Clustering for Decision Support in the Fashion Industry: A Case Study

Ana Monte, Carlos Soares, Pedro Brito and Michel Byvoet

Abstract The scope of this work is the segmentation of the orders of Bivolino, a Belgian company that sells custom tailored shirts. The segmentation is done based on clustering, following a Data Mining approach. We use the K-Medoids clustering method because it is less sensitive to outliers than other methods and it can handle nominal variables, which are the most common in the data used in this work. We interpret the results from both the design and marketing perspectives. The results of this analysis contain useful knowledge for the company regarding its business. This knowledge, as well as the continued usage of clustering to support both the design and marketing processes, is expected to allow Bivolino to make important business decisions and, thus, obtain competitive advantage over its competitors.

1 Introduction

The fashion industry is increasingly characterized by short life cycles, high volatility, low predictability, and high impulse purchasing [1]. In order to have manufacturing systems that support increased competitiveness in a global market, companies need complete and up-to-date knowledge about all the parts of their

A. Monte · P. Brito
Faculdade de Economia, Universidade do Porto, Porto, Portugal

C. Soares (✉)
Faculdade de Engenharia, Universidade do Porto, Porto, Portugal
e-mail: csoares@fe.up.pt

A. Monte · C. Soares · P. Brito
INESC TEC, Porto, Portugal

M. Byvoet
Bivolino, Dipenbeek, Belgium

A. Azevedo (ed.), *Advances in Sustainable and Competitive Manufacturing Systems*, Lecture Notes in Mechanical Engineering, DOI: 10.1007/978-3-319-00557-7_82,

business (e.g., product development, marketing, production) and they need to integrate this knowledge into those systems. This work addresses one part of that problem. We use Data Mining (DM) approaches to extract knowledge from historical data with the goal of supporting the fashion industry in its production/design and marketing decisions.

The data that companies collect about their customers is one of its greatest assets [2]. Companies are increasingly accumulating huge amounts of customer data in large databases [3] and know that, within this vast amount of data, there are all sorts of valuable information that could make a significant difference to the way in which any company run their business and interact with their current and prospective customers. If available, this information can help them gain competitive edge on their competitors [2]. Given the additional fact that companies have to be able to react rapidly to the changing market demands both locally and globally [2], it is urgent that they manage efficiently the information about their customers. This is true in fashion as in most other industries.

One approach that is increasingly being used for that purpose is data mining. DM techniques can be used to extract unknown and potentially useful knowledge about customer characteristics and their purchase patterns [3]. This knowledge can, then, be used predict future trends and behaviors, allowing businesses to make knowledge-driven decisions that will affect the company, both short term and long term [2]. DM is also being used in e-commerce to study and identify the performance limitations, increase sales, and address issues raised by political and physical boundaries [2]. The identification of such patterns in data is the first step to gaining useful marketing insights and making critical marketing decisions [3]. In today's environment of complex and ever changing customer preferences, marketing decisions that are informed by knowledge about individual customers becomes critical [3]. Customers have such a varied tastes and preferences that it is not possible to group them into large and homogeneous populations to develop marketing strategies. In fact, each customer wants to be served according to his individual and unique needs [3]. Thus, the move from mass marketing to one-to-one relationship marketing requires decision-makers to come up with specific strategies for each individual customer based on his profile [3]. This is particularly important in highly customized industries. In extreme cases, such as Bivolino, each product is different from all the others.

In short, DM is a very powerful tool that should be used for increasing customer satisfaction by providing good quality, safe, and useful products at reasonable prices as well as for making the business more competitive and profitable [2]. One of the most important DM tasks is clustering [4, 9]. Clustering algorithms find subgroups of observations in a population that are similar among themselves but very different from the observations in other subgroups. Clustering is often used in segmentation [6]. In this paper, we show how clustering can be useful for segmentation in the textile industry, in particular for highly customized garments. Our goal is to use the clusters to support both the design and the marketing processes. The approach is tested on data provided by Bivolino, a manufacturer of custom

tailored shirts. To the best of our knowledge, there are no publications on the use of this approach in highly customized industries.

