NRC Publications Archive Archives des publications du CNRC

Finding Clothing that Fit through Cluster Analysis and Objective Interestingness Measures

Pena, Isis; Viktor, Herna L.; Paquet, Eric

This publication could be one of several versions: author's original, accepted manuscript or the publisher's version. /
La version de cette publication peut être l'une des suivantes : la version prépublication de l'auteur, la version acceptée du manuscrit ou la version de l'éditeur.

Publisher's version / Version de l'éditeur:

Proceedings of the 11th International Conference on Data Warehousing and Knowledge Discovery (DaWaK 2009), 2009-09-04

NRC Publications Record / Notice d'Archives des publications de CNRC:

<https://nrc-publications.canada.ca/eng/view/object/?id=67372c94-a5ab-4ebd-aacd-42ce18e97d35>
<https://publications-cnrc.canada.ca/fra/voir/objet/?id=67372c94-a5ab-4ebd-aacd-42ce18e97d35>

Access and use of this website and the material on it are subject to the Terms and Conditions set forth at

<https://nrc-publications.canada.ca/eng/copyright>

READ THESE TERMS AND CONDITIONS CAREFULLY BEFORE USING THIS WEBSITE.

L'accès à ce site Web et l'utilisation de son contenu sont assujettis aux conditions présentées dans le site

<https://publications-cnrc.canada.ca/fra/droits>

LISEZ CES CONDITIONS ATTENTIVEMENT AVANT D'UTILISER CE SITE WEB.

Questions? Contact the NRC Publications Archive team at

PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca. If you wish to email the authors directly, please see the first page of the publication for their contact information.

Vous avez des questions? Nous pouvons vous aider. Pour communiquer directement avec un auteur, consultez la première page de la revue dans laquelle son article a été publié afin de trouver ses coordonnées. Si vous n'arrivez pas à les repérer, communiquez avec nous à PublicationsArchive-ArchivesPublications@nrc-cnrc.gc.ca.

Finding Clothing that Fit through Cluster Analysis and Objective Interestingness Measures

Isis Peña¹, Herna L Viktor¹, Eric Paquet^{1,2}

¹ School of IT and Engineering, University of Ottawa, Ottawa, Canada

² National Research Council of Canada, Ottawa, Canada

ipena011@uottawa.ca, hlviktor@site.uottawa.ca, eric.paquet@nrc-cnrc.gc.ca

Abstract. Clothes should fit consumers well, be aesthetically pleasing and comfortable. However, repeated studies of customers' levels of satisfaction indicate that this is often not the case. For example, more robust males often find it difficult to find pants that are the correct length and fit their waists well. What, then, are the typical body profiles of the population? Would it be possible to identify the measurements that are of importance for different sizes and genders? Furthermore, assuming that we have access to an anthropometric database would there be a way to guide the data mining process to discover only those relevant body measurements that are of the most interest for apparel designers? This paper describes our results when addressing these questions through cluster analysis and interestingness measures-based feature selection. We explore a database containing anthropometric measurements as well as 3-D body scans, of a representative sample of the Dutch population.

Keywords: Utility-based data mining, anthropometry, interestingness measures, feature selection, cluster analysis, classification, relational and 3-D data

1 Introduction

Apparel manufacturers develop sizing systems with the goal of satisfying consumers' 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, it follows that the sizes must correspond to real grouping. However, this is often not the case, as indicated by the results obtained by Shofield and LaBat [1]. Their study of forty size charts for women's clothing showed that the different sizes are defined using arbitrary constant intervals between sizes, all vertical measurements increase as the size increases and that the differences between the principal girths are constant for all sizes. Considering this situation, it is easily understandable that repeated studies of the degree of satisfaction with apparel show that consumers' needs are not being met. For example, a North American study found

that about 50% of women and 62% of men cannot find satisfactorily fitting clothes [2]. According to Ashdown et al. [3] two main issues have limited 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. Second, there is a lack of information about the principal aspects to consider when designing garments for a variety of body sizes and shapes.

