104.1$)	XX Large	(42/1)	89.4	PART

Table 5.7: High coverage rules for each clothing/body size for Dutch males.

5.3 Rules as constraints for design and manufacturing

Here we show the rules found in the previous section may be used as constraints for cloth design and manufacture. These rules describe the most relevant attributes for each size, and we may see that for each size the set of attributes is different. In this sense, these rules may be seen as the most important features to consider when designing clothes for a determined size.

When considering the rules for each population, we observe that in general for the *XX-Large* size, the rules mainly consider the weight. This could be explained by the fact that individuals whose size is *XX-Large* are often difficult to characterize due to the nature of their body shape. Moreover, this fact may explain the reason why for tailors it is difficult to design garments that fits these individuals properly [Viktor et al., 2006].

Another important observation is that the stature changes (increase or decrease) from one size to another as opposed to many size charts, where either the stature is a fixed range for all the sizes, or increases as the size increases [Schofield and LaBat, 2005]. We notice that for the male population both Italian and Dutch as well as for the Italian females, *XX-Large* individuals are shorter than *X-Large* individuals. This implies that when designing for example pants for the *XX-Large* subjects, the legs should have short lengths. Otherwise, the mass market may be left with a large amount of garments that cannot sell. We also observe that the attributes to be taken into consideration are different from size to size, as discussed next.

American female population

In the American population, the main feature we observe is that, in all sizes, the measurements to be considered are from the top body. This indicates that the torso is the most variable feature (in all the sizes) for the American females. This is observed in the American centroids in figure 4.3, where the most noticeable differences between sizes are in the torso area. It may also observed that the *Small* size subjects have short torso. This makes sense, since *Small* subjects are usually shorter than the subjects of other sizes. A noteworthy observation is that, as apposed to the male population, in the female population, the *XX-Large* subjects are taller than the *X-Large*. Consequently, when designing clothes for the American population, this has to be taken into account to produce garments that fit the population properly.

Italian female population

By examining the Italian female population we find the following important considerations. First, the main difference between the *X-Small* and *Small* sizes is the height. Since the *Small* size subjects are taller, it follows that they have longer arms, as may be observed in table 4.7. This again represents a major consideration when designing garments such as jackets, as opposed to bust circumference or shoulder breadth, which are very similar in both sizes. The second main characteristic of the Italian female population is that the thigh circumference become relevant in garment design for the *Large* and *X-Large* sizes. From the clothing point of view, this represent a constraint when designing, for instance, pants for these sizes, while not necessary for the smaller sizes. Third, we observe that the subjects in the *X-Large* size have the longest torso. This again represents a major consideration when designing garments, in order to fit them properly.

Dutch female population

The Dutch female population presents several interesting features. First notice that the tallest individuals are among the *Medium* and *XX-Large* sizes. Also, the longest arm length is among the *Medium* and *XX-Large* populations, surpassing importantly the *Large* and *X-Large* size subjects. This may be interpreted as follows. The *Large* and *X-Large* subjects do not have the same constitution as the *Medium* size individuals. They are shorter and more robust than the *Medium* size subjects. This provides valuable information about garment design, for different sizes. For *Medium* size, the clothes have to be designed mainly for tall subjects, while the design of garments for the *Large* and *X-Large* sizes, the girths are the primary aspect to take into account.

Italian male population

For the Italian males, the rules show that the chest, waist, and neck base circumference of the *Large* and *X-Large* subjects are similar. The main feature that makes these sizes different is that the *X-Large* subjects have longer arms. Notice the impact this fact has on clothing design. Consider for instance designing jackets: when planning the manufacture for *Large* and *X-Large* sizes, the main consideration becomes the arm length. Other measurements such as the chest circumference remain practically the same in both sizes. We also notice that even though the measurements of the *X-Large* subjects are similar to the *Large* size subjects, the former are taller than the *Large* and *XX-Large* subjects.

Thus, when designing, for example, pants for the *X-Large* subjects, these need to have a long length.

Dutch male population

For the Dutch male population the following observation are noteworthy. The rules for the *X-Large* subjects consider mainly heights. That is, the stature, crotch height and sitting height are the measurements to be considered when designing clothes for subjects in this group. This may due to the fact that the *X-Large* subjects are the tallest of the population. We also observe that the *Large* and *X-Large* subjects have similar waist circumference, but the former have slightly wider chest. Subsequently, when designing jackets or shirts for the *Large* size subjects these should be wider and shorter than the ones designed for the *X-Large* subjects. We also notice that the waist circumference is an important feature to be considered, since this measure is addressed for the *Medium*, *Large* and *XX-Large* sizes.

By inspecting the rulesets, we notice the positive coverage of the rules indicate they are general enough to be applied when designing clothes. When a large amount of attributes or constraints need to be considered in the design and manufacture of clothes, the production cost may be impacted [Viktor et al., 2006]. Nevertheless, the number of constraints in the rulesets we obtain is sufficient to constrain the design and manufacturing of clothes, yet the number is small enough to be applied without impacting the production costs.

5.4 Chapter summary

In this chapter, we have shown the results of applying multi-view classification using the cluster body sizes as class labels. Multi-view classification allowed us to directly mine the data from the CAESAR^{TM} multi-relational database. Through this process, we obtained a set of rules that describe the relationship between body measurements. We showed the set of rules generated are consistent in terms of the overall accuracy. From these rules sets we have identified the most significant rules in terms of the positive coverage and the information they provide. We showed these rules may be applied as constraints when designing and manufacturing clothes. The positive coverage of the rules indicate they are general enough to be applied when tailoring clothes. We also have shown the

number of rules is sufficient to constrain the design without impacting the production costs.

As discussed previously, utility-based data mining is concerned with the economic factors that impact the overall data mining process. To this end, in the next chapter we aim to reduce the number of body measurements, and consequently the costs, while preserving the overall accuracy as if we consider the complete set of body measurements. In order to achieve this, we engage principal component analysis and feature selection. This allows us to identify the set of most significant measurements and eliminate the ones with little or no predictive information.

Chapter 6

Reducing the number of body measurements

Recall that the goal of utility-based data mining is to consider all economic aspects of the data mining process and maximize the utility of this process. In the previous two chapters, we used in our analysis, the total number of body measurements as contained in the CAESAR^{TM} database. We now aim to reduce the costs and time of the mining process by reducing the number of body measurements. Reducing the number of measurements increases the efficiency of the learning process, enhances comprehensibility of the learned results and improves the learning performance (predicted accuracy) [Han and Kamber, 2006]. Moreover, the reduced set of body measurements may help identify the measures that require special attention when designing garments reducing the cost and complexity of this process. To this end, we perform dimension reduction techniques such as *feature selection* and *feature extraction.* These techniques allow the reduction of the dataset size by removing irrelevant or redundant dimensions, or attributes, according to some metric e.g. an interestingness measure.

6.1 Dimension reduction

The CAESAR^{TM} database contains forty-five anthropometric body measurements for the females and forty-four for the males. We now aim to find the subset of most important measurements, using dimension reduction techniques. Dimension reduction is the process of reducing the number of attributes or dimensions under consideration. Dimension reduction techniques are usually divided into *feature selection* and *feature extraction.*

6.1.1 Feature selection

Feature selection, also know as *variable selection*, *feature reduction* or *attribute subset selection* [Han and Kamber, 2006, Witten and Frank, 2005], is a technique used to remove features or attributes in the data that are statistically uncorrelated with the class labels. The goal of feature selection is therefore to find a minimum set of the original attributes such that the resulting probability distribution of the classes is as close as possible to the distribution obtained using all attributes [Han and Kamber, 2006].

Feature selection algorithms fall into two broad categories, the *wrapper model* and the *filter model* [Das, 2001, Witten and Frank, 2005, Yu and Liu, 2003]. The *wrapper model* searches through the space of feature subsets using a learning algorithm to decide which features to choose. In the wrapper model, the accuracy of the learning algorithm for each feature that may be added to or removed from the feature subset is estimated. That is, the performance of the learning algorithm is used to evaluate and determine which features are selected. The *filter model* relies on general characteristics of the data to select the features without involving any learning algorithm. That is, the filter model selects a feature set to be used for any learning algorithm. The biases of the learning algorithm do not interfere with the feature selection. The search proceeds until a pre-specified number of features is selected or some threshold is reached.

In general the wrapper model tends to find features better suited to the predetermined learning algorithm, but it also tends to be more computationally expensive than the filter model. Then, when the number of features is large, the filter model is usually chosen due to its computational efficiency [Das, 2001, Yu and Liu, 2003].

Within the filter model, different feature selection algorithms may be further categorized into two groups, namely, *feature ranking* and *subset selection* [Yu and Liu, 2003], based on whether they evaluate the goodness of features individually or through feature subsets. In *feature ranking* the features are individually evaluated by a metric, e.g. an interestingness measure, and ranked based on their relevance to the target concept. All features that do not achieve an adequate score are eliminated. On the other hand, *subset selection* searches the set of possible features for the optimal subset. This search is guided by a certain evaluation measure which captures the goodness of each candidate subset.

Subset selection involves searching the space of attributes for the optimal subset. Typically, the space is searched greedily in one of two directions: top to bottom (*forward selection*) or bottom to top (*backward elimination*). At each stage, a local change is made

to the current attribute subset by either adding or deleting a single attribute. *Forward selection* starts with no attributes and adds them one at a time. *Backward elimination* starts with the full set and deletes one at a time. In forward selection, each attribute that is not already in the subset is tentatively added to it and the resulting set is evaluated e.g. using the consistency measure. This evaluation produces a numeric measure of the expected performance of the subset. The effect of adding each attribute in turn is qualified by this measure, and the best one is subsequently chosen. The process continues until any of the remaining attributes does not produce an improvement when added to the current subset. An analogous process is followed by the backward elimination [Witten and Frank, 2005].

6.1.2 Feature extraction

Unlike feature selection, which reduces the attribute set by retaining a subset of the original set of attributes, feature extraction methods transform the original features of a dataset into an alternative, more compact set of dimensions, while retaining as much information as possible. In feature extraction, a mapping of the multidimensional space into a space of fewer dimensions or attributes is applied. This means that the original feature space is transformed by applying e.g. a linear transformation.

Well-known feature extraction methods include *Factor Analysis*, *Singular Value Decomposition* (SVD) and *Principal Component Analysis* (PCA) [Cunningham, 2007]. Factor analysis is a statistical method used to explain variability among observed random variables in terms of fewer unobserved random variables called *factors*. The observed variables are modeled as linear combinations of the factors, plus *error* terms. SVD is closely related to PCA, the main difference is that in PCA the mean of the data is removed, while for SVD, it is not [Tan et al., 2005]. PCA is widely used in practice because it produces better results (in the mean-square error sense) than other linear dimension reduction techniques [Fodor, 2002].

Recall from section 2.3, that, in essence, PCA seeks to reduce the dimension of the data by finding a few orthogonal linear combinations (the principal components) of the original variables with the largest variance. The first principal component is the linear combination with the largest variance. The second principal component is the linear combination with the second largest variance and orthogonal to the first principal component, and so on. There are as many principal components as the number of the original variables. However, for most datasets, the first principal components explain most of

the variance, so that the rest may be disregarded with minimal loss of information. A detailed description of the PCA procedure was given in section 2.3.

6.2 Reduced set of measurements

Here we present the results of applying dimension reduction techniques. For feature selection, we use PCA, a widely used reduction technique that was found to produces good results [Fodor, 2002]. As discussed in section 2.3, we present PCA in the context of utility-based data mining, since the eigenvalues may be seen as an interestingness measure of a given feature. For feature selection we follow the *filter model* approach, because the assessment is performed independently of the learning algorithm. This produces more general results, and different learning algorithms may be applied after. Moreover, the filter model is computationally more efficient in large datasets than the wrapper model [Das, 2001, Yu and Liu, 2003]. Within the filter model, the algorithms follow either the feature ranking or the subset selection approach. In our experimentation, we consider both approaches.

For the feature ranking, we use the measures *Information Gain* [Quinlan, 1986], *Gain Ratio* [Quinlan, 1993] and *Chi Squared* [Mills, 1955]. These three measures have been widely used in the context of feature selection and have been found to produce good results [Cunningham, 2007]. These measures evaluate the attributes by measuring their information gain, gain ration and the chi squared statistic respectively, with respect to the class. To perform information gain in numeric attributes, WEKA first discretizes them using a MDL-based discretization method [Witten and Frank, 2005].

For subset selection, some existing evaluation measures that have been shown effective in removing both irrelevant and redundant features include the *Consistency subset evaluator* and *CFS subset evaluator*. The *consistency subset evaluator* [Dash et al., 2000] evaluates the attribute subsets by the degree of consistency in class values with respect to the original data attributes. That is, attempts to find a minimum number of features that separate classes as consistently as the full set of features do. An inconsistency is defined as two instances having the same attribute values, but different class labels. The consistency of any subset of attributes may never be higher than the one of the full set, so this method seeks the smallest subset whose consistency is the same as that of the full attribute set. The *CFS subset evaluator* [Hall, 2000] applies a correlation measure to evaluate the goodness of feature subsets based on the hypothesis that a good feature subset is one that contains features highly correlated to the class, but uncorrelated

to each other. That is, it assesses the predictive ability of each attribute individually and the degree of redundancy among them, preferring sets of attributes that are highly correlated with the class but have low intercorrelation.