We start by motivating this study in Sect. 2. In Sect. 3 we present a short introduction to clustering. In Sect. 4 we present and discuss the results. The paper finishes with some conclusions and some ideas for future work (Sect. 5).

2 Clustering

According to Jain [4], clustering can be defined, in operational terms, as follows: "Given a *representation* of n objects, find k groups based on a measure of similarity such that the similarities between objects in the same group are high while the similarities between objects in different groups are low."

2.1 K-Medoids Algorithm

According to Velmurugan and Santhanam [5], the basic strategy of K-Medoids clustering algorithms is to find k clusters in n objects by (1) arbitrarily finding a representative object (the medoid) for each cluster; (2) associate each remaining object with the medoid to which it is the most similar; (3) update the medoids by choosing the most representative object in each of the k clusters; and (4) repeat steps 2 and 3 until convergence or a stopping criterion is met. Unlike the more common method, K-Means, the K-Medoids method uses representative objects as reference points. The algorithm takes the input parameter k, the number of clusters, to be partitioned among a set of n objects.

Velmurugan and Santhanam [5] present a typical K-Medoids algorithm for partitioning based on medoids, i.e., examples from the data that represent the corresponding clusters. Due to lack of space, we omit the details of the algorithm, which can easily be found in the literature (e.g., [5]). A very informal summary is:

Input: $K =$ number of clusters and $D =$ dataset containing n objects.
Output: A set of k clusters that minimizes the sum of the dissimilarities of all the objects to their nearest medoid.
Method: Arbitrarily choose k objects in D as the initial representative objects.
Repeat: Assign each remaining object to the cluster with the nearest medoid; randomly select a non medoid object O_{random}; compute the total points S of swap point O_j with O_{random}; if $S < 0$ then swap O_j with O_{random} to form the new set of k medoid until no change occurs.

3 Case Study

This work was motivated by the need of segmenting Bivolino shirts orders, based on data from 2011. The goal was to obtain knowledge to support design and marketing business decisions. Bivolino[1] is a company that produces and sells customized shirts on the web, both for men and women.

Our goal is to demonstrate the applicability of data mining methodology, tasks, techniques, and tools to support both design and marketing processes in the fashion industry. The data mining methodology adopted was CRISP-DM.[2] The data mining task that is suitable for the problem at hand is clustering. Clustering partitions data based on a certain similarity criteria or distance measure. The distance measure used was the Mixed Euclidean Distance, given the fact that we have numerical and categorical variables. This measure uses the classical Euclidean distance for numerical variables and a binary distance for symbolic variables (i.e., distance is 0 if both values are the same and 1, otherwise). The clustering algorithm applied to the data was K-Medoids because it works well with large amounts of data, is less sensitive to outliers, and can handle categorical data. The data mining software used to segment the shirt orders was Rapid Miner.

The problem is the identification of the fashion profiles in Bivolino shirts orders to support the company in terms of design and marketing decisions. The aim is to extract useful information and obtain important knowledge about the business of Bivolino using data mining tools and techniques.

4 Results

We present two different analyses. The first one is focused on the perspective of production/design of shirts while the other takes a marketing perspective. The difference between them is in the inputs, i.e., the sets of attributes used to characterize the transactions. Different variables naturally lead to different results. The analysis of the first results is focused on the shirts and orders attributes, namely in the typical values they assume in each cluster, as given by the respective medoid. Due to lack of space, we do not describe the results and the data in detail. We simply illustrate the results with some examples, namely by identifying the most salient features of the clusters.

[1] http://www.bivolino.com

[2] http://http://khabaza.codimension.net/index_files/crispdm.htm

4.1 Clustering for Design

In these experiments, the goal is to analyze the characteristics of the shirts that were purchased, searching for knowledge that is potentially useful for shirt design and production. We try to identify the fashion trends on shirts based on customer choices of shirts and their characteristics.