Recent work has addressed, to some extent, these problems. Anthropometric surveys such as the *CAESARTM project* [4], *Size USA* [5] and *Size UK* [6] have been carried out on civilian populations. The *CAESARTM project* [4] includes several body measurements such as waist circumference, hip circumference, height, weight, etc. together with 3-D body scans of each participant. In *Size USA* [5] and *Size UK* [6] subjects were scanned in 3-D, and the body measurements were then extracted from the 3-D scans. Some recent studies attempt to find the most important aspects to be taken into account when designing garments. Viktor et al. [7] finds body size groupings in a sample of the North American male population. Veitch et al. [8] aim to produce a well fitting bodice for Australian women. After selecting twelve out of fifty-four measures and applying Principal Component Analysis (PCA), they define thirty-six categories: twelve sizes and three body shapes within each size. Hsu et al. [9] identify three body types and thirty-eight sizes for the female adult Taiwanese population by applying PCA on eleven anthropometric measures.

Although the abovementioned work attempt to address the problem of identifying the main aspects that should be considered for the design of garments, they focus only either on a specific body part, or on a gender. Moreover, they do not account for the economic factors of the data mining process. That is, they do not attempt to find the subset of body measurements that would be of the most interest when designing apparel for different sizes within each gender. This paper addresses this issue. Our goals are as follows. We aim to understand the typical consumers' body profile by identifying the natural body size groups and their distinctive characteristics. Also, we attempt to find the most important body measurements that define each size, and study how these measurements interrelate. Importantly, we aim to reduce the cost of the mining process, and the subsequent cost of apparel design, by reducing the number of body measurements to be used. To this end, we employ interestingness measures to identify the minimal sets of body measurements that are relevant for the different sizes, within each gender. In this way, we obtain reduced body measurements (both anthropometric and 3-D) of high utility, to be used to optimize apparel design.

This paper is organized as follows. Section 2 introduces utility-based data mining and, in particular, objective interestingness measures, which is used to guide our data mining process. This section explains the need for, and usefulness of, these measures when evaluating the results of the data mining process and when used for feature selection. This is followed, in Section 3, with an introduction to the *CAESARTM* anthropometric database. This section also explains our methodology and results when characterizing the population. In this section, we describe the cluster analysis of both the anthropometric measurements and 3-D shapes. In Section 4, we present the approach followed when reducing the number of body measurements, through the use of interestingness measures-based feature selection together with feature extraction.

This section also discusses the results obtained for the various body sizes, within each gender. Section 5 concludes the paper.

2 Finding Interesting Patterns

Over the past years, research in data mining mainly focused on producing accurate models with little consideration of the economic aspects of the data mining process. For most real scenarios such as our case study, however, it becomes important to consider economic aspects of the mining process. This has led to the emerging field of *utility-based data mining* which aims to consider all utility aspects in the data mining process, and maximize the utility of the entire process [10, 11]. In particular, a subset of utility-based approaches focuses on reducing the time and mining space cost by using *interestingness measures*. Interestingness measures are “measures” that narrow the search space and find the truly interesting patterns allowing pruning, evaluating, selecting and ranking patterns according to their potential interest to the user. Thus, using interestingness measures to identify interesting patterns allows the time and space cost of the mining process to be reduced [12]. Some criteria have been proposed to determine whether a pattern is interesting. These criteria are: *Generality, Reliability, Conciseness, Peculiarity, Diversity, Novelty, Surprisingness, Utility* and *Actionability* [12, 13]. The first five criteria are considered objective while the remaining four are considered subjective. Based on these criteria, interestingness measures are then divided into two categories: *objective* and *subjective*. For classification rules, the most important role of objective interestingness measures is to act as heuristics. However, interestingness measures such as *Information gain, Gain ratio* and *Chi-squared* are not only used as heuristics, but are also used for *feature selection* [14, 15]. In our case study, we consider the results of the data mining process to be interesting if it is able to correctly and accurately characterize the bodies of different sizes, within each gender. Furthermore, our aim is to find only those body measurements which are crucial when designing a garment for a size and therefore need special consideration to ensure that the clothes fit well.