In searching the attribute space for the optimal subset, several methods have been proposed. These include *Best First*, *Greedy stepwise*, *Genetic search*, *Rank search*, *Exhaustive search* and *Random search* [Han and Kamber, 2006, Witten and Frank, 2005]. Since *Exhaustive search* and *Random search* are computationally expensive for large datasets [Witten and Frank, 2005], we perform our experimentation using *Best First*, *Greedy Stepwise*, *Genetic search* and *Rank search* methods.

Best First search performs greedy hill climbing with backtracking. It is able to search forward from the empty set of attributes, backward from the full set, or start at an intermediate point. This search method keeps a list of all attribute subsets evaluated so far, sorted in order of performance measure, so it may revisit an earlier configuration if the performance starts to drop. *Greedy Stepwise* searches greedily through the space of attribute subsets, like Best First, it may progress forward or backward. Unlike Best First, this search method does not backtrack and terminates as soon as adding or deleting the best remaining attribute decreases the evaluation metric. *Genetic search* uses a simple genetic algorithm [Goldberg, 1989]. This method is based on the principle of natural selection; it evolves good feature subsets by using random perturbations of a current list of candidate subsets. *Rank search* sorts attributes using a single-attribute evaluator, e.g. Gain Ratio, and then ranks promising subsets using an attribute subset evaluator. It starts by sorting the attributes with the single-attribute evaluator and then evaluated subsets of increasing size using the subset evaluator.

In chapter 5 we present the results of performing classification rule learning with the full set of anthropometric measurements. We now present the results of training the classification rule learners *PART*, *RIPPER* and *C4.5* with the reduced set of attributes obtained from PCA and feature selection. Since we aim to find the subset that produced the best trade-off between the number of attributes and the predicted accuracy, we proceed in the following way.

For feature ranking, we select the subset of most significant attributes according to each interestingness measure. Using this subset we train the classification rule learners as follows: first train the learner using the whole subset. In the next step, we remove the lowest ranked attribute and re-run the learning algorithm. We repeat removing an attribute and re-running the learning algorithm until just two attributes are left. We finally choose the subset of attributes that produced the highest predicted accuracy.

We proceed similarly with the principal components obtained using PCA. We run the learning algorithms with the set of most important components, and at each step we remove the one with the least eigenvalue and re-run the learning algorithms. We keep the set of principal components that maximize the predicted accuracy.

For subset selection we proceed in a different way. We use the subsets returned by each of the four search methods (Best First, Greedy Stepwise, Genetic search and Rank search) to train the learning algorithms, without removing any attribute. We again choose the subset that produces the highest predictive accuracy. The results for the anthropometric and 3-D data are presented in sections 6.2.1 and 6.2.2.

6.2.1 Anthropometric data

Here we present the results of applying PCA and feature selection on the anthropometric data. Recall that the original number of body measurements is forty-four for the males, and forty-five for the females (under bust circumference was also recorded for the females). Tables 6.1 and 6.2 show the results for the American males and females, respectively. Shown are the predictive accuracy and, in parenthesis, the number of attributes in the subset.

	Original	**PCA**	**Info Gain**	**Gain Ratio**	**Chi Squared**	**Consistency Subset**	**CFS Subset**
PART	78.7%	76.8% (10)	82.6% (8)	81.2% (5)	82.6% (8)	83.1% (8)*	80.0% (26)*
RIPPER	79.2%	76.3% (11)	82.9% (6)	81.2% (8)	82.9% (15)	79.7% (8)†	78.7% (26)†
C4.5	80.1%	74.6% (11)	83.8% (6)	81.9% (7)	83.1% (7)	84.5% (8)*	79.5% (26)†

Subset produced by *Best First search, †Greedy Stepwise search.

Table 6.1: Results of the attribute reduction for the American males.

For the American male population we observe that the subsets produced by Info Gain, Gain Ratio, Chi Squared and Consistency subset evaluator considerably improve the predicted accuracy with less body measurements. This is specially evident for the subset produced by the Consistency subset evaluator, where the accuracy is higher than 83%, when using PART and C4.5. This subset then contains the eight most important measures to define the body sizes for the male population.

For the American females, it may be seen that the subsets obtained using Info Gain, Chi Squared and the subsets produced by Best First using the Consistency measure sig-

	Original	PCA	Info Gain	Gain Ratio	Chi Squared	Consistency Subset	CFS Subset
PART	75.8%	79.3% (15)	82.0% (24)	77.3% (11)	80.9% (27)	82.4% (8)*	79.7% (25)‡
RIPPER	76.6%	75.0% (10)	80.5% (19)	79.3% (19)	81.3% (12)	77.7% (13)‡	78.1% (25)*
C4.5	76.2%	78.1% (5)	80.1% (6)	78.5% (7)	80.1% (8)	83.6% (8)*	77.7% (21)†

Subset produced by *Best First search, †Greedy Stepwise search, ‡Genetic search.

Table 6.2: Results of the attribute reduction for the American females.

nificantly improve the predicted accuracy. We also observe that the subsets produced using the Consistency subset evaluator contains, in average, a smaller number of body measurements than the subsets produced using Info Gain and Chi Squared measures. Moreover, the accuracy is maximized using the subset produced by Best First containing only eight body measurements. We then select this subset as containing the most significant body measurements to define the body sizes for the American females. The reduced sets of body measurements for the American population are shown in table 6.3.

Males	Females
1. Acromial Height Sitting	1. Arm Length (Shoulder-Wrist)
2. Arm Length (Shoulder-Wrist)	2. Arm Length (Shoulder-Elbow)
3. Arm Length (Spine-Wrist)	3. Bust Circumference under Bust
4. Hand Length	4. Buttock Knee Length
5. Knee Height Sitting	5. Stature
6. Stature	6. Subscapular Skinfold
7. Thumb Tip Reach	7. Thumb Tip Reach
8. Weight	8. Weight

Table 6.3: Reduced sets of body measurements for the American population.

We next show the results for the Dutch population. Tables 6.4 and 6.5 present the results of applying PCA and feature selection for the males and females, respectively.

From table 6.4 it may be seen that for the Dutch males, PCA and all feature selection methods produce good results. That is, the subsets contain a small number of attributes (except by the subsets obtained using the CFS evaluator) and improve the accuracy we obtained using the full set of attributes. We observe that, in general, the highest accuracy

	Original	PCA	Info Gain	Gain Ratio	Chi Squared	Consistency Subset	CFS Subset
PART	80.3%	82.9% (12)	82.7% (13)	82.2% (14)	81.1% (14)	80.3% (12)§	81.3% (25)‡
RIPPER	80.6%	81.5% (7)	81.3% (9)	82.4% (14)	81.3% (13)	82.0% (8)*	80.3% (25)‡
C4.5	78.5%	82.0% (7)	81.8% (9)	82.2% (6)	80.6% (13)	80.4% (12)§	80.3% (31)§

Subset produced by *Best First search, ‡Genetic search, §Rank search.

Table 6.4: Results of the attribute reduction for the Dutch males.

	Original	PCA	Info Gain	Gain Ratio	Chi Squared	Consistency Subset	CFS Subset
PART	81.1%	80.4% (19)	83.9% (12)	82.3% (13)	84.7% (7)	83.1% (7)†	81.9% (35)§
RIPPER	77.9%	77.3% (19)	83.4% (13)	82.9% (17)	82.7% (7)	81.3% (7)*	81.9% (25)*
C4.5	77.4%	77.4% (19)	82.9% (13)	83.3% (11)	82.7% (7)	81.7% (7)†	79.3% (25)†

Subset produced by *Best First search, †Greedy Stepwise search, §Rank search.

Table 6.5: Results of the attribute reduction for the Dutch females.

is achieved using the subsets produced by Gain Ratio and PCA. Although PCA produces accurate results, its application in a tailoring scenario presents additional challenges, because PCA do not produce a subset of the original attributes. Instead, PCA produces a linear combination of the original set of attributes, preventing the direct application of PCA results in the tailoring process. We therefore select the subset containing six attributes produced by Gain Ratio, because this produces the best trade-off between accuracy and the number of attributes.

For the Dutch females the best results are obtained using Info Gain, Gain Ratio and Chi Squared. These three interestingness measures produced subsets that highly improve the accuracy. However, the number of attributes in the subsets generated by Info Gain and Gain Ratio is larger than the number of attributes in the subset produced by Chi Squared. We therefore select the subset with seven attributes produced by Chi Squared. The reduced sets of body measurements for the Dutch population both males and females is presented in table 6.6.

Finally, we present the results obtained for the Italian population. Tables 6.7 and 6.8 show the results for the male and female population respectively.

By inspecting table 6.7 we observe that for the Italian males, PCA and Gain Ratio

Males	Females
1. Chest Girth at Scye	1. Arm Length (Spine-Wrist)
2. Hip Breadth Sitting	2. Bust Circumference
3. Stature	3. Chest Girth at Scye
4. Vertical Trunk Circumference	4. Stature
5. Waist Circumference	5. Thumb Tip Reach
6. Weight	6. Vertical Trunk Circumference
	7. Weight

Table 6.6: Reduced sets of body measurements for the Dutch population.

	Original	**PCA**	**Info Gain**	**Gain Ratio**	**Chi Squared**	**Consistency Subset**	**CFS Subset**
PART	73.6%	80.9% (13)	82.1% (18)	82.1% (15)	81.1% (12)	79.7% (20)§	78.7% (25)‡
RIPPER	74.3%	83.3% (13)	78.5% (15)	80.2% (10)	77.5% (14)	78.9% (20)§	78.0% (25)‡
C4.5	73.9%	81.4% (12)	77.0% (9)	78.2% (8)	76.3% (8)	75.8% (20)§	76.3% (25)‡

Subset produced by ‡Genetic search, §Rank search.

Table 6.7: Results of the attribute reduction for the Italian males.

	Original	**PCA**	**Info Gain**	**Gain Ratio**	**Chi Squared**	**Consistency Subset**	**CFS Subset**
PART	78.1%	78.4% (18)	83.8% (7)	83.3% (10)	81.7% (7)	81.2% (12)§	80.2% (24)†
RIPPER	75.8%	75.5% (7)	81.7% (7)	81.4% (7)	82.5% (8)	80.2% (8)*	78.1% (31)§
C4.5	76.8%	76.8% (12)	81.7% (5)	81.7% (8)	83.0% (6)	79.1% (12)§	79.9% (24)*

Subset produced by *Best First search, †Greedy Stepwise search, §Rank search.

Table 6.8: Results of the attribute reduction for the Italian females.

produce the best results. As mentioned previously, PCA presents additional challenges when applying directly in the garments design. We then select the subset generated by Gain Ratio that maximizes the accuracy. That is, the subset containing fifteen attributes.

For the Italian female population, we notice that the subsets produced by Info Gain and Chi Squared significantly increase the accuracy. The highest accuracy is achieved using the subset generated by Info Gain that contains seven attributes. We therefore select this subset of body measurements for the Italian female population. The reduced sets containing the most important body measurements for the Italian population are shown in table 6.9.

Males		**Females**
1. Arm Length (Shoulder-Wrist)	9. Hip Circ Max Height	1. Arm Length (Shoulder-Wrist)
2. Arm Length (Spine-Wrist)	10. Knee Height Sitting	2. Arm Length (Spine-Wrist)
3. Buttock Knee Length	11. Stature	3. Knee Height Sitting
4. Chest Circumference	12. Thumb Tip Reach	4. Stature
5. Chest Girth at Scye	13. Waist Circumference	5. Thumb Tip Reach
6. Crotch Height	14. Waist Height	6. Vertical Trunk Circumference
7. Hip Breadth Sitting	15. Weight	7. Weight
8. Hip Circumference		

Table 6.9: Reduced sets of body measurements for the Italian population.

By inspecting the results of the three populations, we observe that the body measurements may be reduced significantly by applying feature selection and feature extraction. Actually, for the American, Dutch and Italian female populations, the reduced sets contain only around eight body measurements. For the Italian male population, the number of body measurements is slightly higher, but there still is an important reduction. We notice that these reduced sets of body measurements improve the predictive accuracy. These subsets therefore contain the most important body measurements for defining the body sizes.

We also observe that, even though there are some measurements are common to all populations such as the length of the arm, stature and weight, the reduced set of body measurements also consider specific body measurements that are different for each population. This may indicate that the set of body measurements that require special attention e.g. when designing clothes, are different from one population to another. This makes sense, if we consider the fact that each population has its own distinctive

characteristics. Recall that in Chapter 4, we found that the shortest and thinnest population (of the three that we consider in our study) are the Italians, the tallest population corresponds to the Dutch, and Americans represent the most robust population.

When considering the aforesaid situation, it becomes natural that different body measurements describe better each of the populations. For the American males the most important body measurements are the acromial height and knee height together with the length of the arm. Special attention should be paid to the knee and acromial heights when designing long or short pants, in order to thus take the position of the knee into consideration. Furthermore, the length of the arm is important when designing shirts that fit this population well. For the females, the circumference under the bust and the buttock knee length become crucial when defining the body size. Hence, when designing clothes for the American females, the circumference under the bust should receive more attention than other measurements that are mainly used in garment design, such as the bust circumference. Moreover, the subscapular skinfold, a measurement of subcutaneous fat accumulation, is considered in the reduced set of body measurements for the American females. This confirms our previous results that the Americans are the most robust population.