Table 1 summarizes the results of the clustering performed on data concerning 10,775 shirts orders characterized by 29 attributes (4 numerical and 25 categorical), with the number of clusters $k = 6$. All the clusters obtained have medoids with at least one attribute that is clearly distinctive, i.e., that identifies each cluster in a unique way. The medoids are defined by the most "typical" value of each attributes in the corresponding cluster. For numerical attributes, the typical value is the mean while for the categorical ones it is the most common value, i.e., the mode.

Analyzing the results, we can group the clusters into two main groups. The first group is composed by clusters 1, 2, and 4 corresponding to work shirts and representing 65 % of total shirts orders. The second group is composed by clusters 3, 5, and 6 corresponding to fashion shirts and representing 35 % of total shirts orders. Analyzing these groups in more detail, we additionally observe the following: costumer choices are conditioned by a certain formal business dress code; cluster 2 represents the most common choices in terms of shirt attributes; and the binary[3] attributes (e.g., has pocket, has monogram) assume, in most of the cases, the value *no*. In terms of the second group, we conclude that costumer choices were conditioned by their physical attributes (e.g., mature men have preference for pockets on shirts unlike the young men). This relation is illustrated in Fig. 1.

These observations can be useful for the product designers because understanding what the preferences of the customers are and how they evolve will lead to the development of products which are better suited to their preferences.

4.2 Clustering for Marketing

In the second experiment, we focus on the customers rather than on the shirts, which is more suitable for marketing purposes. In marketing terminology, we perform a segmentation of the costumers. This perspective aims to identify and define in which segments the company should focus its efforts and marketing strategies.

First, we have to define the segmentation bases, i.e., the attributes that are used to characterize the observations. We have added to the attributes that characterize the shirt, several attributes that characterize the customers. In Table 2 we show the classical segmentation bases suggested by Kotler and Keller [6].

[3] Binary attributes in Rapid Miner's terminology are binomial attributes.

Table 1 Clustering results in a technical perspective (distinct medoids attributes values). The second line contains the number of observations and the corresponding proportion. The lines below that, contain differentiating features of the clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
1,282 12%	4,889 45%	872 8%	1,570 15%	500 5%	1,662 15%
Work_shirt	Work_shirt	Fashion_shirt	Work_shirt	Fashion_shirt	Fashion_shirt
Fabric_color: *Multicolor*	Fabric: *Greenwich*	Fabric_structure: *Twill*	Collar: *Italian Semi-spread*	Collar/Cuff white: *Yes*	Fabric: *Miro_3*
Fabric: *Sheffield*		Fabric: *Red & Bordeaux*	Fabric: *London_4*	Fabric: *Kiwi_9*	Fabric_structure: *Herringbone*

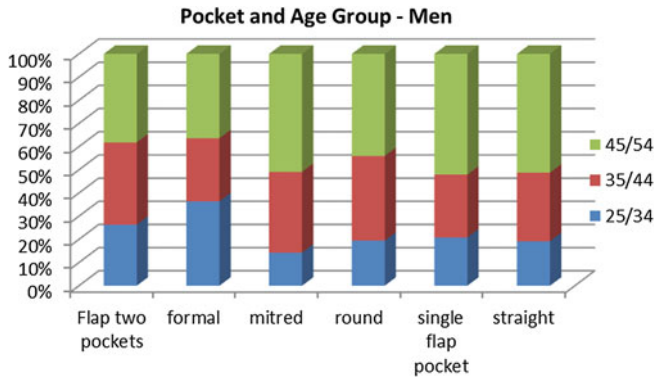

Fig. 1 Relation between age and shirt pocket

Table 2 Segmentation bases

Demographic (who they are)	Geographic (where they live)	Psychographic (how they behave)	Behavioristic (why they buy)
Gender: men (91%), women (9%) Age: [25-34], [35-44], [45-54] Country/Nationality: UK (United Kingdom), fr (France), de (Germany)	Nationality: UK (United Kingdom), fr (France), de (Germany) Postal code: any different	Lifestyle: activities—work; social events; or entertainment Interests—work; fashion	Price sensitivity: has voucher (yes; no)

So, the company can segment its customers according to the three demographic variables gender, age, and nationality. This category of segmentation variables is, according to Kotler and Keller [6], very popular in the way that these variables are often associated with consumer needs and wants and are easy to measure. If a company chooses to segment its customers by gender, it has to take into account that "men and women have different attitudes and behave differently, based partly on genetic makeup and partly on socialization" [6] and that "they have different expectations of fashion products" [7]. If a company segments its customers on the basis of the age, the shirts have to be "designed to meet the specific needs of certain age groups" because "customer wants and abilities change with the age" [6]. If a company decides to segment its customer market based on nationality, it has to pay attention to the "identity attributes because of social and cultural values that inform the self" [7].

