KU LEUVEN
FACULTEIT ECONOMIE EN
BEDRIJFSWETENSCHAPPEN

Business-Driven Data Mining: New Algorithms and Applications

Proefschrift voorgedragen tot
het behalen van de graad van
Doctor in de Toegepaste
Economische Wetenschappen

door

Alex Seret

Nummer 474 2015

Committee

Chairperson	Prof. dr. Inneke Van Nieuwenhuyse	KU Leuven
Supervisor	Prof. dr. Bart Baesens	KU Leuven
	Prof. dr. Jan Vanthienen	KU Leuven
	Prof. dr. Arnulfo Azcarraga	De La Salle University
	Prof. dr. Richard Weber	Universidad de Chile
	Prof. dr. Sebastián Maldonado	Universidad de los Andes

Acknowledgments

I would like to express my sincere gratitude to my Promotor, Bart Baesens. As my advisor, you gave me all the tools I needed and even more. You always gave me the right balance of freedom and support necessary to achieve my goals and, most of all, you gave me your trust. I will always remember and value this trust and hope to be able to pass this forward to other junior researchers that may cross my path. I believe you would consider this to be a nice pay back, isn't it boss? Consider me as an asset on which you can count in the future, whatever the topic, whatever the place.

I also want to express special thanks to my committee members. To Jan Vanthienen and Arnulfo Azcarraga who followed and guided my journey from the first moments, I sincerely thank you and hope to be able to collaborate with you in the future. Your wisdom and advice were of great help during my PhD and I hope to make good use of it in what follows. To Richard Weber and Sebastian Maldonado who joined the committee later on the road, I want to thank you for two reasons. First, because you helped me to finish properly this journey by providing me the feedback and inspiration I needed. Second, and not least, because you prepared me a new journey starting soon in Chile. Our conversations were worth the travel to Santiago and I hope that the distance will never be a reason to stop it.

A PhD is often a solo trip, but not always. I wish the best to all my friends and colleagues and hope to meet you in a near future. Be sure that you can always count on me, especially if you come with an Orval.

As usual, I was able to count on my family during this journey. I particularly want to thank my mother, my father, my brother and my grandmother for their daily support. You are a big part of this.

Finally, I want to thank my lovely wife, Ségolène, for her presence during this adventure. Thank you for everything you did and for what is coming.

To my wife and son,
I love you.

Alex Seret - Leuven, February 2015.

Contents

Chapter 1

Introduction

1.1 General context

The explosive growth of the amount of available data and the reliance on data mining techniques have led to the creation of a myriad of new business models and opportunities. The field of direct marketing is not an exception and explores ways of getting competitive advantages by supporting research on the development of innovative and value-adding techniques.

Although data mining techniques have been successfully applied in different companies (often big companies), it is still difficult for smaller organizations to monetize or at least explore the data they collected. From this perspective, a need for comprehensible stepwise business-oriented exploration techniques can be identified. In the remainder of this document, different approaches applied on cases with such smaller companies or departments are presented and discussed. The focus of this thesis is on providing the business with business-oriented, comprehensible, step-wise and visual exploration techniques and methodologies. Since the step-wise aspect of the approaches presented in this work is of a crucial importance for business acceptance reasons, some of the techniques are reused and re-discussed in different chapters.

1.2 Research context

The research reported in this document is a subset of the output generated during a PhD sponsored by Ticketmatic, a leading ticketing company based in Belgium and active in central Europa. Providing their customers with ticketing solutions in a SAAS fashion, Ticketmatic is also capturing and storing daily transactional data about sold tickets for each of his customers. These customers being small to large event organizers dealing with thousands of event attenders per year, the databases of Ticketmatic are rapidly growing. Being an innovative company, Ticketmatic decided to investigate the opportunities associated with the use of data mining techniques in such a context. The long term commercial goal of Ticketmatic would then be to provide the event organizers with data-driven tools enabling the monetization of their databases.