For the Dutch males, the reduced set of measurements indicates that the most significant measurements are the waist circumference, the chest girth at scye and the vertical trunk circumference. When tailoring shirts, sweaters or jackets, for the Dutch males, these measurements should be considered carefully to produce garments that fit this population properly. For the Dutch females, the most important measurements are the bust circumference and, as in the case of the males, the chest girth at scye and the vertical trunk circumference. The vertical trunk circumference takes into direct account the fact that we already know about this population, i.e. they are the tallest population, since, generally, the longer the trunk the taller the person. Furthermore, for the Dutch females, unlike the American females, the bust circumference becomes critical when determining the body size. Therefore, when tailoring clothes for the Dutch females, the bust circumference requires special attention in order to design garments that fit the population better.

Finally, for the Italian females, the most important measurements are the vertical trunk circumference and the knee height, which are relevant when, for example, tailoring blouses, skirts or pants. The vertical trunk circumference is important when deciding how long a jacket or blouse should be, in order to produce garments that are not too short or long for this population. For the Italian males, the reduced set of measurements

considers the chest, waist and hip circumferences along with the crotch, waist and hip heights. The measurements, then, address both the height and girths. This indicates that not only the height, but also the chest, waist and hip circumferences should receive special attention when designing clothes for the Italian males. This, again, confirms our results that the main characteristics of the Italian population are related to height and girths. That is, the Italians are the shortest and thinnest population.

6.2.2 3-D data

Here we present the results of applying PCA and feature selection on the 3-D data. As discussed in Chapter 4, in our analysis of the 3-D data we only consider the Dutch population. This is due to the fact that a better sampling was archived for this population. Moreover, there is a high percentage of missing 3-D scans for the Italian population.

Tables 6.10 and 6.11 show the results for the Dutch males and females, respectively. The original number of indices used to describe the 3-D body scans is hundred and twenty for both, males and females.

	Original	**PCA**	**Info Gain**	**Gain Ratio**	**Chi Squared**	**Consistency Subset**	**CFS Subset**
PART	78.8%	81.6% (19)	85.2% (58)	81.0% (56)	83.1% (51)	82.3% (24)‡	81.0% (52)*
RIPPER	76.8%	80.6% (15)	80.8% (54)	79.3% (54)	81.2% (51)	78.8% (8)*	79.2% (52)†
C4.5	79.3%	78.8% (21)	80.3% (56)	79.5% (56)	80.4% (53)	80.1% (8)†	80.4% (72)§

Subset produced by *Best First search, †Greedy Stepwise search, ‡Genetic search, §Rank search.

Table 6.10: Results of the attribute reduction of 3-D data for the Dutch males.

	Original	**PCA**	**Info Gain**	**Gain Ratio**	**Chi Squared**	**Consistency Subset**	**CFS Subset**
PART	78.6%	75.6% (32)	78.4% (57)	79.2% (54)	78.0% (54)	78.0% (44)§	81.9% (59)†
RIPPER	75.1%	76.0% (9)	80.4% (55)	80.1% (55)	78.4% (55)	79.8% (8)*	79.0% (88)§
C4.5	78.7%	74.1% (12)	77.5% (58)	77.2% (54)	77.1% (54)	78.1% (24)‡	79.8% (59)†

Subset produced by *Best First search, †Greedy Stepwise search, ‡Genetic search, §Rank search.

Table 6.11: Results of the attribute reduction of 3-D data for the Dutch females.

From table 6.10 it may be observed that for the Dutch male population, Info Gain and Chi Squared achieve the best accuracy. However, the number of attributes in the subsets produced by these measures is much bigger than the number of attributes in the subsets produced by the Consistency subset evaluator. Moreover, there is a set produced by the Consistency subset evaluator that contains only eight indices, and the accuracy achieved by this set it is just slightly lower than the one obtained using the subsets generated by Info Gain and Chi Squared. The best trade-off between accuracy and the number of attributes is then achieved by the subset produced using the Consistency subset evaluator containing eight indices.

For the females, it may be observed that the highest accuracy is achieved by the subsets produced by the CFS subset evaluator. Nevertheless, the number of attributes in these subsets are big. By inspecting the table, we may notice that PCA and the Consistency subset evaluator produced subsets containing a smaller number of attributes. Moreover, the accuracy, in the case of the subset produced by the Consistency subset evaluator containing eight indices, is comparable to the accuracy obtained using the subsets produced by the CFS subset evaluator. Therefore, the subset produced by Consistency subset containing eight indices, shows the best trade-off between the number of attributes and accuracy.

By inspecting the results of applying PCA and feature selection techniques on the 3-D data, the following observations are noteworthy. Similar to the results of the anthropometric data, the results of the 3-D data indicate that the accuracy is improved using a smaller set of attributes. In comparison with the reduction achieved on the anthropometric data, the reduction of the 3-D data is higher. The reduced sets of attributes contain around eight attributes in both the anthropometric and 3-D data, but the original number of indices used to describe the 3-D body scans is hundred and twenty, while the original number of body measurement is forty-five for the females and forty-four for the males.

Even though a smaller reduction ratio is achieved on the anthropometric data compared with the 3-D data, this is still significant, since more of the 80% of the body measurements are of less importance. Moreover, this reduction improves the efficiency of the classification results and reduces the complexity of the mining process by using around 20% of the original attributes.

6.3 Chapter summary

This chapter shows the results when aiming to reduce the cost of the mining process by reducing the number of body measurements. A reduced set of body measurements may improve the performance of the mining process and enhance the comprehensibility of the mined results.

To this end, we performed dimension reduction techniques. Dimension reduction techniques are broadly divided into feature extraction and feature selection, based on whether they produce a subset of linear combinations of the original attributes, or whether they produce a subset of the original attributes. Feature selection methods are divided into filter models and wrapper models. In our analysis we choose to use the filter model, since this method is independent of the learning algorithm. The filter model techniques are further divided into feature ranking and subset selection, based on whether they evaluate the features individually or through feature subsets. In our analysis we applied both methods. For the feature ranking we used Information Gain, Gain Ratio and Chi Squared. For subset selection we used the Consistency subset evaluator and the CFS subset evaluator. To perform feature extraction, we used PCA.

Our results indicate that the number of body measurements may be reduced significantly by applying feature selection and feature extraction. Moreover, these new sets of body measurements reduce the cost of the mining process and improve significantly the predictive accuracy. These sets therefore contain the most important body measurements for defining the body sizes, and may be used in garment design to identify the measurements that require special attention for each population. For example, for the American females, we found that the under bust circumference should receive more attention than the bust circumference, which is the measurement that is mainly used when designing garments.

In the next chapter we present the results of our analysis when aiming to understand the demographic nature of the individuals within each size. To this end, we perform association analysis on the demographic data. This data mining technique allows us the discovery of relationships between different demographic attributes.

Chapter 7

Determining the target populations

In previous chapters we analyze the anthropometric and 3-D data as contained in the CAESAR^{TM} database. Now, we aim to understand the demographic nature of the individuals within each size. Understanding the demographic nature of these individuals may define new market opportunities e.g. for the clothing industry, as follows. When defining a target market, demographic aspects such as age, gender, income and lifestyle choices are considered. This demographic profile may be used to determine, for example, the potential customers of expensive line of garments and which are the main lifestyle characteristics within this group. This information may next be used to determine when and where advertising should be placed so as to obtain maximum results.

In order to achieve this objective, we perform *association analysis* on the demographic data. This data mining technique allows the discovery of relationships between different demographic attributes. The discovery of these relationships helps achieving a better understanding of the demographic nature of potential customers. Since, even from a small dataset, the number of discovered rules may be very large, we apply *interestingness measures* to narrow the search space and find the truly interesting patterns.

In this chapter we provide an overview of association analysis and the classical algorithms for association rule mining. We present our experimental design and the interestingness measures we apply to improve the quality of the mined rules. We show the results of our demographic analysis for each population and suggest how to take advantage of this knowledge in the design and manufacture of garments.

7.1 Association analysis overview

Association analysis or *association rule mining* [Han and Kamber, 2006, Tan et al., 2005] is a data mining technique that is used to find interesting relationships among data attributes or items in a given dataset. Generally the uncovered patterns are represented in the form of association rules. Recall from section 2.2.3, that an association rule is an implication of the form $A \rightarrow C$, where A and C are nonintersecting sets of items. Association rules identify collections of data attributes that are statistically related in the underlying data.

Association rules are similar to classification rules in the sense that both may be expressed as if-then rules. However, association rules imply any attribute, not just the class attribute. Association rules also imply combinations of attributes. Moreover, association rules are not intended to be used together as a set, as classification rules are. Different association rules express different regularities present in the dataset, and they usually imply different relationships [Witten and Frank, 2005]. In general, association rule mining may be viewed as a two-step process [Han and Kamber, 2006]:

1. Find all frequent itemsets: an itemset is called frequent when its support is above a predefined minimum value. That is, the set of data attributes that appear frequently together in a database.

2. Generate strong association rules from the frequent itemsets: by definition strong or interesting rules must satisfy minimum *support* and minimum *confidence*. However, many different association rules satisfying the minimum *support* and *confidence* may be mined from even a small dataset. Additional interestingness measures then may be applied to reduce the number of them, and find the rules that are truly interesting.

Apriori is a classical algorithm for learning association rules proposed by Agrawal et al. [Agrawal and Srikant, 1994]. Apriori employs an iterative approach known as a *level-wise search*, where k-itemsets are used to find $(k + 1)$-itemsets. First, the set of frequent 1-itemsets is found by scanning the database to accumulate the count for each item, and collecting those items that satisfy the minimum support. The resulting set is denoted by L_1. Next, L_1 is used to find L_2, the set of frequent 2-itemsets. This in time is used to find L_3, and so on, until no more frequent k-itemsets may be found. To improve the efficiency of the level-wise generation and reduce the search space, Apriori

uses a frequent itemset property called the *apriori property*. This property states that all nonempty subsets of a frequent itemset must also be frequent.

Another well know algorithm for mining association rules is *Predictive Apriori* proposed by Scheffer [Scheffer, 2001]. The most significant difference between Apriori and Predictive Apriori relates to how the interestingness of an association rule is measured. In the Apriori algorithm, rules are usually ranked according to its *confidence* value. Predictive Apriori algorithm tries to maximize the expected accuracy that an association rule will have for future data, instead of the accuracy on the training data (as measured by confidence in Apriori). In order to achieve this, Scheffer [Scheffer, 2001] combines confidence and support into a single measure called the *expected predictive accuracy*, which provides information about the accuracy of the rule found.

Other algorithms for mining association rules that follow a different approach have been proposed. These include *Frequent pattern growth* (*FP-growth*) and *ECLAT*. *FP-growth* [Han and Kamber, 2006] is a method of mining frequent itemsets without candidate generation. Instead, it constructs an *FP-tree*, a compact data structure that compresses the original transaction database. *ECLAT* [Tan et al., 2005], on the other hand, is a method that uses the vertical data format to find the frequent itemsets together with the apriori property.

7.2 Demographic data analysis

Recall from section 3.5, that, the CAESAR^{TM} database includes demographic data such as fitness, education, occupation, marital status, income, etc. Most of this information corresponds to a typical anthropometric survey questions. However, data about the subject's car was also recorded. The demographic data was collected through a demographic questionnaire that was filled out by the participant. Even though the information was clarified with the participant in case of apparent inconsistencies [Blackwell et al., 2002], the demographic data provided depend on the truthfulness and objectivity of the participants. Moreover, in some populations there was a high percentage of missing values for certain questions, e.g. family income.

The sampling of the population is another issue that should be consider when analyzing the demographic data. Although the goal of stratified sampling is to have an equal number of subjects on each stratum, for the minority groups this was not achieved. This minority groups are, then, not representative of the whole population belonging to these groups. This is specially evident in the American and Italian sample (section 3.6).

Furthermore, the sampling of the Italian population is biased toward young single people whose occupation is student and therefore have low income. The sampling in the Dutch population shows a more balanced number of subjects per strata and it is not biased toward any age range. Therefore, the best sampling was achieved for the Dutch population.

In our analysis of demographic data we use the Apriori [Agrawal and Srikant, 1994] and Predictive Apriori [Scheffer, 2001] association rule learners. Apriori is a simple and efficient algorithm for finding association rules [Sotiris Kotsiantis, 2006]. Predictive Apriori follows the same principle as Apriori, and improves the process of how the interestingness of a rule is measured [Scheffer, 2001]. As mentioned previously, the default measure for evaluating the interestingness of a rule in Apriori is confidence. *Confidence* [Agrawal et al., 1993] is the proportion of the records covered by the antecedent that are also covered by the consequent. A problem with confidence is that is sensitive to the frequency of the consequent. Consequents with higher support produce higher confidence values even if there is no association between the items [Sheikh et al., 2004]. We then also use the interestingness measures *lift* [Brin et al., 1997a], *conviction* [Brin et al., 1997b] and *leverage* [Piatetsky-Shapiro, 1991]. These measures proved to be useful in finding interesting patterns [Sheikh et al., 2004].