The company can also divide its customers market into different countries and also into specific regions given the postal code. This kind of segmentation does not ensure that all customers in a location will make the same buying decision; however, it helps in identifying some general patterns [6].

A psychographic segmentation is based on variables that are inferred such as personality traits (consumerism, dogmatism, locus of control, cognitive style, and

religion), personal values, and lifestyle (activities: work, hobbies, social events, entertainment, etc.; interests: family, home, job, fashion, etc.; opinions: of oneself, social issues, economics, culture, etc.). As we did not have access to such kind of personal information, we could only make some inference about it. Therefore, we propose that the company segments its customers market according to their lifestyle based on their activities (e.g., work and social events or entertainment) and interests (e.g., work and fashion).

Behavioral variables are considered by marketers "the best starting point for constructing market segments" [6] and "are related to buying and consumption behavior" [8]. This category comprises variables such as occasions, benefits expectations, brand loyalty, price sensitivity, usage rate, end use, attitude, preferences, etc. Some of these variables can be directly measured and others have to be inferred. As for the psychographic variables, we did not have enough information. However, we identified the price sensitivity as a behavioristic variable which can be measured by the use or not of gift vouchers that offers discount on payment.

Table 3 presents the clustering results concerning 10,775 shirts orders characterized by 59 attributes this time (30 more that in the first experiment, where 27 are numerical and 32 categorical). As before, we set the number of clusters to $k = 6$, and the algorithm returned 6 clusters with at least one clearly distinctive attribute.

Table 3, like Table 1, represents the distinctive attribute values that identify each cluster in a unique way. In this case, we can group the clusters in 3 groups. They are the work shirts (clusters 1, 4, and 5), the party shirts (clusters 2 and 3), and the fashion shirts (cluster 6). Several interesting observations can be made from these results. For instance, they show that young men (age between 25 and 34] years old) prefer shirts of slim fit, while mature men (age between 45 and 54 years old) prefer shirts of comfort fit.

Another interesting segment is given by cluster 2, which represents mostly women. The characteristics of the shirts in this segment make it possible for the

Table 3 Clustering results in a marketing perspective (distinct medoids attributes values). The second line contains the number of observations and the corresponding proportion. The lines below that contain differentiating features of the clusters

Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5	Cluster 6
2,687 25%	5,116 47%	2,316 21%	80 1%	564 5%	12 1%
Work_shirt	Party_shirt	Party_shirt	Work_shirt	Work_shirt	Fashion_shirt
Collar_group: <36	Gender: women	Postal code: sg49aq	Postal code: co45bq	Postal code: cv313nd	Country: de (Germany)
	Has voucher: yes				
	Country: fr (France)				
	Affiliate: Bivolino				

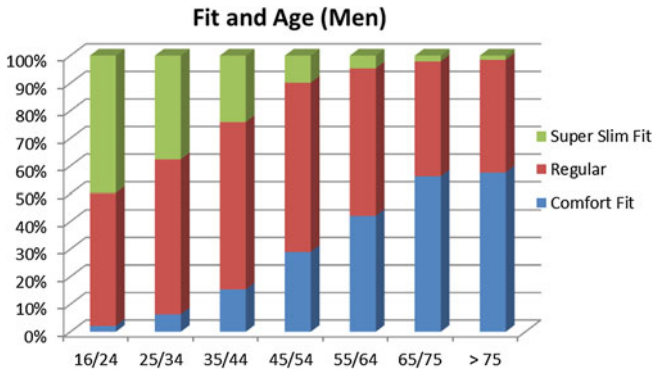

Fig. 2 Relation between shirt fit and customer age

company to segment its consumer market on the basis of the gender of the customers, as men and women have different expectations of fashion products.