3 Characterization of the Population

One of the most important challenges for the clothing industry is to produce garments with quality fit. Poor fitting garments may never be sold or customers may return them. In order to produce better fitting garments, accurate and up-to-date measurements need to be further analyzed in order to be able to better characterize the population [3]. To address the aforementioned issue, we aim to find the natural body size groupings using the anthropometric measurements and the 3-D shapes as contained in the CAESARTM anthropometric database. From these groups we identify size *Archetypes* and their most important characteristics.

The CAESARTM database [4] includes traditional anthropometric measurements of a large number of individuals from North America, Italy and the Netherlands. This number is forty-four (44) for the males, and forty-five (45) for the females, since

under bust circumference is not appropriate for the male subjects. These measurements include height, weight, acromial height, waist circumference, thigh circumference and foot length, amongst others, which were recorded by domain experts. Additionally, the shape of each person was scanned in three dimensions using a full body scanner. That is, a laser scanning device measured and recorded detailed geometry of the subjects' body surface. The 3-D body scans were described using a global shape-based descriptor, which is an abstract and compact representation of the three-dimensional shape of the corresponding body. In essence, each scan is represented by a set of three histograms, which constitutes a 3-D shape index or descriptor for the human body [16]. This index characterizes the radial and angular distribution of the surface elements associated with a given body. The index is designed to be orientation invariant and robust against pose variation [16].

All our experiments are implemented in WEKA, a collection of machine learning algorithms for data mining tasks [17]. We use the Cleopatra visual data mining tool [16] to determine whether our results correspond to reality. We proceed as follows. Firstly, we use the anthropometric data to find the clusters therein, as described in Section 3.1. Next, we determine the Centroids of each cluster. For each Centroid, we use the Cleopatra system to retrieve the human body corresponding to it, together with the N nearest neighbours, where N is the number of bodies in this cluster. We verify the cluster membership by visually inspecting the results. This process is repeated for the 3-D body scans, as shown in Section 3.2. Next, we apply dimensionality reduction, as discussed in Section 4, to find reduced sets of anthropometric body measurements, and 3-D descriptors, respectively. In this study we consider the Dutch population, both males and females. The data was then first separated based on the gender of the subjects. The resulting sets consist of 567 males and 700 females.

3.1 Clustering of Anthropometric Measurements

In order to identify the natural body size groupings, we first apply cluster analysis techniques to the anthropometric data. Cluster analysis is an unsupervised learning data mining technique used to partition a set of physical or abstract objects into subsets or clusters based on data similarity [14, 17].

In the context of tailoring, the ideal scenario is to cover the greatest number of people with the fewest number of sizes [9]. Therefore, we aim to find the minimum number of clusters that fully characterize the population. 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 three clusters. Then, by inspecting the cluster distribution we decide whether is worth splitting the clusters, as described in [17]. This process is repeated until the clusters appear well-defined.

Fig. 1. Cluster visualization for (a) male population and (b) female population. The y axis represent the clusters, the x axis is the weight range.

For our cluster analysis experimentation, a number of clustering algorithms was considered. These included partitioning, hierarchical, density-based, model-based and grid-based approaches. By inspection of the cluster distribution and through the analysis of the results using Cleopatra, we found that the male population is best characterized with five clusters and the female population with six clusters. We also observed that the best partition is achieved using a *density-based* algorithm with k-means components. The clusters obtained with this algorithm are shown in figure 1, where it may be observed that the clusters are compact and well defined. In order to visually validate the quality of the clusters produced by the density-based algorithm with k-means components, we use the cluster information to navigate through the 3-D body scans using the Cleopatra system and retrieve the body scans of the Archetypes or Centroids, as identified from the anthropometric data. Tables 1 and 2 indicate some of the characteristics of the Centroids of the male and female populations, respectively. Figures 2 and 3 show the 3-D body scans of the human subjects that correspond to these measurements, highlighting the difference in body types of the clusters. Thus, by inspecting the mean values in tables 1 and 2, the cluster distribution, and through the analysis of the results using Cleopatra, we observe that the anthropometric clusters discriminate between the different body sizes.

Table 1. Body measurements of male Centroids. The table shows the means (in cm), the standard deviation in parenthesis, and the number of subjects on each cluster.