Recall from section 2.2.3, that Lift is the ratio of the observed frequency of the consequent occurring in the context of the antecedent over that expected if the two were independent. Lift is a symmetric measure, i.e. $lift(A \rightarrow C) = lift(C \rightarrow A)$. Conviction is similar to lift, except that measures the effect of the consequent not being true. Conviction in contrast to lift is not a symmetric measure. Leverage measures the proportion of additional cases covered by both the antecedent and the consequent above those expected if they were independent of each other. Values above 0 for leverage and greater than 1 for lift and conviction indicate that the implication described by the rule did not occur by chance, but because there is a relationship between the antecedent and the consequent, and therefore the rule is interesting.

As shown by Sheikh et al. [Sheikh et al., 2004], any measure alone cannot determine the interestingness of a rule. Instead, a combination of different measures have to be used in order to get the rules that are really interesting. We select then the association rules whose lift and conviction values are greater than 1 and leverage value greater than 0. We also filter out rules with less than 85.0% of confidence. For the Predictive Apriori algorithm, we select only the rules with 75% or greater accuracy. Tables 7.1 to 7.6 show the most interesting rules found for each population, according to these criteria.

Age	Fitness	Education	Occupation	Children	Marital Status	Income (US$)	Clothing Size
21-25	High	Bachelor		0	Single	30,000-45,000	Small
	Medium			<=2	Married	80,000+	Small
			Engineer	<=2	Married	80,000-100,000	Small
			Management	<=2	Single or Married	Over 100,000	Small
	High		Engineer	0	Single	45,000-60,000	Medium
36-40		Masters		<=2	Married	80,000-100,000	Medium
30-40	Low		Management		Married	Over 100,000	Medium
			Engineer	<=2	Married	Over 100,000	Medium
21-25	High	Bachelor		0	Single	30,000-45,000	Large
31-35	High	Masters			Married	Over 100,000	Large
31-35	Low	Masters			Married	80,000-100,000	Large
36-40	Low	Bachelor	Management		Married	80,000+	Large
		Bachelor	Engineer	<=2	Married	60,000-80,000	XLarge
		Bachelor			Married	60,000-80,000	XLarge
	Low	Bachelor	Engineer		Married	80,000-100,000	XXLarge
46-50					Married	Over 100,000	XXLarge

Table 7.1: Most interesting rules for the American male population.

Age	Fitness	Education	Occupation	Children	Marital Status	Income (US$)	Clothing Size
	High	Bachelor		0	Single	30,000-45,000	Small
			Administrative support		Married	45,000-60,000	Small
	Medium	Bachelor		0	Single	30,000-45,000	Medium
		Masters			Married	Over 100,000	Medium
	High	Bachelor		0	Single	30,000-45,000	Large
					Married	80,000+	Large
			Management			Over 100,000	Large
		Technical training	Administrative support			30,000-45,000	XLarge
26-30					Married	60,000-80,000	XLarge
		High School				80,000-100,000	XLarge
36-40		Masters		0	Married	80,000-100,000	XXLarge

Table 7.2: Most interesting rules for the American female population.

Age	Fitness	Education	Occupation	Children	Marital Status	Income (US$)	Clothing Size
18-20				0	Single	Less than 12,000	Small
20-30		Middle level		0	Single	25,000-38,000	Small
36-40	Low	Lower level			Married	25,000-38,000	Small
56-60					Married	58,000-77,000	Small
21-25				0	Single	Less than 12,000	Medium
26-30		Bachelor		0	Married	38,000-58,000	Medium
31-35		Bachelor		0	Married	25,000-38,000	Medium
			Transportation & Communications	<=2	Married	25,000-38,000	Large
36-40			Industry	<=2	Married	25,000-38,000	Large
		Bachelor		<=2	Married	38,000-58,000	Large
			Financial Institution	0		25,000-38,000	XLarge
20-30				0	Living together	25,000-38,000	XLarge
30-55					Married	38,000-58,000	XLarge
	Low	Middle level			Married	25,000-38,000	XXLarge
	Medium	Middle level			Married	38,000-58,000	XXLarge

Table 7.3: Most interesting rules for the Dutch male population.

Age	**Fitness**	**Education**	**Occupation**	**Children**	**Marital Status**	**Income (US$)**	**Clothing Size**
21-25				0	Single	Less than 12,000	XSmall
26-30		Bachelor		0	Single	19,000-25,000	XSmall
				<=3	Married	25,000-38,000	XSmall
30+				0	Married	58,000-77,000*	XSmall
				0	Single	Less than 12,000	Small
26-30				0	Single	19,000-25,000	Small
	Low				Married	38,000-58,000	Small
18-25				0	Single	Less than 12,000	Medium
30+					Married	77,000-103,000	Medium
		High School				25,000-38,000	Large
45-55		High School	Homemaker	<=2	Married	25,000-38,000	Large
40-45	Low		Homemaker		Married	25,000-38,000	XLarge
45-55					Married	38,000-58,000	XXLarge

Table 7.4: Most interesting rules for the Dutch female population.

Age	Fitness	Education	Occupation	Children	Marital Status	Income (US$)	Clothing Size
			Student	0	Single	Less than 7,000	Small
18-20			Student			Less than 7,000	XLarge
51-55						33,000-44,000	XLarge

Table 7.5: Most interesting rules for the Italian male population.

Age	Fitness	Education	Occupation	Children	Marital Status	Income (US$)	Clothing Size
26-30			Administrative support			22,000-33,000	Large

Table 7.6: Most interesting rules for the Italian female population.

7.3 Demographic profile

Using the information provided by the association rules, we now analyze the interrelationships between age, fitness, education, occupation and income, amongst others. This demographic profile allows a better understanding of the customer and some of their preferences.

American male population

By inspecting the rules in table 7.1, it may be seen there is a relationship between the level of fitness and the marital status and number of children. In general, single males with no children have higher level of fitness than married males. Also, we observe that a 30 aged or older male is usually married, while a male younger than 30 years is most likely to be single. In general, married males have higher income than single males. This is specially evident for the Medium sized males, where a married Engineer earns significantly more than his single counterpart. Moreover, for the Large sized group, a 31-35 aged married male holding a master's degree with high level of fitness have higher income than an unfit person with the same background.

We also notice that the males whose occupation is management have high income irrespective of their marital status, number of children and fitness level. An expensive clothing brand then may want to design e.g. exclusive business suits that fit well this segment of the population.

For the XX-Large sized males, it may be seen that there is a segment of 46-50 aged married males that have high income. In chapter 5 we showed that, in general, XX-Large people is difficult to characterize and it is therefore problematic to produce garments that fit them well. Again, a clothing designer may want to study further the characteristics of the individuals within this group in order to produce exclusive garments that target this segment of the population.

Finally, we notice there is a strong relationship between the males whose occupation is management and the Infiniti luxury model car. The same relationship is observed between Engineers and Ford automobiles. Moreover, males holding a master's degree tend to prefer Nissan cars. This information may be used to further define the clothing preferences of the customer. For example, a person driving a sport car would be more likely to buy a more casual suit than a person driving a more classic car.

American female population

For the American female population we observe that they have some patterns in common with the male population. From table 7.2 it may be seen that the income of a married female is higher than her single counterpart. Unlike the male population where a male younger than 30 years is single, females aged 25 or older are usually married. We also notice that the fitness level of a single female is higher than a married female. Moreover, a single female holding a bachelor degree usually has medium or high level of fitness. A clothing company then may want to design e.g. sportswear for this group; or a gym may target them for advertising a special fitness program.

As observed in the male population, there is a pattern indicating that females whose occupation is management have high income. Once more, a clothing company may design expensive professional clothes to target this segment of the population.

Unlike the male population, a married female holding a master's degree has an annual income over $100,000 dollars. A further study of this segment of the population may reveal lifestyle choices that may make them potential customers for certain products.

Another interesting pattern holding for the female population is that a person whose annual income is equal or higher than $60,000 dollars, is a potential customers for buying a new car.

Dutch male population

From the association rules in table 7.3, it may be seen than in general married males have higher income than single males. An exception to this pattern may be observed for the Small sized males, where a 20-30 single male have a similar income than a 36-40 aged married male. The key factors for this exception may be the difference in age and level of education. We also observe, in contrast to the American males, that for the Medium sized males, a married 26-30 aged male holding a bachelor degree earns significantly more than a married male with the same background, but 31-35 aged.

For the XX-Large sized males the following patterns may be observed. A married male with middle level education and medium level of fitness have higher income than a male with the same background, but with lower level of fitness. Moreover, some rules indicate that an XX-Large sized and 56-60 aged male has a high probability of being unemployed.

Unlike the American male population, for the Dutch males there is no clear relationship between the level of fitness and the marital status or the number of children. That is, fitness seems to be independent from the marital status or the number of children. Another interesting pattern for the Dutch male population is that some rules indicate that the income is related to the place of residence. In general people with high annual income live in Utrecht city.

Dutch female population

For the Dutch female population, we observe there is some relationship between the size and the age. In general younger females tend to be smaller in size, while older females are larger in size. Similar to the Dutch male population, there is no clear relationship between the level of fitness and the number of children or the marital status. The pattern between the marital status and the income is also present for the females. That is, single females earns on average less than the married counterpart.

Also, it may be seen that a married female who is 30 years or older and is X-Small or Medium in size has a high income. A clothing company may wish to investigate further this segment of the population and design exclusive garments for them. Similar to the Dutch male population, there is a strong pattern between the annual income and the place of residence. Utrecht city is indicated as the place where people have higher income. This is the case for the married Medium sized females whose income is in the range of $77,000-103,000 dollars.

Italian population

As may be seen from tables 7.5 and 7.6, for the Italian population (both males and females) the number of rules is very small. This is caused by the bias of the data toward young people. Actually, around 65 % of the population is 30 years or younger. Most of them are students that are single with no children, and low income. Therefore, the association rules generated from these datasets mainly show this pattern. Moreover, defining a demographic profile of the Italian population is additionally difficult by the fact that around 67% of the income values and about 50% of the automotive values were missing.

7.4 Chapter summary

In this chapter we presented the results of our analysis when aiming to understand the demographic nature of the individuals within each size. To this end, we performed association rule mining and applied different interestingness measures to reduce the search space and improve the quality of the results. In our analysis of the demographic data, we used the Apriori and Predictive Apriori rule learners. For the Apriori algorithm the default measure to evaluate the interestingness of a rule is confidence. Since confidence is sensitive to the frequency of the consequent, we used a combination of interestingness measures that includes lift, conviction and leverage. For the Predictive Apriori algorithm the measure used to evaluate the interestingness of a rule is predictive accuracy. This is a measure that combines confidence and support into a single measure.

As a result of our demographic analysis we identified some relationships between age, fitness, marital status, education and income, amongst others. For the American and Dutch populations our demographic profile indicates that there are some segments of the population that may be potential customers for the clothing industry. We also identified some of the main lifestyle choices for these segments. In the case of the Italian population, the number of rules found was very small, due to the bias of the data toward young people. Moreover, the number of missing values for some key attributes is high, making it difficult to define a demographic profile for this population.

In the next chapter we present the conclusions of our study and a summary of the contributions of this thesis. We also show the practical application of our study and discuss future work.

Chapter 8

Conclusions

In this chapter we present a brief discussion and a practical application of our study. We also summarize the contributions of this thesis and discuss future work.

8.1 Discussion

One of the greatest challenges for the apparel industry is to produce garments that fit the customers properly, are esthetically pleasing and comfortable. This is of crucial importance not only for the expensive designer labels, but also for the mass market. Ashdown et al. [Ashdown and Loker, 2005] identified two main issues that have limited the ability of the apparel companies to produce garments with quality fit. First, the lack of up-to-date anthropometric data to describe the civilian population. Second, the lack of information about the principal aspects to consider when designing garments for a variety of body sizes and shapes. Thus, in order to produce better fitting garments, accurate and up-to-date body measurements of civilian populations are needed and a better characterization thereof. This implies that the different sizes must correspond to real body shapes, in the sense that one or more archetypes should represent any individual belonging to the same size [Viktor et al., 2006].

Consequently, it is important to define clusters that may be characterized by one archetype, i.e. a truly representative of all other individuals that belong to the same cluster. Based on the assumption that the cluster has a spherical or quasi spherical symmetry, the archetype then corresponds to the closest individual to the Centroid of the cluster. We might also choose one of the individuals belonging to the sub-region with the highest density in terms of number of individuals. If the resulted clusters are not

spherical, more than one archetype might be necessary to fully characterize the cluster. In the context of tailoring, however, the optimal scenario is to cover the greatest number of people with the fewest number of sizes. In this context then, it is preferred to have only one archetype, since each new size or sub-size involves more costs and increases the complexity in the manufacturing. It is also important that each cluster represents a large proportion of the population, otherwise clothes would be manufactured for a size that fits virtually nobody, causing financial lost. A cluster with small number of members may be justified, however, for very expensive clothes for which a good fit is an essential condition [Viktor et al., 2006].