Since a large part of the data collected by Ticketmatic could be used to serve marketing purposes, different marketing departments of major customers of Ticketmatic were approached. After discussions with the different marketing managers, a clear need for exploration techniques and data-driven segmentation approaches was identified. Facing departments with no experience with data mining, general research questions were stated and answered through different projects.

1.2.1 Outline and contribution

This section is an outline of this manuscript highlighting the main contributions and linking the different chapters. Each chapter is based on a specific project conducted with partners from the industry. These projects have been first reported and published or submitted to international peer reviewed journals before being merged for the purpose of this manuscript.

In the second chapter, entitled "A new SOM-based method for profile generation: theory and an application in direct marketing", a SOM-based profile generator is presented, consisting of a generic method leading to value-adding and business-oriented profiles for targeting in-

dividuals with predefined characteristics. The profile generator is then applied and the performance of it is assessed (see Chapter 2). After conducting this project, new insights were obtained and it was decided that advanced visual clustering approaches as SOM could be really value-adding for the marketers involved in the project. Although the predictive aspects triggered the interest of the different partners and made it possible to understand the impact of the volume of available data on the analysis, only few customers of Ticketmatic were mature enough to go that far in the proposed methodology. Since the design and application of accessible and comprehensible techniques were key aspects for the partners involved, one decided to focus on the descriptive aspects and concentrated on the segmentation, hence clustering, steps.

This chapter is based on a paper published in the European Journal of Operational Research:

- Seret, A., Verbraken, T., Versailles, S., Baesens, B. (2012). A new SOM-based method for profile generation: theory and an application in direct marketing. European Journal of Operational Research, 220 (1), 199-209.

Following this project in a marketing context, a similar analysis in a microfinance context has been performed and published in World Development:

- Louis, P., Seret, A., Baesens, B. (2013). Financial efficiency and social impact of microfinance institutions using self-organizing maps. World Development, 46, 197-210.

Because of the unsupervised character of traditional clustering approaches, some of the obtained segmentations were not satisfying the business partners. While trying to understand the reasons, it appeared that an a priori knowledge about the importance of subsets of the variables existed. This a priori knowledge was dependent on the analyst or analysis and his goals. Incorporating this a priori knowledge into a clustering algorithm in order to guide it and lead to a better perceived clustering motivates the third chapter, entitled “A new

knowledge-based constrained clustering approach: theory and application in direct marketing" (See Chapter 3). A formalization of the fact that an intuitive a priori prioritization of the variables might exist, is presented in this chapter and applied in a direct marketing context. By providing the analyst with a new approach offering different clustering perspectives, this chapter proposes a straightforward way to apply constrained clustering with soft attribute-level constraints based on feature order preferences.

This chapter is based on a paper published in Applied Soft Computing:

- Seret, A., Verbraken, T., Baesens, B. (2014). A new knowledge-based constrained clustering approach: theory and application in direct marketing. Applied Soft Computing, 24, 316-327.

The need to incorporate a priori business knowledge was shared by different marketers in the concert industry. Moreover, being in contact with the customer intelligence department of one of the main international banks, a segmentation project guided by business knowledge was initiated, allowing us to assess the applicability of our approach in another industry. The results of this project are reported in the fourth chapter, entitled "Business knowledge based segmentation of online banking customers" (see Chapter 4). Offering an increasing amount of internet services, better understanding the online customer in order to facilitate the appropriate value proposition is a major concern for nowadays banking companies. Data mining tools have proven their efficiency in addressing this challenge by providing unsupervised quantitative techniques to identify those customer segments with similar characteristics. To contrast with these traditional approaches, this chapter focusses on segmenting an online banking customer base in a meaningful way for the business by enhancing an unsupervised quantitative technique approach with business knowledge.