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

Table 2. Body measurements of female Centroids. The table shows the means (in cm), the standard deviation in parenthesis, and the number of subjects on each cluster.

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

By inspecting the body measurements of the male Centroids shown in table 1, it may be observed that the *XX-Large* subjects are generally 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. 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.

Fig. 2. Cluster Centroids for the male population. (a) Small, (b) Medium, (c) Large, (d) X-Large and (e) XX-Large.

Fig. 3. Cluster Centroids for the female population. (a) X-Small, (b) Small, (c) Medium, (d) Large, (e) X-Large and (f) XX-Large.

For the female population is noticeable 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 the *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.

3.2 Clustering of 3-D Shapes

Following the same approach as with the anthropometric data, we subsequently performed cluster analysis on the 3-D data, which allows for the analysis of individuals based on their 3-D shape. There are many motivations for using the 3-D shapes directly. For products for which the shape is of critical importance like protective equipments, considering primarily the 3-D scans may yield superior results. In this case, one is most interested in determining which 3-D *shape features* are potentially the most relevant. We proceeded as follows. The 3-D data was separated into two sets based on the gender of the subject. The resulting sets consist of 542 males and 663 females and we applied the density-based algorithm with k-means components to the data. The number of subjects per cluster is given in table 3. We verify the cluster membership through querying the 3-D scans using the Cleopatra system, in order to determine whether our results correspond to the reality. The 3-D body scans of the Centroids of the male and female populations are shown in figures 4 and 5, 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.

Table 3. Number of subjects per cluster.

	Males	Females
X-Small	-	132 (20%)
Small	135 (25%)	63 (10%)
Medium	79 (15%)	103 (16%)
Large	166 (31%)	149 (22%)
X-Large	106 (20%)	147 (22%)
XX-Large	56 (10%)	69 (10%)

By inspecting the number of subjects per cluster in tables 1 to 3, we notice the number of subjects is not the same when the clustering is performed using the anthropometric measurements and the 3-D data. In our attempt to find the reasons for this situation, we analyze the anthropometric measurements and 3-D scans information of a set of representative individuals. Our results indicate that the clustering of 3-D body scans, as may be expected, group together subjects with similar body shape. On the other hand, when the clustering is performed on the anthropometric data, the algorithm finds the clusters containing those subjects with similar height, weight, bust circumference, and so on. A possible “change” in size, by a subject, is thus 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 and are therefore complementary to one another.

Fig. 4. Cluster Centroids from 3-D data for the male population. (a) Small, (b) Medium, (c) Large, (d) X-Large and (e) XX-Large.

Fig. 5. Cluster Centroids from 3-D data for the female population. (a) X-Small, (b) Small, (c) Medium, (d) Large, (e) X-Large and (f) XX-Large.

In the clothing industry, the design and manufacture of garments is based on body measurements. Therefore, the clustering obtained from the anthropometric measurements is most suitable for the direct application to tailoring and garments design. The clustering of 3-D shapes is, on the other hand, more useful when designing e.g. masks for protection against hazardous materials or custom made products such as artificial legs or helmets, where the goal is to produce a good fit for different shapes. Here, the anthropometric measurements are not enough to constraint the shape of an individual and consequently cannot replace a description based on the 3-D shape. Consequently, it is important, from the start, to determine if the particular application is based on measurements or shape.

4 Reducing the Number of Measurements

Recall from Section 2 that utility-based data mining accounts for the economic aspects that impact the mining process, and aims to maximize the utility of the process. In the previous section, we used in our analysis the total number of body measurements. We now aim to reduce the costs and time of the mining process by utilizing interestingness measures to reduce the number of body measurements and the size of the 3-D index. Our reasoning is as follows. The CAESARTM database contains a large number of attributes, making difficult for domain experts to interpret. Furthermore, even though it is the most comprehensive study of its kind to date, our

aim is to develop a scalable solution for future studies which may involve a very large number of participants. Reducing the number of measurements increases the efficiency of the learning process, enhances comprehensibility of the learned results and improves the learning performance (predictive accuracy) [14]. Moreover, the reduced set of body measurements may help identify the measures that require special attention when designing garments, thus potentially reducing the cost and complexity of this process. To this end, we perform two kinds of dimension reduction techniques, namely *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 [14].