The method we utilize in this work satisfies the aforementioned requirements, since we were able to group the individuals into clusters with a well defined Centroid. Our verification using the Cleopatra system, indicate that the cluster membership correspond to the reality, in the sense that the bodies corresponded to our expectations of the cluster membership. We also created sets of rules that describe the relationships between the different body measurements within each body size. These rules then describe the most relevant measurements for each size, and indicate which of these measurements need to be taken into account when tailoring clothes. Moreover, we found reduced sets of body measurements. These sets contain the most important body measurements for defining the body sizes. When designing clothes, these measurements require special attention to produce garments that best fit the population. We also analyzed the demographic data to better understand the demographic nature of the individuals within each size. We found interrelationships between fitness, education, marital status and income. This analysis allowed us identifying potential target markets for the clothing industry.

8.2 Application

In this work we used the full set of measurements describing the whole body, and therefore we obtained a characterization of the whole body. However, for designing specific garments e.g. pants, we may only consider the bottom body measurements. Aldrich [Aldrich, 2006] provides a list of standard body measurements for clothing design. These measurements are shown in figure 8.1.

From figure 8.1, we may observe that for tailoring pants we mainly need to consider the waist circumference, hip circumference, crotch height, height and trouser waist which is 4 cm. below the natural waist. Using these measurements one may apply the method we described in this work, define the clusters that best describe the population and iden-

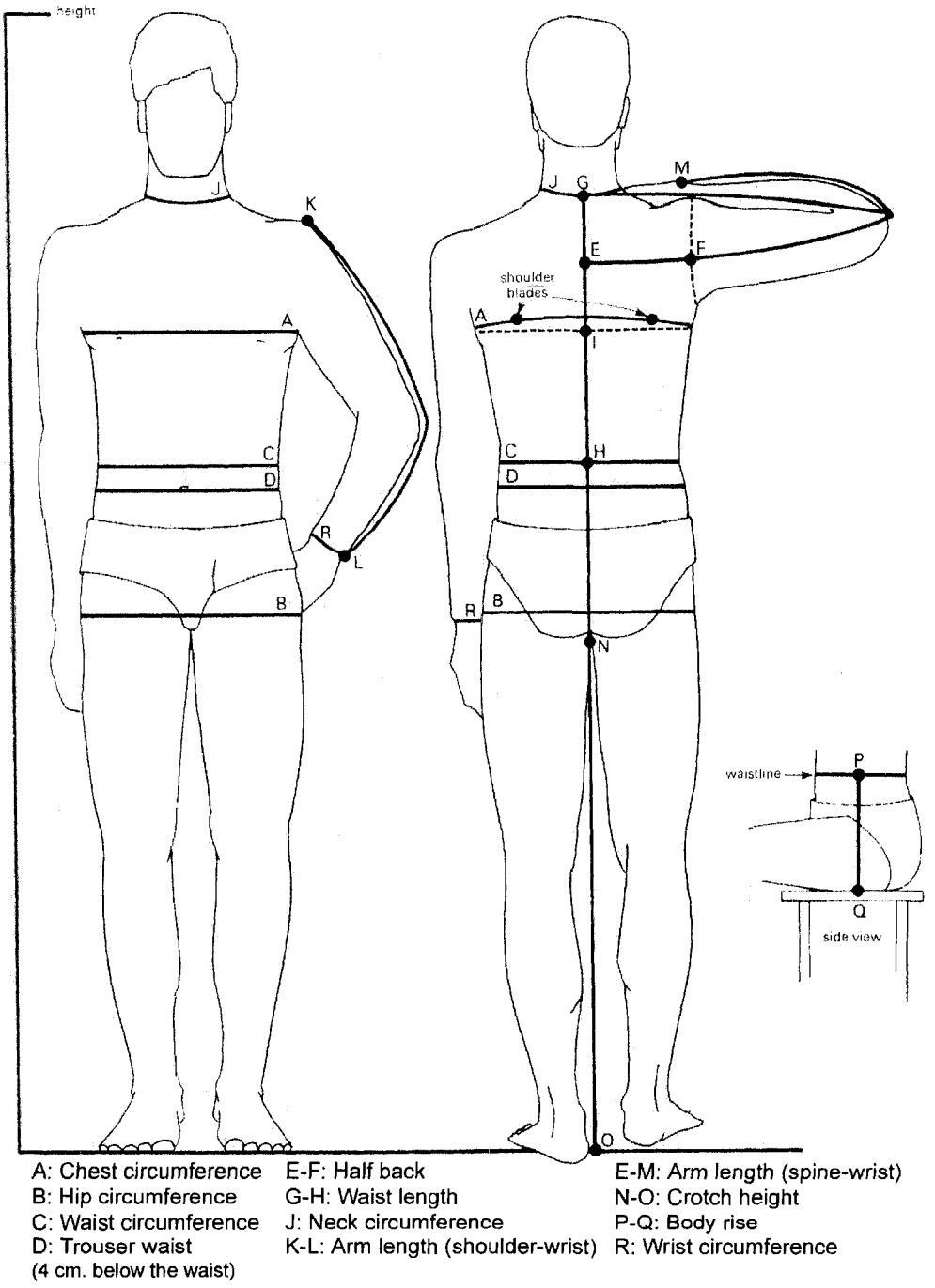


Figure 8.1: Standard body measurements for tailoring as presented in [Aldrich, 2006].

tify the archetypes. These archetypes may then be used for designing clothes that fit the population better. Next, rules describing the relationships between these measurements within each size may be used as constraints for the manufacture of clothes. One also may want to reduce the set of body measurements and identify the measurements that need special attention when designing garments. Finally, to better understand the demographic nature of potential customers and their needs, a clothing company may analyze demographic data and identify lifestyle characteristics within this group. This information may next be used to determine when and where advertising should be placed to obtain the maximum results.

In the previous scenario, we defined a specific garment, but a clothing designer may also wish to design clothes for specific sectors of the population e.g. women aged 55 and above (this group reported lack or poor fit in clothes in the US population [Salusso et al., 2006]) or XX-Large sized married males aged 45-50. From our study, we know XX-Large individuals are difficult to characterize and therefore to produce garments that fit them properly. From the demographic analysis we found XX-Large married males aged 45-50 reported high income. These groups are then potential customers for clothes that fit them well.

8.3 Contributions

This thesis focuses on applying utility-based data mining techniques to facilitate the design and manufacture of garments. We extend the work of Viktor et al. by applying the method they proposed in [Viktor et al., 2006] to another two populations and by considering utility-based data mining techniques. Through extending this work, this thesis makes four major contributions:

Comparative study among three populations: Viktor et al. [Viktor et al., 2006] analyzed the American male population. In this study, we analyzed the American female population, and the Italian and Dutch populations (males and females). After our analysis, we made a comparative analysis among the three populations based on the archetypes. We found that the male populations (American, Italian and Dutch) as well as the American female population are best characterized by five sizes, while the Italian and Dutch female populations are best described by six sizes. We also found that the Italians are the thinnest and shortest population, the Dutch are the tallest population and the Americans are the most robust population.

Comparative analysis of anthropometric and 3-D data clustering results: In this study we also analyzed the 3-D data and identified the archetypes. Next, we performed a comparative analysis of the anthropometric and 3-D data clustering results. We found that the clustering results of anthropometric and 3-D data are complementary, in the sense that each of them provides a different perspective on the subject characterization. That is, clustering of 3-D data groups together similar body shapes, while the clustering of anthropometric data groups together subjects with similar body measurements independently of the body shape. For applications such as tailoring, the most suitable approach is to consider the anthropometric data. For the design of e.g. masks and helmets, 3-D data may be more suitable because the goal, in this case, is to produce a good fitting for different shapes.

Account for economic factors of the data mining process: The objective of utility-based data mining is to consider all utility aspects in the data mining process and maximize the utility of the entire process. In this study, we reduce the cost and time of the mining process by identifying a reduced set of body measurements that preserves, or even improves, the performance of the full set of measurements. From the tailoring perspective, this reduced set of measurements may be viewed as the set of measurements that require special attention when designing clothes. As part of the account of utility factors, we also used interestingness measures in the analysis of demographic data to reduce the cost of the mining process and improve the quality of the results. This is further discussed in the next contribution.

Thorough demographic data analysis: In our study we were able to identify interrelationships between fitness, education, marital status, age, occupation and income within each size for the three populations. In our analysis we applied different objective interestingness measures. This allowed us to reduce the search space and, as consequence, improve the efficiency of the mining process. Applying interestingness measures also improved the quality of the mined results. That is, applying interestingness measures allowed us to identify the most interesting association within the demographic data. Using the analysis results, we defined a demographic profile and identified potential target markets.

8.4 Future work

Although we have made significant achievements as mentioned in the section of contributions, many research directions are still possible. Suggestions for future work include:

Subjective interestingness measures: In this work, we aimed to find general and reliable association rules. We therefore apply objective interestingness measures based on probability. As a future work we may apply subjective interestingness measures in order to identify novel, unexpected and actionable rules.

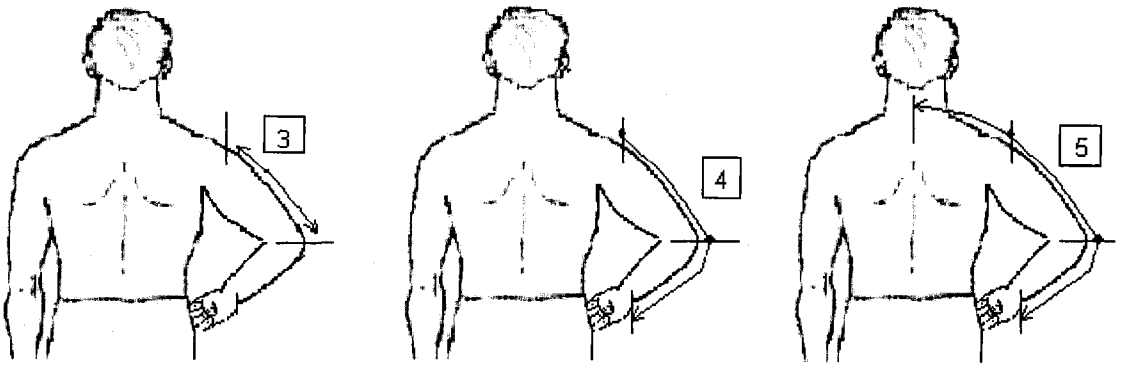
Sizing for the clothing industry: Sizing is the process to establish a size chart of key body measurements for a range of apparel sizes. Grading is the standard method of applying increases and decreases at points of a pattern to make the pattern larger or smaller. A grading system is developed from sizing specifications, and sizing specifications are derived from anthropometric data. As a future work, we may develop a grading system and propose size charts for the clothing industry.

Anthropometry of the disabled and elderly: The need for knowledge about variation of body dimensions, range of motion and strength data of elderly and disabled is an important issue, since these groups are much more dependant on the quality of their equipment and products than able-bodied people. As a future work, we aim to analyze anthropometric data for the disabled and elderly people.

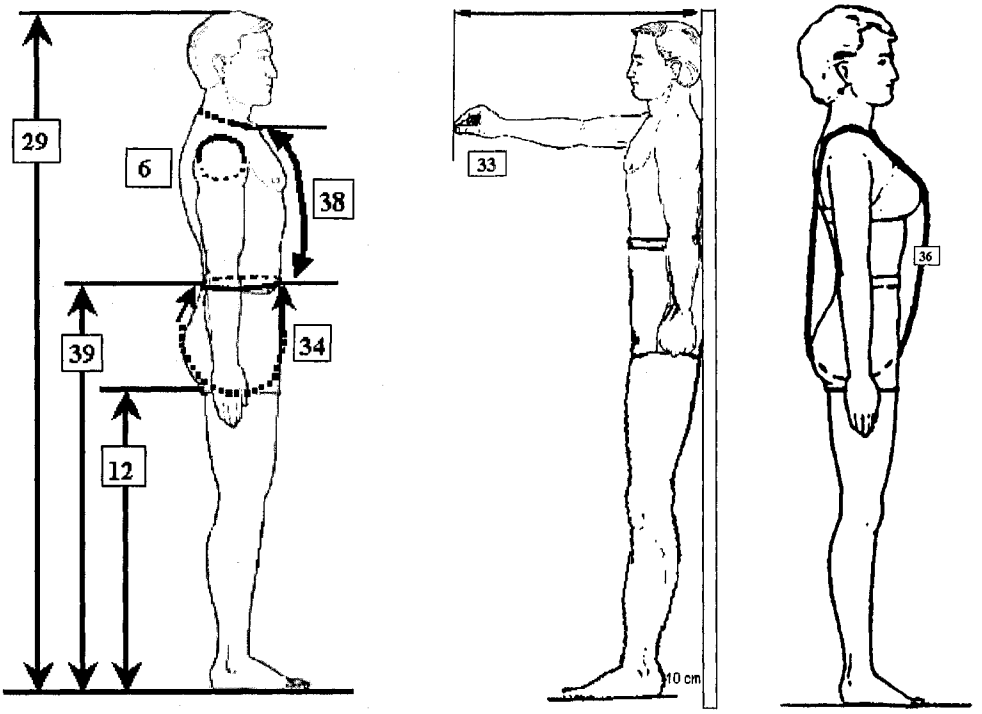
Appendix A

Traditional anthropometric measurements

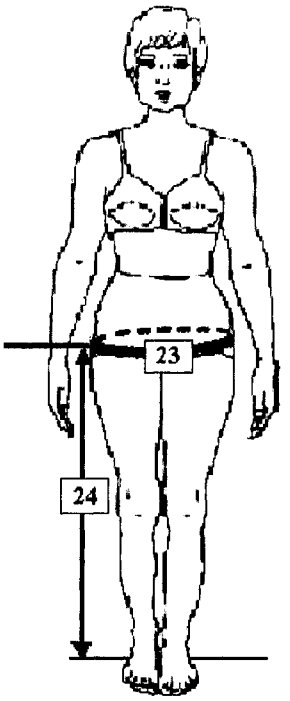
In this appendix we provide a visual description of the traditional anthropometric measurements as presented in [Robinette et al., 2002].