This chapter is based on a paper submitted to Intelligent Data Analysis:

- Seret, A., Bejinaru, A., Baesens, B.. Business knowledge based

segmentation of online banking customers. Intelligent Data Analysis. Under review.

As a logical second step following the understanding of the customers at a specific moment, we then investigated the impact of the time dimension in such a segmentation analysis. This interest was triggered by the fact that customers may, soon or later, change their behaviors and hence move from one segment to another. Understanding these movements in a high dimensional space while answering advanced business questions building on the knowledge based clustering algorithm previously studied became thus the next research step. A new project based on data from the main customer of Ticketmatic started and resulted in a chapter entitled "A dynamic understanding of customer behavior processes based on clustering and sequence mining" (see Chapter 5). In this chapter, a new approach towards enabling the exploratory understanding of the dynamics inherent in the capture of customers' data at different points in time is outlined. The proposed methodology combines the previously designed clustering techniques with a tuned sequence mining method to discover prominent customer behavior trajectories in data bases, which – when combined – represent the "behavior process" as it is followed by particular groups of customers. The framework is applied to a real-life case of an event organizer; it is shown how behavior trajectories can help to explain consumer decisions and to improve business processes that are influenced by customer actions.

This chapter is based on a paper published in Expert Systems with Applications:

- Seret, A., vanden Broucke, S., Baesens, B., Vanthienen, J. (2014). A dynamic understanding of customer behavior processes based on clustering and sequence mining. Expert Systems with Applications, 41(10), 4648-4657.

Finally, the different projects of this work leading to new insights for the departments involved, a need for updating and enriching mechanisms of these insights was identified. In this context, the sixth chapter of this manuscript, entitled "Identifying next relevant variables for

segmentation by using feature selection approaches", investigates the problem of updating and improving an existing clustering model by adding relevant new variables (see Chapter 6). A relevant variable is here defined as a feature which is highly correlated with the current structure of the data, since our main goal is to improve the model by adding new information to the current segmentation, but without modifying it significantly. For this purpose, a general framework is proposed, and subsequently applied in a real business context involving an event organizer facing this problem. Based on extensive experiments based on real data, the performance of the proposed approach is compared to existing methods using different evaluation metrics, leading to the conclusion that the proposed technique is performing better for this specific problem.

This chapter is based on a paper accepted in Expert Systems with Applications:

- Seret, A., Maldonado, S., Baesens, B. (2015). Identifying next relevant variables for segmentation by using feature selection approaches. Expert Systems with Applications. Accepted.

In the last chapters, the impact on the business is highlighted (see Chapter 7) and general conclusions are drawn while identifying and discussing some tracks for future research (see Chapter 8).

1.2.2 Research methodology

The research methodology used for the elaboration of this manuscript consists of five projects or case studies, each of them answering or approaching specific questions. The general approach used to build each of the chapters was based on a standard set of steps.

- The first step of each project consisted in the formalization of the business context and problem. This step involved the participation of different business partners mainly working in marketing contexts.

- In a second step, data was collected business-side and preprocessed and transformed in interaction with the business involved.
- As a third step, a literature study of the existing methods and approaches related to the formalized problem was performed, allowing to identify gaps or opportunities to improve current state-of-art.
- In a fourth step, new methodologies, algorithms and performance measures were proposed in order to solve the formalized problem.
- In a fifth step, software pieces were developed in order to implement the proposed methodologies, mainly using *Matlab* and *R*.
- Finally, an application in a real business context was typically reported while discussing the impact of different parameters used during the experiments.

As a crucial element guiding our methodology, we always tried to validate our decisions and approaches by continuously looking for peer review, targeting top international journals. This process clearly improved the quality of the different chapters while validating our general approach.