Recall, from section 2, that interestingness measures are used in feature selection to remove the attributes with little or no predictive information. In our case study, this means that we use interestingness measures to identify the subset of the body measurements, or the 3-D descriptor, which is of most importance when describing an Archetype. For feature selection we thus apply *Information Gain*, *Gain Ratio*, *Chi Squared*, the *Consistency subset evaluator* and the *CFS subset evaluator*. These are measures that have been widely used in the context of feature selection and have been found to produce good results [18]. For feature extraction we use *Principal Component Analysis* (PCA), a well-known feature extraction method.

In order to verify our results, we consider the full set of anthropometric measurements (forty-four for the males, and forty-five for the females since under bust circumference was only recorded for the females) and the subsets produced by the feature selection and feature extraction. To this end, we constructed a number of classifiers, where the clusters we discovered during the characterization phase acted as class labels. For our experimentation, we consider three different classifiers, namely *RIPPER*, *C4.5* and *PART*. In the experiments we used ten-fold cross validation to test the accuracy of the classification models. The results of applying PCA and feature selection on the anthropometric data are shown in tables 4 and 5 for males and females, respectively. Shown are the predictive accuracy and, in parenthesis, the number of attributes in the subset.

Table 4. Results of the attribute reduction for the male population.

	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)

From table 4 it may be seen that for the males, PCA and all feature selection methods produce good results. That is, the subsets contain a small number of attributes (except for the subsets obtained using the CFS evaluator) and improve the accuracy we obtained using the full (original) set of attributes. We observe that, in general; the highest accuracy 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.

Table 5. Results of the attribute reduction for the female population.

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

For the females (table 5), the best results are obtained using Information 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 Information 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 both males and females are presented in table 6. For the 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 male population, these measurements should be considered carefully to produce garments that fit this population properly. For the 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. Therefore, when tailoring clothes for the Dutch females, the bust circumference requires special attention in order to design garments that fit the population better.

Table 6. Reduced set of anthropometric body measurements

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

We also apply the abovementioned approach to the 3-D shapes. The goal here is to determine which part of the radial-angular distribution of the surface elements is the most relevant, both in terms of design and in terms of time to search. It follows that, in very large databases; the smaller the index, the faster the search. Tables 7 and 8 show the results for the 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. Shown are the predictive accuracy and, in parenthesis, the number of attributes in the subset.

Table 7. Results of the attribute reduction of 3-D shapes for the male population.

	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)

Table 8. Results of the attribute reduction of 3-D shapes for the female population.

	Original	PCA	Info Gain	Gain Ratio	Chi Squared	Consistency Subset	CFS Subset
PART	78.6%	75.5% (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)

Table 7 shows that for the male population, Information Gain and Chi Squared achieve the best accuracy. However, the number of attributes in the subsets produced by these measures is much higher 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 is just slightly lower than the one obtained using the subsets generated by Information 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 (table 8), the highest accuracy is achieved by the subsets produced by the CFS subset evaluator. However, the numbers of attributes in these subsets are big. We also 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. Recall also that PCA provides linear combinations of the various elements of the descriptor, which might be difficult both to interpret and implement in the production line. Consequently, CFS would be preferred in most industrial applications. The subset of elements obtained with CFS provides us not only with a more compact index which higher descriptive power. In addition, it also gives us a set of distances and directions that are the most relevant in general in a 3-D shape; i.e. the distances and directions on which a design may be based. This information may subsequently be used in the scanning process, in order to generate an “intelligent” scan in which the most relevant regions of the body are scanned with higher resolution.

By inspecting the results of applying PCA and feature selection techniques on the anthropometric and 3-D data, it may be observed that in both anthropometric and 3-D data the accuracy is improved using a smaller set of attributes. When comparing 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 approximately 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 only around 20% of the original attributes.