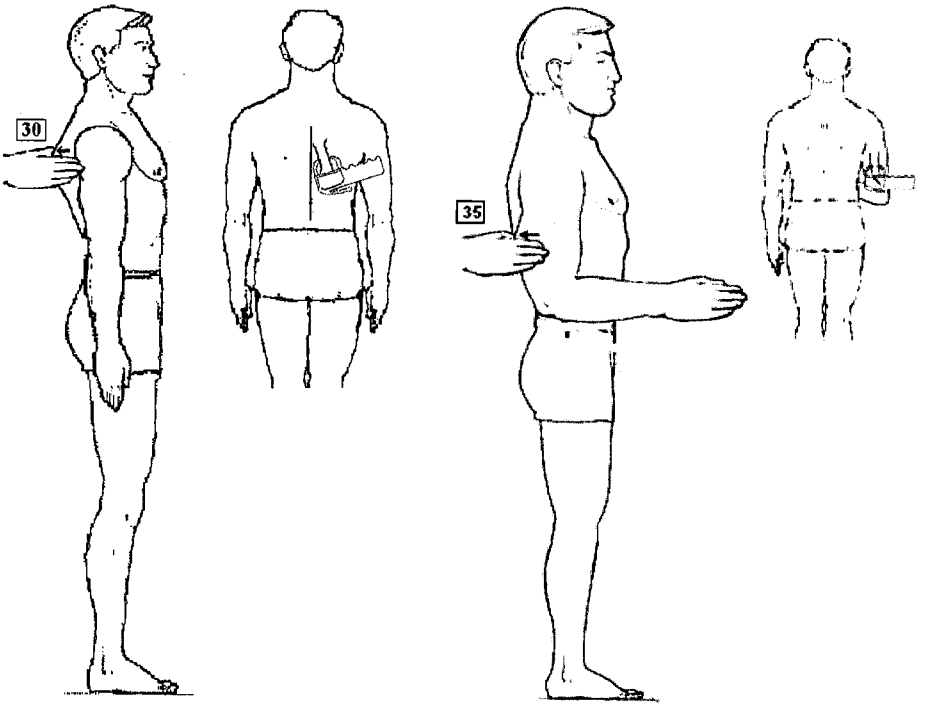


3. Arm Length (Shoulder-Elbow)
4. Arm Length (Shoulder-Wrist)
5. Arm Length (Spine-Wrist)

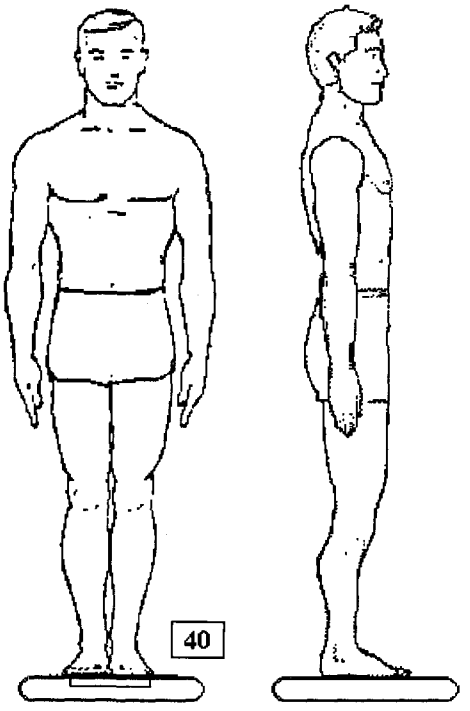


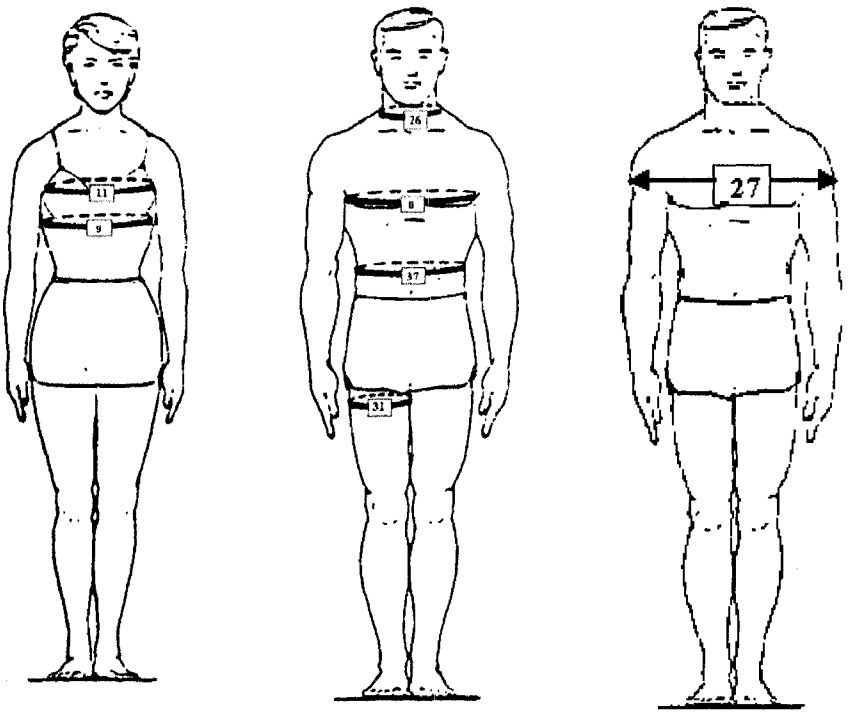
6. Armscye Circumference
(Scye Circumference over Acromion)
12. Crotch Height
23. Hip Circumference, Maximum
24. Hip Circumference, Maximum, Height
29. Stature (Body Height)
33. Thumb Tip Reach, Right
34. Total Crotch Length
36. Vertical Trunk Circumference, Right
38. Waist Front Length
39. Waist Height, Preferred



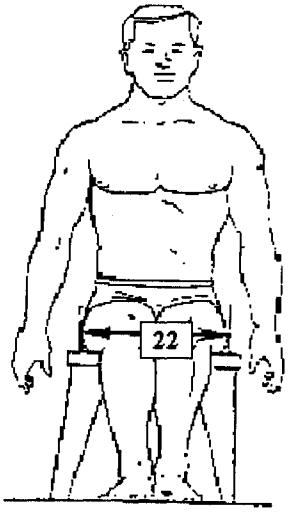


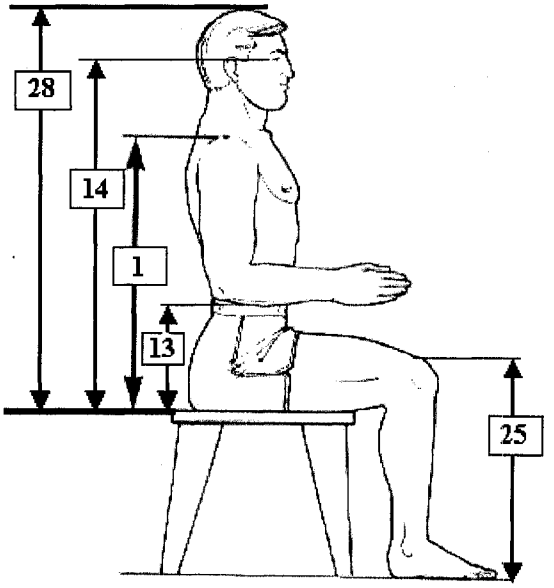
30. Subscapular Skinfold, Right
35. Triceps Skinfold
40. Weight (Mass)

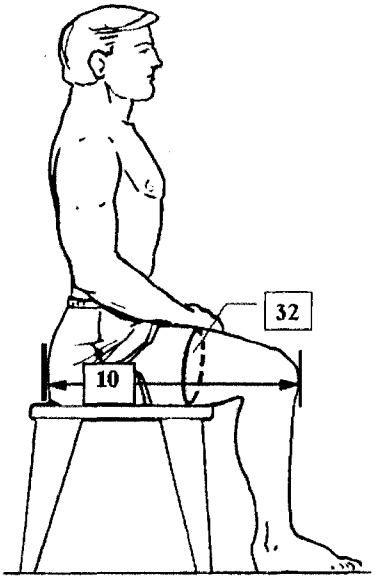




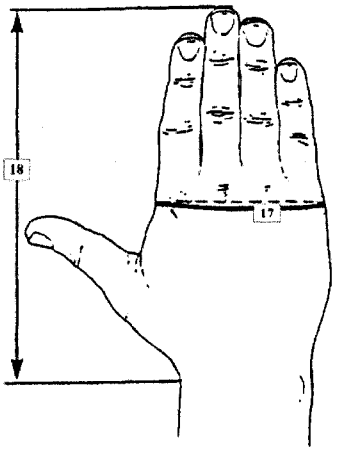
8. Bust/Chest Circumference
9. Bust/Chest Circumference under Bust
11. Chest Girth (Chest Circumference at Scye)
22. Hip Breadth, Sitting
26. Neck Base Circumference
27. Shoulder Breadth
31. Thigh Circumference, Maximum, Right
37. Waist Circumference, Preferred

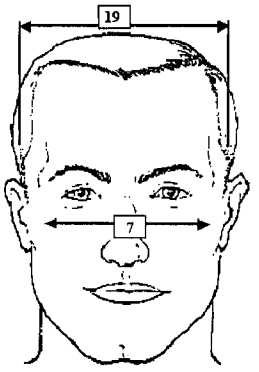


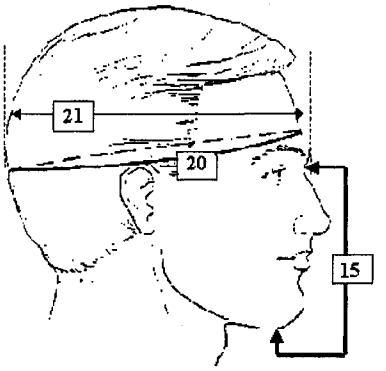




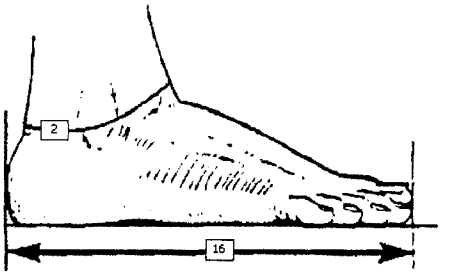
1. Acromial Height, Sitting
10. Buttock-Knee Length, Right
13. Elbow Height, Sitting, Right
14. Eye Height, Sitting, Right
17. Hand Circumference, Right
18. Hand Length, Right
25. Knee Height, Sitting, Right
28. Sitting Height
32. Thigh Circumference, Maximum, Sitting, Right







2. Ankle Circumference
7. Bizygomatic Breadth
15. Face Length (Menton-Sellion Length)
16. Foot Length, Right
19. Head Breadth
20. Head Circumference
21. Head Length



Appendix B

Cluster analysis

In this appendix we provide the clustering results of the different algorithms for the Italian and Dutch populations.

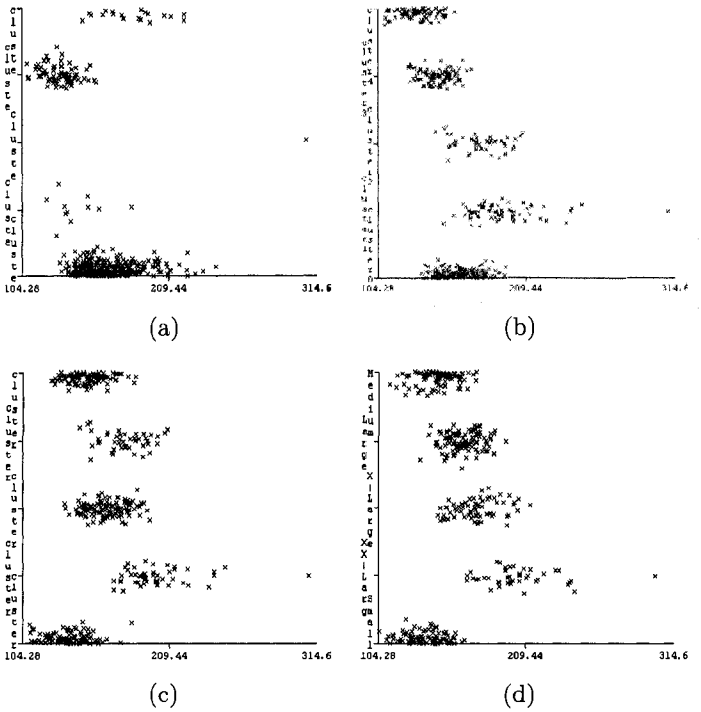


Figure B.1: Cluster visualization of the different algorithms for Italian males. (a) Farthest First algorithm, (b) EM algorithm, (c) K-means algorithm and (d) Density-based with k-means components algorithm.