Chapter 2

A new SOM-based method for profile generation: theory and an application in direct marketing

2.1 Introduction

The explosive growth of the amount of available data and the reliance on data mining techniques have led to the creation of a myriad of new business models and opportunities. The field of direct marketing is not an exception and explores ways of getting competitive advantages by supporting research on the development of innovative and value-adding techniques. Self-organizing maps (SOM) are one of these techniques and have been applied for as many purposes as domains. Giving a powerful encapsulated facility for the analysis of complex databases by reducing the curse of dimensionality, this technique provides the direct marketer with the required tools to take accurate, quick and value-adding decisions. The inexhaustible source of applications has been widely discussed in the literature and has been combined with existing techniques in order to verify the statement that *The whole is greater than the sum of its parts*. Combined with clustering techniques,

it is then possible to widen the scope of the analysis and obtain a better insight about the studied data. The extraction of so-called salient dimensions permits the direct marketer to identify and segment its prospects.

In this chapter, the authors propose a generic method aiming at generating profiles based on the SOM technology and the extraction of salient dimensions, enabling the direct marketer to formalize his feelings and insights on a dataset while generating value-adding and business-oriented profiles which target individuals with predefined characteristics. The developed generic method is applied to a real life case study, conducted in cooperation with Ticketmatic, a Belgian provider of ticketing solutions. Data from the concert industry is analyzed, and the performance of the proposed method is discussed and challenged in order to evaluate its potential while identifying further interesting research topics.

This chapter is structured as follows. Section 2.2 provides the necessary background about customer segmentation and direct marketing, introducing the concepts of segmentation bases, customer profitability, the RFM framework and three techniques of segmentation, namely self-organizing maps, k-means algorithm and salient dimensions extraction. In Section 2.3 the SOM-based profile generator is presented and completed with ad hoc definitions of performance measures. Section 2.4 presents an application of the proposed method and an analysis of the performance of the generated profiles. A more extensive discussion of the impact of different parameters on the performance, the managerial aspects and different topics for further research is to be found in Section 2.5.

2.2 Customer segmentation & Profile generation

In the field of direct marketing, many techniques have been used to identify the most profitable customers, or the customers which are most likely to respond to a specific campaign. However, such analyses

only enable the direct marketer to predict the behavior of the already known customers. A more interesting goal for customer segmentation is the identification of customer profiles, so that one can predict the behavior of unknown customers. With such customer profiles, interesting applications in direct marketing emerge, such as targeting specific geographic zones or social groups (e.g. readers of a certain journal, listeners of a certain radio channel, etc.). Whether or not the main goal of the segmentation is to build customer profiles, two major characteristics have to be defined: the segmentation bases and the technique used to identify segments. Given that this chapter proposes a new segmentation technique based on the generation of profiles, the two following sections will focus on the techniques used while referring the interested reader to [1] for additional information on the segmentation bases.

2.2.1 Techniques used for the segmentation task

Data mining techniques are often used for the difficult task of segmentation in order to provide the domain experts with key information on the structure of the data they are dealing with. Different techniques are discussed in the literature and an important distinction has to be made between supervised and unsupervised learning. Supervised learning problems involve labeled data and aim at finding models predicting the labels of new unlabeled training patterns. Different supervised techniques such as optimization models ([2]), Bayesian neural networks ([3] and [4]) and decision trees ([5]) have been used and discussed in the literature, offering different approaches for the task of segmentation by creating rules which capture the information hidden in the data. However, unsupervised learning techniques such as clustering have encountered more success because only unlabeled data are necessary which ease their collection and allow exploratory analysis. Clustering techniques are still widely applied, discussed and improved in the literature (good examples are [6] , [7], [8], [9], [10], [11]) and find applications in all domains where data grouping and summarization using prototypes or profiles make sense. The evolution

of these techniques, outlined in [12], offers new ways of dealing with existing problems such as the segmentation of a customer base. Moreover, new approaches combining existing techniques (e.g. [13]) reveal synergy possibilities that have to be exploited. This chapter proposes a method consisting of a sequence of existing unsupervised techniques and a new approach in order to go further in the analysis.