We also observe that the company can segment its market based on the age of the customers, since the changes in body characteristics and shape have a great impact on fashion choices. One example is that in terms of shirt fit (level of shirt tightness to the body) choices, the young customers prefer the "super slim fit" and the older customers the "comfort fit", although the "regular fit" remains the most common choice. This relation is illustrated in Fig. 2. Other example is that the older customers are the ones that like to use pocket on shirts the most, especially of a certain type.

The company can also segment the market geographically by country or postal code, since customers from different places have different requirements for fashion and clothing products, and their choices are often influenced by social and cultural values. We observe that the more representative country in terms of Bivolino sales in 2011 was the United Kingdom (UK). This fact is illustrated in Fig. 3.

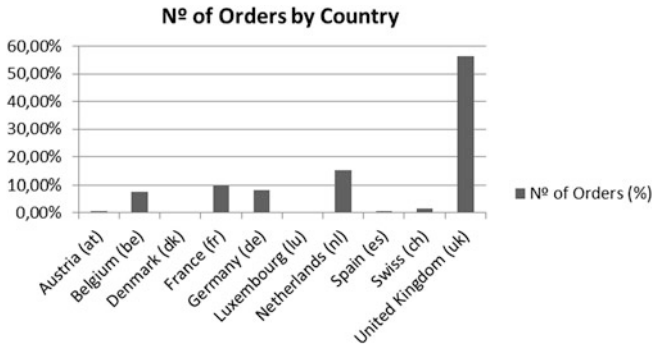

Fig. 3 Relation between shirt fit and customer age

Another alternative for segmenting the Bivolino costumers is on the basis of the purpose or intention of buying according to, for instance, the categories "work shirt", "fashion shirt", and "party shirt" (configurator or collection type). We can then infer, for example, that their buying is motivated by professional requirements (e.g., segments 1, 4, and 5), interest on fashion (e.g., segment 6), or by social events requirements (e.g., segments 2 and 3).

We have also identified another segmentation variable that is the price sensitivity. This could be measured by the usage of gift vouchers to get a discount on payment. We concluded that women are more price sensitive than men given that more than 82 % used a voucher for payment. Women are generally more receptive to promotions of this type and more open to experimenting new products than men. However, we do not know if this was related to a specific promotion, because Bivolino only introduced the female collection in 2011. On the other hand, if this is a gender-related pattern, then the company should study other ways of attract the female public and improve sales.

4.3 Limitations

During our study we faced several challenges. The first one is related to the limitations of the method adopted. The K-Medoids clustering algorithm, despite being less sensitive to outliers than K-Means because of the use of the median instead of the mean, still requires the a priori definition of the number of clusters. Deciding the optimal k number of clusters is well known to be a difficult task [9]. Common approaches to this problem consist of running the algorithm multiple times with different parameter values (i.e., k), and the best configuration obtained from all of the runs is used as the final clustering. This method is extremely time consuming. In our experiments, each clustering process took a significant amount of time (a few hours) on Rapid Miner due to the quantity of examples as well as to the number of attributes. Another difficulty concerns the evaluation of the result: how to assess whether the clustering obtained is good or not. All clustering algorithms will, when presented with data, produce clusters regardless of whether the data contain clusters or not. If the data does contain clusters, some clustering algorithms may obtain 'better' cluster than others [9]. However, in this study, this was not an important problem because, given the type of the attributes used in the study (binary and polynomial, i.e., non-numeric), the choice of algorithm was very limited. Most of the clustering algorithms available on Rapid Miner do not process categorical data (data separable into categories that are mutually exclusive, such as age groups). However, the algorithm selected proved to be suitable for the problem at hand.

In summary, the most important difficulties faced in the study were:

- Deciding the number of clusters k;
- High computational cost of the methods;

- Limitations imposed by the types of the attributes, which strongly constrain the selection of the algorithm;
- Subjective nature of the evaluation process;
- Insufficient data to characterize customers.