	Small	Medium	Large	X-Large	XX-Large
Farthest First	76 (18%)	11 (3%)	306 (74%)	19 (5%)	1 (0%)
EM	71 (17%)	89 (22%)	142 (34%)	42 (10%)	69 (17%)
K-means	95 (23%)	108 (26%)	117 (28%)	44 (11%)	49 (12%)
Density-based	88 (21%)	107 (26%)	107 (26%)	69 (17%)	42 (10%)

Table B.1: Number of subjects per cluster for Italian males.

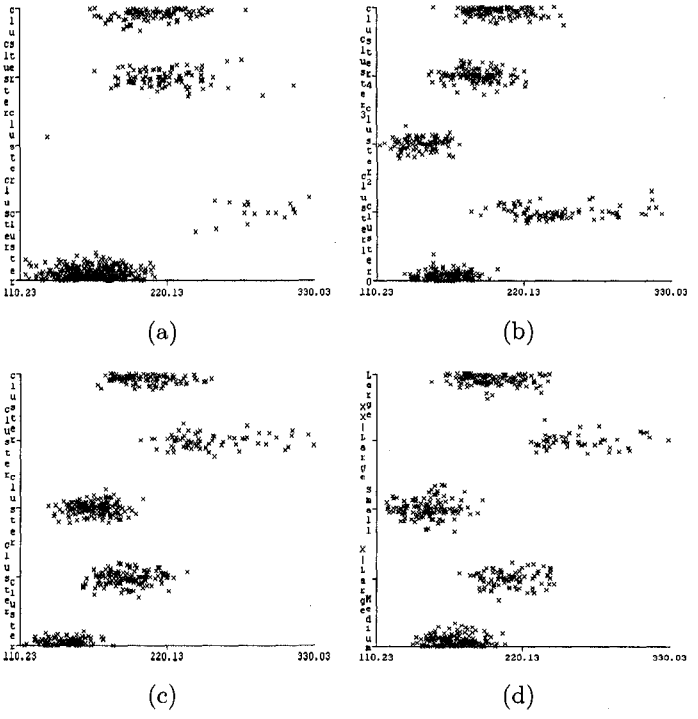


Figure B.2: Cluster visualization of the different algorithms for Dutch males. (a) Farthest First algorithm, (b) EM algorithm, (c) K-means algorithm and (d) Density-based with k-means components algorithm.

	Small	Medium	Large	X-Large	XX-Large
Farthest First	361 (64%)	1 (0%)	101 (18%)	87 (15%)	17 (3%)
EM	95 (17%)	148 (26%)	118 (21%)	123 (22%)	83 (15%)
K-means	96 (17%)	167 (29%)	131 (23%)	106 (19%)	67 (12%)
Density-based	126 (22%)	173 (31%)	139 (25%)	82 (14%)	47 (8%)

Table B.2: Number of subjects per cluster for Dutch males.

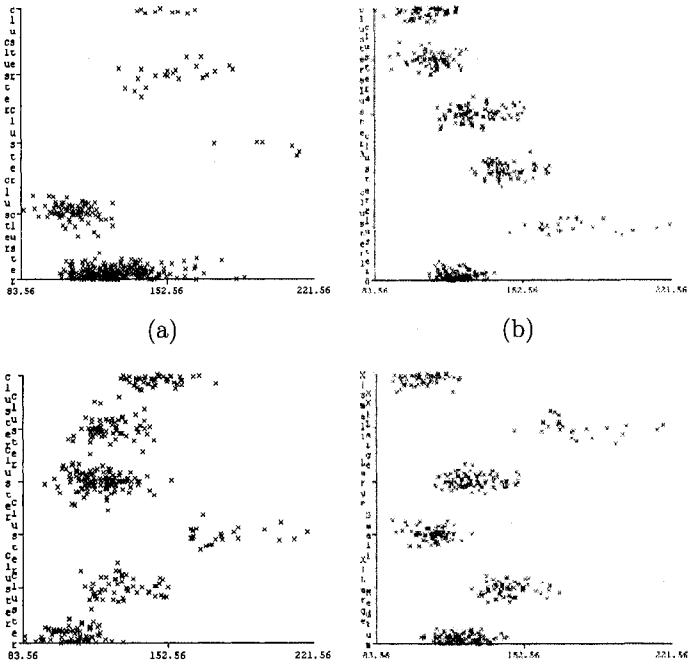


Figure B.3: Cluster visualization of the different algorithms for Italian females. (a) Farthest First algorithm, (b) EM algorithm, (c) K-means algorithm and (d) Density-based with k-means components algorithm.

	X-Small	**Small**	**Medium**	**Large**	**X-Large**	**XX-Large**
Farthest First	47 (12%)	134 (35%)	170 (44%)	22 (6%)	9 (2%)	6 (2%)
EM	53 (14%)	70 (18%)	99 (26%)	83 (21%)	59 (15%)	24 (6%)
K-means	74 (19%)	124 (32%)	65 (17%)	52 (13%)	50 (13%)	23 (6%)
Density-based	51 (13%)	65 (17%)	110 (28%)	82 (21%)	56 (14%)	24 (6%)

Table B.3: Number of subjects per cluster for Italian females.

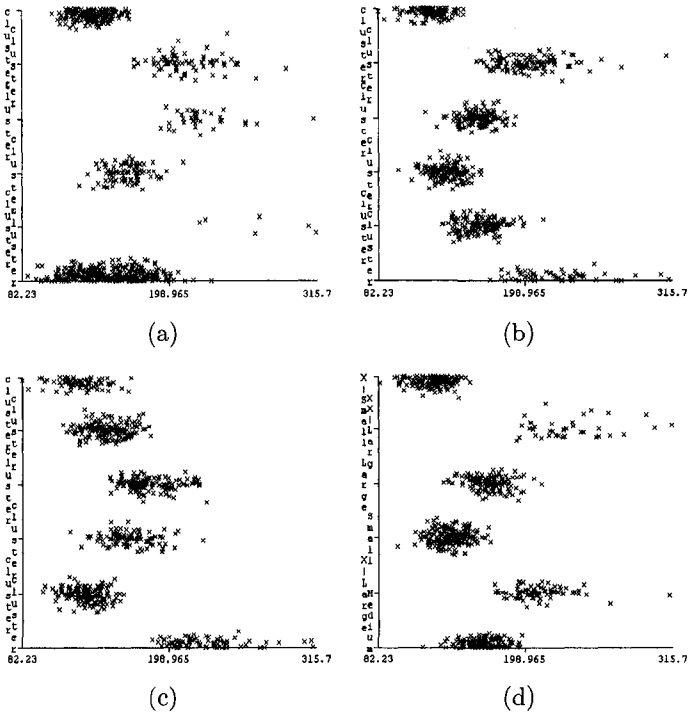


Figure B.4: Cluster visualization of the different algorithms for Dutch females. (a) Farthest First algorithm, (b) EM algorithm, (c) K-means algorithm and (d) Density-based with k-means components algorithm.

	X-Small	**Small**	**Medium**	**Large**	**X-Large**	**XX-Large**
Farthest First	329 (47%)	174 (25%)	87 (12%)	71 (10%)	33 (5%)	6 (1%)
EM	117 (17%)	158 (23%)	132 (19%)	139 (20%)	99 (14%)	55 (8%)
K-means	71 (10%)	148 (21%)	178 (25%)	98 (14%)	131 (19%)	74 (11%)
Density-based	130 (19%)	198 (28%)	125 (18%)	125 (18%)	83 (12%)	39 (6%)

Table B.4: Number of subjects per cluster for Dutch females.

Bibliography

[wik,] Wikipedia. Resource available at http://en.wikipedia.org/wiki/Anthropometrics.

[siz, 2001] (2001). Size UK. UK National Sizing Survey. Resource available at http://www.size.org/.

[siz, 2003] (2003). Size USA. The US National Size Survey. Resource available at http://www.sizeusa.com/.

[siz, 2007] (2007). Size China. Perfect fit for China. Resource available at http://www.sizechina.com/.

[Abdali et al., 2004] Abdali, O., Viktor, H. L., Paquet, E., and Rioux, M. (2004). Exploring anthropometric data through cluster analysis. *SAE 2004 Transactions Journal of Aerospace*, 113(1):241–244.

[Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *SIGMOD '93: Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, pages 207–216, New York, NY, USA. ACM.

[Agrawal and Srikant, 1994] Agrawal, R. and Srikant, R. (1994). Fast algorithms for mining association rules in large databases. In *VLDB '94: Proceedings of the 20th International Conference on Very Large Data Bases*, pages 487–499, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Aldrich, 2006] Aldrich, W. (2006). *Metric Pattern Cutting for Menswear*. Blackwell Publishing, fourth edition.

[Ashdown and Dunne, 2006] Ashdown, S. and Dunne, L. (2006). A Study of Automated Custom Fit: Readiness of the Technology for the Apparel Industry. *Clothing and Textiles Research Journal*, 24(2):121–136.

[Ashdown and Loker, 2005] Ashdown, S. and Loker, S. (2005). Improved Apparel Sizing: Fit and Anthropometric 3D Scan Data. *Annual Report NTC Project: S04-CR01. National Textile Center.*

[Azouz et al., 2006] Azouz, Z. B., Rioux, M., Shu, C., and Lepage, R. (2006). Characterizing human shape variation using 3D anthropometric data. *Visual Computer: International Journal of Computer Graphics*, 22(5):302–314.

[Becerra-Fernandez et al., 2002] Becerra-Fernandez, I., Zanakis, S. H., and Walczak, S. (2002). Knowledge discovery techniques for predicting country investment risk. *Comput. Ind. Eng.*, 43(4):787–800.

[Bhatnagar et al., 2005] Bhatnagar, V., Al-Hegami, A. S., and Kumar, N. (2005). Novelty as a measure of interestingness in knowledge discovery. *International Journal of Information Technology*, 2(1).

[Blackwell et al., 2002] Blackwell, S., Robinette, K. M., Daanen, H., Boehmer, M., Fleming, S., Kelly, S., Brill, T., Hoeferlin, D., and Burnsides, D. (2002). Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report, Volume II: Descriptions. *AFRL-HE-WP-TR-2002-0173, United States Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division, 2255 H Street, Wright-Patterson AFB OH 45433-7022.*

[Blum and Mitchell, 1998] Blum, A. and Mitchell, T. (1998). Combining labeled and unlabeled data with co-training. In *COLT' 98: Proceedings of the eleventh annual conference on Computational learning theory*, pages 92–100, New York, NY, USA. ACM.

[Brin et al., 1997a] Brin, S., Motwani, R., and Silverstein, C. (1997a). Beyond Market Baskets: Generalizing Association Rules to Correlations. In Peckham, J., editor, *SIGMOD 1997, Proceedings ACM SIGMOD International Conference on Management of Data, May 13-15, 1997, Tucson, Arizona, USA*, pages 265–276. ACM Press.

[Brin et al., 1997b] Brin, S., Motwani, R., Ullman, J. D., and Tsur, S. (1997b). Dynamic itemset counting and implication rules for market basket data. In *SIGMOD '97: Proceedings of the 1997 ACM SIGMOD international conference on Management of data*, pages 255–264, New York, NY, USA. ACM.

[Brunsman et al., 1997] Brunsman, M., Daanen, H. M., and Robinette, K. M. (1997). Optimal postures and positioning for human body scanning. In *NRC '97: Proceedings of the International Conference on Recent Advances in 3-D digital Imaging and Modeling*, pages 266–273, Los Alamitos, CA, USA. IEEE Computer Society.

[Chae et al., 2001] Chae, Y. M., Ho, S. H., Cho, K. W., Lee, D. H., and Ji, S. H. (2001). Data Mining approach to Policy Analysis in a Health Insurance Domain. *International Journal of Medical Informatics*, 62(2-3):103–111.

[Chen et al., 1996] Chen, M.-S., Han, J., and Yu, P. S. (1996). Data mining: an overview from a database perspective. *IEEE Transactions on Knowledge And Data Engineering*, 8:866–883.

[Chen and Wu, 2006] Chen, X. and Wu, Y.-F. (2006). Personalized knowledge discovery: Mining novel association rules from text. In *Proceedings of the 2006 SIAM International Conference on Data Mining*, pages 67–71, Bethesda, Maryland.

[Cohen, 1995] Cohen, W. W. (1995). Fast effective rule induction. In Prieditis, A. and Russell, S., editors, *Proc. of the 12th International Conference on Machine Learning*, pages 115–123, Tahoe City, CA. Morgan Kaufmann.

[Cunningham, 2007] Cunningham, P. (2007). Dimension Reduction. *Technical Report UCD-CSI-2007-7, University College Dublin*, pages 1–24.

[Das, 2001] Das, S. (2001). Filters, Wrappers and a Boosting-Based Hybrid for Feature Selection. In *ICML '01: Proceedings of the Eighteenth International Conference on Machine Learning*, pages 74–81, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Dasgupta and Long, 2005] Dasgupta, S. and Long, P. M. (2005). Performance guarantees for hierarchical clustering. *Journal of Computer and Systems Science*, 70(4):555–569.