5 Conclusions

One of the greatest challenges for the apparel industry is to produce garments that fit the customers properly, are aesthetically pleasing and comfortable. In order to produce garments that fit us well, better characterizations of our populations are needed. Furthermore, the different sizes must correspond to real body shapes, in the sense that one or more archetypes should represent the individuals belonging to the same size accurately. 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 convex or quasi convex 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 convex, 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 largest 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 tailoring and increases the complexity in the manufacturing.

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, when using the Cleopatra system, indicates that the cluster membership corresponds to the reality, in the sense that the bodies correspond to our expectations of the cluster membership. Also, our results indicate that the number of body measurements may be significantly reduced by applying interestingness measure-based feature selection and feature extraction. Moreover, these new sets of reduced body measurements improve 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 those body measurements that require special attention, when tailoring clothes for a specific population and gender.

What then, are the general conclusions and lessons to be learned from this data mining effort? Our experience shows that techniques, such as PCA, that combine attributes are difficult to interpret or, at least, to apply in a real world situation, such as when tailoring clothes. In our case study, we obtained better accuracy through dimensional reduction while improving the understandability of the model. It follows

that using less attributes is more efficient from a “production” perspective, both from a monetary and procedural point of view. Furthermore, when considering both descriptive (relational) and multimedia data, the relationship is not necessarily intuitive. That is, it not clear if one should replaced the other and which one are more suitable for use; rather, they seem complimentary.

References

1. Schofield, N. A. and LaBat, K. L. Exploring the relationships of grading, sizing and anthropometric data. *Clothing and Textiles Research Journal*, 23(1):13–27. (2005).
2. DesMarteau, K. "CAD: Let the fit revolution begin." *Bobbin*, 42:42–56. (2000).
3. Ashdown, S., Loker, S., and Rucker M. Improved Apparel Sizing: Fit and Anthropometric 3-D Scan Data. *Annual Report NTC Project: S04-CR01-07. National Textile Center*. (2007).
4. Robinette, K. M., Blackwell, S., Daanen, H., Fleming, S., Boehmer, M., Brill, T., Hoeferlin, D., and Burnsides, D. 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*. (2002).
5. Size USA. The US National Size Survey. Resource available at <http://www.sizeusa.com/>.
6. Size UK. UK National Sizing Survey. Resource available at <http://www.size.org/>.
7. Viktor, H. L., Paquet, E., and Guo, H. Measuring to fit: Virtual tailoring through cluster analysis and classification. In *PKDD 2006: Knowledge Discovery in Databases*, pages 395–406. (2006).
8. Veitch, D., Veitch, L., and Henneberg, M. Sizing for the Clothing Industry Using Principal Component Analysis - An Australian Example. (2007).
9. Hsu, C.-H., Lin, H.-F., and Wang, M.-J. Developing female size charts for facilitating garment production by using data mining. *Journal of Chinese Institute of Industrial Engineers*, 24(3):245–251. (2007).
10. Zadrozny, B., Weiss, G., and Saar-Tsechansky, M. UDBM 2006: Utility-Based Data Mining 2006 workshop report. *SIGKDD Explor. Newsl.*, 8(2):98–101. (2006).
11. Weiss, G., Saar-Tsechansky, M., and Zadrozny, B. Report on UDBM-05: Workshop on Utility-Based Data Mining. *SIGKDD Explor. Newsl.*, 7(2):145–147. (2005).
12. Geng, L. and Hamilton, H. J. Interestingness measures for data mining: A survey. *ACM Comput. Surv.*, 38(3):9. (2006).
13. McGarry, K. A survey of interestingness measures for knowledge discovery. *Knowl. Eng. Rev.*, 20(1):39–61. (2005).
14. Han, J. and Kamber, M. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, second edition. (2006).
15. Kim, Y., Street, W. N., and Menczer, F. Feature selection in data mining. *Data mining: opportunities and challenges*, pages 80–105. (2003).
16. Paquet, E., Robinette, K. M., and Rioux, M. Management of three-dimensional and anthropometric databases: Alexandria and Cleopatra. *Journal of Electronic Imaging*, 9:421–431. (2000).
17. Witten, I. H. and Frank, E. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, second edition. (2005).
18. Cunningham, P. Dimension Reduction. *Technical Report UCD-CSI-2007-7, University College Dublin*, pages 1–24. (2007).