2.2.2 Techniques used in the proposed method

This section will focus on the three major techniques used in this study, namely self-organizing maps (SOM), k-means algorithm and salient dimensions extraction in order to provide the reader with the necessary background.

2.2.2.1 Self-organizing maps

Kohonen maps, also called self-organizing maps (SOM), have been introduced in 1981 by Kohonen. Fields like data exploratory analysis, web usage mining [14], industrial and medical diagnostics [15], and corruption analysis [16] are contemporary examples of SOM analysis applications and successes. This section is based on [17] and aims at giving a theoretical background to the reader. An application of the technique can be found in Section 2.4.1. The main objective of the SOM algorithm is the representation of a high dimensional input dataset on lower dimensional maps. This gives the possibility to explore the data and to use techniques like visual correlation analysis or clustering analysis in an intuitive manner. To do so, a feedforward Neural Network (NN) is trained on the input data. The output layer is a map with a lower dimensionality and a given number of neurons. During each iteration of the algorithm, an input data vector n_i is compared with the neurons m_r of the output map using Euclidian distances. The neuron m_c with the smallest distance with regard to the input vector is identified as the Best Matching Unit (BMU):

$$\|n_i - m_c\| = \min_r\{\, \|n_i - m_r\|\}. \tag{2.1}$$

The weights of the BMU are then modified in the direction of the input vector, leading to a self-organizing structure of the neurons. A learning rate $\alpha(t)$ and a neighborhood function $h_{cr}(t)$ are defined as parameters of the learning function:

$$m_r(t+1) = m_r(t) + \alpha(t)h_{cr}(t)[n(t) - m_r(t)]. \tag{2.2}$$

The learning-rate will influence the magnitude of the BMU's adaptation after matching with an input vector n_i, whereas the neighborhood function defines the range of influence of the adaptation. In order to guarantee the stability of the final output map, decreasing learning rates and neighborhood functions are often used at the end of the training. In order to evaluate the resulting output, the mean quantization error (MQE) and a topographic function are commonly used ([18,19]). On the one hand, the MQE measures the quality of the quantization by calculating the average distance separating the different input vectors from their BMUs. On the other hand, the topographic function will capture the degree of topology preservation of the output by analyzing the location of the neurons and their respective input vectors. An exhaustive discussion of the influence of the parameters such as the number of neurons, the shape of the map, or the initial weights of the neurons is to be found in [17].

2.2.2.2 The k-means algorithm

The k-means algorithm is a typical iterative distance-based clustering approach which iteratively creates k clusters based on the distances between data points. First, the number of clusters, k, has to be specified and k initial points are chosen as initial cluster centers. Different approaches exist in order to fix the number of clusters and to choose the initial centers (see [20]). However, the parameter k is often based on business knowledge and the k initial points are typically k data points randomly selected in the data set at hand. In a second step, all data points are assigned to their closest center according to the Euclidian distance. The mean, also called centroid, of each cluster is then calculated and used as new centers for the k clusters. The process

is repeated using the updated centers until the algorithm converges, meaning that the data points are assigned to the same centers in consecutive iterations. By choosing the cluster centers to be the centroids, the algorithm minimizes the total squared distance from each of the cluster's points to its center. This clustering method is simple and effective but it is important to notice that the obtained partition depends on the original cluster centers. Different initializations of the algorithm can lead to different results being local optima. It is thus advised to run the algorithm multiple times with different seeds as initial centers in order to augment the probability to find a global optimum.