5 Conclusions

In this paper, we use clustering for segmentation in a textile manufacturer of highly customized garments. We investigated how the clusters obtained can be used to support both the design and the marketing processes.

The approach was tested on data provided by Bivolino, a manufacturer of custom tailored shirts, containing 10,775 examples which correspond to the number of shirt orders in 2011. The clustering was obtained using a K-Medoids algorithm with $k = 6$ clusters. Two sets of experiments were carried out. In the first, the transactions were characterized using 29 attributes that describe the orders, the shirts, and the customers, while in the second, 59 attributes were used, including attributes that characterize the customers in terms of demographic, geographic, psychographic, and behavioristic features.

In the first set of experiments, we focused on the characteristics of the shirts because the goal was to obtain knowledge (e.g., profiles and trends) that is useful for the design process. The clusters obtained represent attribute values in shirt orders that are specific to a subgroup of transactions, which can be used by the fashion designers to better perform their tasks.

In the second set of experiments, we included features that are typically used in marketing studies to describe customers. Thus, the clusters found by the K-Medoids algorithm enabled us to analyze the profile of Bivolino customers and their relation to the product (shirts). We concluded that their choices in terms of shirts attributes are greatly influenced and conditioned by numerous factors such as, their physical attributes, age, gender, nationality (country), and purpose of buying. The profiles found can be used by the company to adjust its product and marketing strategies to the different segments identified.

In summary, despite some challenges that were discussed, this study illustrates how DM tools and techniques, namely clustering, are indeed valuable instruments to better understand consumer tastes and preferences allowing companies to be more efficient and responsive to customer requests and gaining a competitive advantage, particularly in highly customized textile industries.

Taking into account the difficulties encountered during the study and the limitations they imposed on it, we find that future work is needed and should focus on the following issues, which are mainly related to the problems of the clustering process itself:

- Develop a heuristic solution to find the optimal number of clusters for any given dataset;

- Reduce the time complexity when dealing with large number of dimensions and large number of data items;
- Develop methods to support the systematic interpretation of the result of a clustering algorithm (that in many cases can be arbitrary itself);
- Transform the data to enable the use of different clustering algorithms, compare the different results, and decide which one is the best;
- Reduce the dependency of the effectiveness of the clustering method on the definition of "distance" (for distance-based clustering);
- Alternatively to the previous suggestion, it would be interesting to find a way of defining a distance measure, that is specific to the business problem and aligned with the evaluation criteria.

Acknowledgments The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007) agreement n° [260169] (Project CoReNet), from the ERDF—European Regional Development Fund through the COMPETE Programme (operational programme for competitiveness), and from QREN (National Strategic Reference Framework) as part of project PTXXI.

References

1. Lo W, Hong T, Jeng R (2008) A framework of E-SCM multi-agent systems in the fashion industry. Int J Prod Econ 114:594–614
2. Ahmed S (2004) Applications of data mining in retail business. In: International conference on information technology: coding and computing (ITCC'04), IEEE
3. Shaw M, Subramaniam C, Tan G, Welge M (2001) Knowledge management and data mining for marketing. Decis Support Syst 31:127–137
4. Jain AK (2009) Data clustering: 50 years beyond K-means. Pattern Recognit Lett 31:651
5. Velmurugan T, Santhanam T (2010) Computational complexity between K-means and K-medoids clustering algorithms for normal and uniform distributions of data points. J Comput Sci 6(3):363–368
6. Kotler P, Keller K (2000) Marketing management, 13th edn. Prentice Hall, New York
7. Rocha M, Hammond L, Hawkins D (2005) Age, gender and national factors in fashion consumption. J Fashion Mark Manage 9(4):390–390
8. Wedel M, Kamakura W (2000) Market segmentation: conceptual and methodological foundations. Kluwer Academic Publishers, Dordrecht
9. Jain AK, Murty MN, Flynn PJ (2000) Data clustering: a review. ACM Comput Surv 31(3):264