[Dash et al., 2000] Dash, M., Liu, H., and Motoda, H. (2000). Consistency based feature selection. In *PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*, pages 98–109, London, UK. Springer-Verlag.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum Likelihood from Incomplete Data via the EM Algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38.

[Devarajan and Istook, 2004] Devarajan, P. and Istook, C. L. (2004). Validation of Female Figure Identification Technique (FFIT) for Apparel software. *Journal of Textile and Apparel, Technology and Management*, 4(1):1–23.

[Dong and Li, 1998] Dong, G. and Li, J. (1998). Interestingness of discovered association rules in terms of neighborhood-based unexpectedness. In *PAKDD '98: Proceedings of the Second Pacific-Asia Conference on Research and Development in Knowledge Discovery and Data Mining*, pages 72–86, Melbourne, Australia. Springer-Verlag.

[Dzeroski and Lavrac, 2001] Dzeroski, S. and Lavrac, N. (2001). editors, Relational Data Mining. Springer-Verlag, Berlin.

[Elkan, 2001] Elkan, C. (2001). The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Articial Intelligence*, pages 973–978.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, 17:37–54.

[Fodor, 2002] Fodor, I. K. (2002). A survey of dimension reduction techniques. *Center for Applied Scientific Computing, Lawrence Livermore National Laboratory*, pages 1–18.

[Frank and Witten, 1998] Frank, E. and Witten, I. H. (1998). Generating accurate rule sets without global optimization. In *ICML '98: Proceedings of the Fifteenth International Conference on Machine Learning*, pages 144–151, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Frawley et al., 1992] Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992). Knowledge discovery in databases - an overview. *AI Magazine*, 13:57–70.

[Geng and Hamilton, 2006] Geng, L. and Hamilton, H. J. (2006). Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9.

[Gennari et al., 1989] Gennari, J. H., Langley, P., and Fisher, D. H. (1989). Models of incremental concept formation. *Artificial Intelligence*, 40(1-3):11–61.

[Goldberg, 1989] Goldberg, D. E. (1989). *Genetic Algorithms in Search, Optimization and Machine Learning.* Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA.

[Gordon et al., 1989] Gordon, C. C., Churchill, T., Clauser, C. E., Bradtmiller, B., and McConville, J. T. (1989). Anthropometric Survey of U.S. Army Personnel: Summary Statistics, Interim Report for 1988. *Technical rept. 1987-1988*, pages 001–335.

[Guo and Viktor, 2005] Guo, H. and Viktor, H. L. (2005). Mining relational databases with multi-view learning. In *MRDM '05: Proceedings of the 4th international workshop on Multi-relational mining*, pages 15–24, New York, NY, USA. ACM.

[Hall, 2000] Hall, M. A. (2000). Correlation-based feature selection for discrete and numeric class machine learning. In *ICML '00: Proceedings of the Seventeenth International Conference on Machine Learning*, pages 359–366, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

[Han and Kamber, 2006] Han, J. and Kamber, M. (2006). *Data Mining: Concepts and Techniques.* Morgan Kaufmann, second edition.

[Heinz et al., 2003] Heinz, G., Peterson, L. J., Johnson, R. W., and Kerk, C. J. (2003). Exploring relationships in body dimensions. *Journal of Statistics Education*, 11(2). Resource available at http://www.amstat.org/publications/jse/v11n2/datasets.heinz.html.

[Hilderman and Hamilton, 2000] Hilderman, R. J. and Hamilton, H. J. (2000). Applying objective interestingness measures in data mining systems. In *PKDD '00: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 432–439, London, UK. Springer-Verlag.

[Hilderman and Hamilton, 2001] Hilderman, R. J. and Hamilton, H. J. (2001). *Knowledge Discovery and Measures of Interest.* Kluwer Academic Publishers, Norwell, MA, USA.

[Hsu et al., 2007] Hsu, C.-H., Lin, H.-F., and Wang, M.-J. (2007). Developing female size charts for facilitating garment production by using data mining. *Journal of Chinese Institute of Industrial Engineers*, 24(3):245–251.

[Hsu and Wang, 2005] Hsu, C.-H. and Wang, M.-J. J. (2005). Using decision tree-based data mining to establish a sizing system for the manufacture of garments. *International Journal of Advanced Manufacturing Technology*, 26(5–6):669–674.

[Infometrics, 2004] Infometrics (2004). New zealand's angel capital market: The supply side. *Final report on New Zealand's Angel Capital Market from a supply side perspective*, pages 1–69.

[Kim et al., 2003] Kim, Y., Street, W. N., and Menczer, F. (2003). Feature selection in data mining. *Data mining: opportunities and challenges*, pages 80–105.

[Krogel and Wrobel, 2003] Krogel, M.-A. and Wrobel, S. (2003). Facets of aggregation approaches to propositionalization. pages 30–39. Department of Informatics, University of Szeged.

[Lenca et al., 2008] Lenca, P., Meyer, P., Vaillant, B., and Lallich, S. (2008). On selecting interestingness measures for association rules: User oriented description and multiple criteria decision aid. *European Journal of Operational Research*, 127(2):610–626.

[Li and Hamilton, 2004] Li, G. and Hamilton, H. J. (2004). Basic association rules. In *Proceedings of the 4th SIAM International Conference on Data Mining*, pages 166–177, Orlando, FL.

[Ling et al., 2002] Ling, C. X., Chen, T., Yang, Q., and Cheng, J. (2002). Mining Optimal Actions for Profitable CRM. In *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*, page 767, Washington, DC, USA. IEEE Computer Society.

[Liu et al., 1999] Liu, B., Hsu, W., Mun, L.-F., and Lee, H.-Y. (1999). Finding interesting patterns using user expectations. *IEEE Transactions on Knowledge and Data Engineering*, 11(6):817–832.

[Lu et al., 2001] Lu, S., Hu, H., and Li, F. (2001). Mining weighted association rules. *Intell. Data Anal.*, 5(3):211–225.

[MacQueen, 1967] MacQueen, J. B. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 1:281–297. University of California Press.

[Maddouri and Elloumi, 2002] Maddouri, M. and Elloumi, M. (2002). A data mining approach based on machine learning techniques to classify biological sequences. *Knowledge-Based Systems*, 15(4):217–223.

[McGarry, 2005] McGarry, K. (2005). A survey of interestingness measures for knowledge discovery. *Knowl. Eng. Rev.*, 20(1):39–61.

[Mills, 1955] Mills, F. (1955). *Statistical Methods*. Pitman.

[Ohsaki et al., 2004] Ohsaki, M., Sato, Y., Kitaguchi, S., Yokoi, H., and Yamaguchi, T. (2004). Comparison between objective interestingness measures and real human interest in medical data mining. In *IEA/AIE'2004: Proceedings of the 17th international conference on Innovations in applied artificial intelligence*, pages 1072–1081. Springer Springer Verlag Inc.

[Padmanabhan and Tuzhilin, 1999] Padmanabhan, B. and Tuzhilin, A. (1999). Unexpectedness as a measure of interestingness in knowledge discovery. *Decis. Support Syst.*, 27(3):303–318.

[Paquet et al., 2000] Paquet, E., Robinette, K. M., and Rioux, M. (2000). Management of three-dimensional and anthropometric databases: Alexandria and Cleopatra. *Journal of Electronic Imaging*, 9:421–431.

[Paquet et al., 2007] Paquet, E., Viktor, H. L., Guo, H., and Sanchez, I. P. (2007). Constrained Virtual Tailoring from Anthropometric Data, 3-D Shape and Data Mining. In *Proceedings of the Second International WEAR Conference (WEAR'02)*, Alberta, Canada.

[Piatetsky-Shapiro, 1991] Piatetsky-Shapiro, G. (1991). Discovery, analysis, and presentation of strong rules. In Piatetsky-Shapiro, G. and Frawley, W., editors, *Knowledge Discovery in Databases*, pages 229–248. AAAI/MIT Press, Cambridge, MA.

[Provost, 2005] Provost, F. (2005). Toward economic machine learning and utility-based data mining. In *UBDM '05: Proceedings of the 1st international workshop on Utility-based data mining*, pages 1–1, New York, NY, USA. ACM.

[Quinlan, 1986] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, 1(1):81–106.

[Quinlan, 1993] Quinlan, J. R. (1993). *C4.5: programs for machine learning.* Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

[Rissanen, 1978] Rissanen, J. (1978). Modelling by the shortest data description. *Automatica*, 14:465–471.

[Robinette et al., 2002] Robinette, K. M., Blackwell, S., Daanen, H., Fleming, S., Boehmer, M., Brill, T., Hoeferlin, D., and Burnsides, D. (2002). Civilian American and European Surface Anthropometry Resource (CAESAR), Final Report, Volume I: Summary. *AFRL-HE-WP-TR-2002-0169, United States Air Force Research Laboratory, Human Effectiveness Directorate, Crew System Interface Division, 2255 H Street, Wright-Patterson AFB OH 45433-7022.*

[Roebuck, 1993] Roebuck, J. A. (1993). *Anthropometric methods: Designing to Fit the Human Body.* Monographs in Human Factors and Ergonomics, first edition.

[Sahar, 1999] Sahar, S. (1999). Interestingness via what is not interesting. In *KDD '99: Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 332–336, New York, NY, USA. ACM.

[Salusso et al., 2006] Salusso, C. J., Borkowski, J. J., Reich, N., and Goldsberry, E. (2006). An alternative approach to sizing apparel for women 55 and older. *Clothing and Textiles Research Journal*, 24(2).

[Scheffer, 2001] Scheffer, T. (2001). Finding association rules that trade support optimally against confidence. In *PKDD '01: Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery*, pages 424–435, London, UK. Springer-Verlag.

[Schofield and LaBat, 2005] Schofield, N. A. and LaBat, K. L. (2005). Exploring the relationships of grading, sizing and anthropometric data. *Clothing and Textiles Research Journal*, 23(1):13–27.

[Sha and Liu, 2005] Sha, D. and Liu, C.-H. (2005). Using data mining for due date assignment in a dynamic job shop environment. *International Journal of Advanced Manufacturing Technology*, 25(11-12):1164–1174.

[Sheikh et al., 2004] Sheikh, L. M., Tanveer, B., and Hamdani, M. A. (2004). Interesting measures for mining association rules. *Multitopic Conference, 2004. Proceedings of INMIC 2004, 8th International*, pages 641–644.

[Silberschatz and Tuzhilin, 1996] Silberschatz, A. and Tuzhilin, A. (1996). What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and Data Engineering*, 08(6):970–974.

[Smith, 2002] Smith, L. I. (2002). A tutorial on principal components analysis. pages 1–26.

[Sotiris Kotsiantis, 2006] Sotiris Kotsiantis, D. K. (2006). Association rules mining: A recent overview. *GESTS International Transactions on Computer Science and Engineering*, 32(1):71–82.

[Tan et al., 2002] Tan, P.-N., Kumar, V., and Srivastava, J. (2002). Selecting the right interestingness measure for association patterns. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 32–41, New York, NY, USA. ACM.

[Tan et al., 2005] Tan, P.-N., Steinbach, M., and Kumar, V. (2005). *Introduction to Data Mining*. Addison Wesley, first edition.

[Veitch et al., 2007] Veitch, D., Veitch, L., and Henneberg, M. (2007). Sizing for the Clothing Industry Using Principal Component Analysis - An Australian Example.

[Viktor et al., 2006] Viktor, H. L., Paquet, E., and Guo, H. (2006). Measuring to fit: Virtual tailoring through cluster analysis and classification. In *PKDD 2006: Knowledge Discovery in Databases*, pages 395–406.

[Wang et al., 2002] Wang, K., Zhou, S., and Han, J. (2002). Profit mining: From patterns to actions. In *EDBT '02: Proceedings of the 8th International Conference on Extending Database Technology*, pages 70–87, London, UK. Springer-Verlag.

[Webb and Brain, 2002] Webb, G. I. and Brain, D. (2002). Generality is predictive of prediction accuracy. In *Proceedings of the 2002 Pacific Rim Knowledge Acquisition Workshop (PKAW 2002)*, pages 117–130.

[Weiss et al., 2005] Weiss, G., Saar-Tsechansky, M., and Zadrozny, B. (2005). Report on UBDM-05: Workshop on Utility-Based Data Mining. *SIGKDD Explor. Newsl.*, 7(2):145–147.

[Weiss and Tian, 2006] Weiss, G. M. and Tian, Y. (2006). Maximizing classifier utility when training data is costly. *SIGKDD Explor. Newsl.*, 8(2):31–38.

[Witten and Frank, 2005] Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques.* Morgan Kaufmann, second edition.

[Yu and Liu, 2003] Yu, L. and Liu, H. (2003). Feature selection for high-dimensional data: A fast correlation-based filter solution. In *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)*, pages 856–863, Washington DC.

[Zadrozny et al., 2006] Zadrozny, B., Weiss, G., and Saar-Tsechansky, M. (2006). UBDM 2006: Utility-Based Data Mining 2006 workshop report. *SIGKDD Explor. Newsl.*, 8(2):98–101.

[Zbidi et al., 2006] Zbidi, N., Faiz, S., and Limam, M. (2006). On mining summaries by objective measures of interestingness. *Mach. Learn.*, 62(3):175–198.