2.2.2.3 Salient dimensions extraction

Extracting salient dimensions (SD) for automatic SOM labeling is a methodology developed by [21] and aims at identifying salient dimensions for clusters of SOM nodes. These salient dimensions are then used to label a SOM in an unsupervised way. The methodology is based on five main stages and starts with the training of a SOM using preprocessed data normalized within an input range of 0 to 1, followed by the clustering of the resulting nodes using any clustering technique. Pruning the nodes within the different clusters will lead to more homogeneous clusters and is the aim of the second step. This pruning phase is based on the mean and the standard deviation of the Euclidian distance between the centroid and the neurons of the different clusters. A parameter z_1 is used to identify the neurons to be pruned (the outliers or unlabeled neurons) and the neurons to be kept. The higher the value of z_1, the smaller the number of neurons pruned. The third step consists of identifying two sets for each cluster. The in-patterns set is defined and gathers all the individual training patterns belonging to the cluster. On the other hand, the out-patterns set consists of all the individual training patterns belonging to the other clusters or being attached to an unlabeled neuron identified in the second step. Using the sets defined in the previous step, the salient dimensions can then be identified for the clusters using a measure of

deviation in the statistical sense of the term. A difference factor is calculated for each dimension of all clusters and is used to identify the salient dimensions. A second parameter, z_2, is used to build a confidence interval around the mean of the difference factors of a cluster. A salient dimension will then be a dimension d, belonging to the set D gathering all the dimensions, for which the difference factor differs too much with regard to other dimensions within a cluster:

$$|df(k,d) - \mu_{df}| \geq z_2 \sigma_{df}(k), \tag{2.3}$$

with $df(k,d)$ being the difference factor for the dimension d of the cluster k, and $\mu_{df}(k)$ and $\sigma_{df}(k)$ respectively the mean and the standard deviation of the difference factors of the cluster k. The smaller the value of z_2, the larger the number of salient dimensions identified. The final step uses the different salient dimensions to label clusters with input from domain-specific experts. The result gives the possibility to label a new pattern using the label of the cluster to which it is attached. The formulas leading to the different statistics are to be found in detail in [21] and are discussed and adapted in Section 2.3.1.

2.3 SOM-based profile generator

This section is composed of Section 2.3.1 which proposes a generic method aiming at generating profiles based on the SOM approach and the salient dimensions extraction and Section 2.3.2 which defines different measures of performance applicable to the generated profiles.

2.3.1 The method

The general idea of the SOM-based profile generator consists of 5 main steps: (1) The generation of indices; (2) SOM training; (3) Clustering and SD extraction; (4) Generation of profiles; and (5) Ranking of profiles. Figure 2.1 schematizes these steps which are discussed in detail in what follows.

The first step consists of the preparation of the different indices that will be used during the training of the SOM. Categorical variables

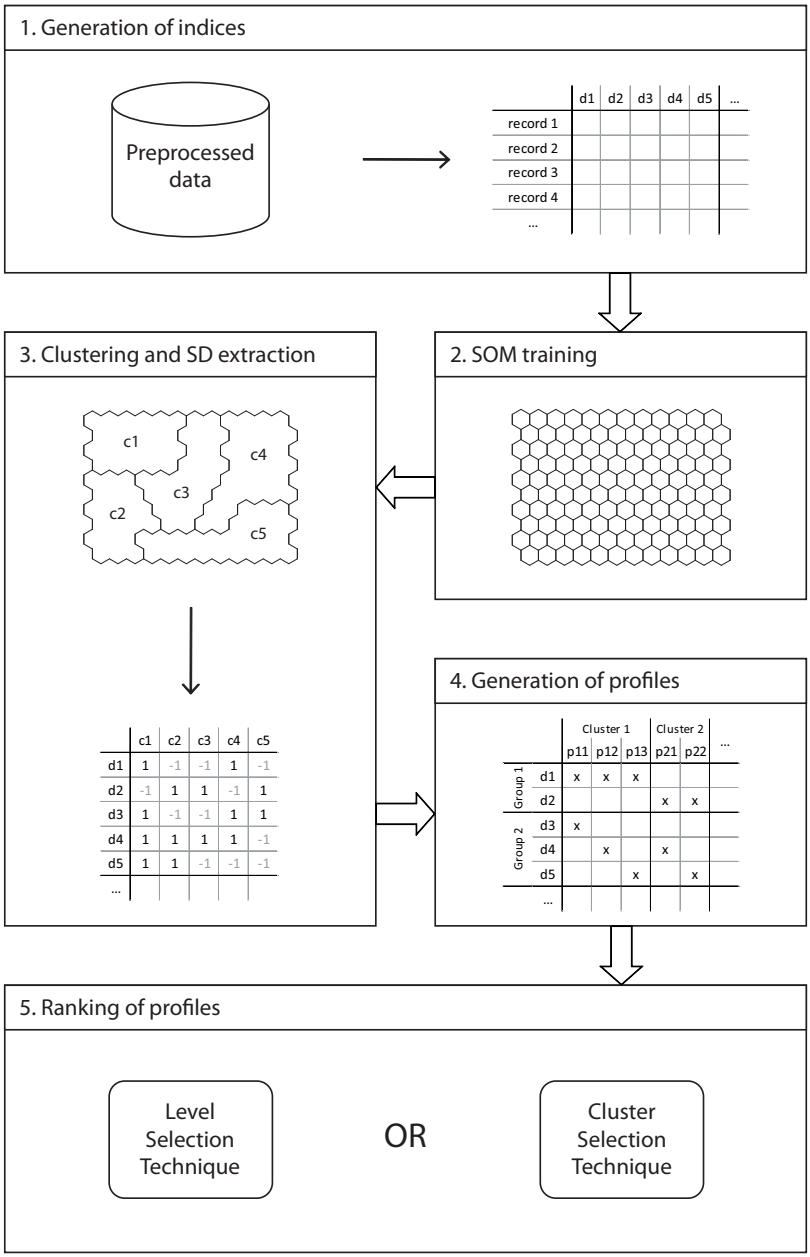


Figure 2.1: Figure schematizing the five steps of the SOM-based profile generator.

are preferred because of the way of using the extraction of the salient dimensions. Using continuous variables, it can only be defined whether a variable has high or low values with regard to other clusters. It is then better to categorize the continuous variables, giving the possibility to identify one or more of these categories as salient for a given cluster. Thus, a set N of input vectors n_i with $|D|$ dimensions is obtained. The value assigned to a dimension of n_i is either 1, if the input vector is characterized by the given dimension, or 0 if not.

During the second step, a SOM is trained using normalized values in the range of 0 to 1 as described in [21]. The reader interested in the details of the parametrization of a SOM analysis is referred to [17].

In the third step, the output map of the previous step is clustered by using any clustering technique, e.g. the k-means clustering widely discussed in the literature (e.g. [20]). The reader is referred to [22] for a justification of clustering in two steps. The method for extracting salient dimensions is a special case of the method developed in [21] with the first parameter, z_1, tending to infinity and the second parameter, z_2, being equal to zero. It corresponds to a case where no pruning of the clusters is performed and where all dimensions are either positive or negative salient dimensions. The in-patterns set $\phi_{in}(k)$ of the cluster k is defined as the set of all individual training patterns belonging to the cluster k:

$$n_i \in \phi_{in}(k) \Leftrightarrow \forall j \in K, \min_j(dist(c_j, n_i)) = dist(c_k, n_i), \tag{2.4}$$

with $dist(c_j, n_i)$ the Euclidian distance between the centroid of cluster j and the individual training pattern n_i, and K the set of all clusters identified using k-means clustering. The out-patterns set $\phi_{out}(k)$ of cluster k is computed by subtracting the in-patterns set of cluster k from the set N of all individual training patterns:

$$\phi_{out}(k) = N \setminus \phi_{in}(k). \tag{2.5}$$

In order to identify the salient dimensions, the following steps have to be processed for each cluster (steps 1, 2 and 3 being adapted from [21]):

1. For each dimension d, compute $\mu_{in}(k,d)$ and $\mu_{out}(k,d)$ as respectively the mean input value for the set of in-patterns $\phi_{in}(k)$ and out-patterns $\phi_{out}(k)$, where n_{id} is the dth component of the input vector n_i:

$$\mu_{in}(k,d) = \frac{\sum_{n_i \in \phi_{in}(k)} n_{id}}{|\phi_{in}(k)|}, \tag{2.6}$$

$$\mu_{out}(k,d) = \frac{\sum_{n_i \in \phi_{out}(k)} n_{id}}{|\phi_{out}(k)|}. \tag{2.7}$$

2. Compute the difference factor $df(k,d)$ of each dimension d as:

$$df(k,d) = \frac{\mu_{in}(k,d) - \mu_{out}(k,d)}{\mu_{out}(k,d)}. \tag{2.8}$$

3. Compute the difference factors mean $\mu_{df}(k)$ over all dimensions d as:

$$\mu_{df}(k) = \frac{\left(\sum_{d=1}^{D} df(k,d)\right)}{|D|}. \tag{2.9}$$

4. The salient dimension sign $sds(k,d)$ of the cluster k for the dimension d can then be computed as:

$$sds(k,d) = 1 \text{ if } df(k,d) \geq \mu_{df}(k), \tag{2.10}$$

or

$$sds(k,d) = -1 \text{ if } df(k,d) < \mu_{df}(k). \tag{2.11}$$

Based on the salient dimension signs, the profiles, each consisting of a set of dimensions, for a given set of targeted dimensions T are generated by using Algorithm 2.1, newly proposed in this chapter. This is the fourth step of the method.

[1] A group is defined as a set of dimensions having a real-world meaning to the user. An example of such a group could be the different dimensions resulting from the categorization of a variable (e.g. the variable could be the age variable, whereas the categories, such as [18..25], [26..35], [36..50], [51..65], and [66..], could be the dimensions of the group).

Algorithm 2.1 Generation of profiles

1: Define a set PD as a subset of D, containing the dimensions to be involved in the profile generation.
2: Define a set G of non overlapping groups[1] of dimensions from PD so that there is no dimension of PD not belonging to one group of G and no dimension of PD belonging to two different groups of G.
3: Define a set T of targeted dimensions such that each group of G is at most represented by one dimension in T.
4: Define a set of targeted groups TG composed of all groups having one dimension in T.
5: Define a set of untargeted groups UG composed of all groups of G not belonging to TG.
6: Define a set of selected clusters SC composed of the clusters having a positive salient dimension sign for all the targeted dimensions of the set T.
7: Create a list LSC composed of the clusters of the set SC.
8: Assign a score computed as the sum of the difference factors of the dimensions in T for each cluster $k \in LSC$ and rank the clusters in LSC in a decreasing order of their respective scores.
9: **for all** cluster k in SC **do**
10: Identify a set PC^k of all the possible combinations of the dimensions belonging to the groups of UG and having positive salient dimension signs, with maximum one dimension in each group of UG and minimum one dimension in a group of UG if there is at least a dimension in that group with a positive salient dimension.
11: Create a list LPC^k composed of the combinations of PC^k.
12: Assign a score to each combination in LPC^k computed as the sum of the difference factors of the dimensions involved in each combination and rank the combinations in LPC^k in a decreasing order of their respective scores.
13: **end for**

The fifth and last step of the method consists of the ranking of the profiles generated in the previous step using one of the two priority rules implemented by Algorithms 2.2 and 2.3, Cluster selection technique (CST) and Level selection technique (LST) respectively. The CST allows for an *intensification* approach. When a score is calculated for the different clusters obtained so far, the CST will select the best cluster and rank all related profiles before going to the next best cluster. The profiles generated using the best cluster have a higher rank than those generated by the second best cluster, etc. On the other hand, the LST allows for a *diversification* approach. The best profile of the best cluster is first considered, then the best profile of the second best cluster and so on until all selected clusters have been considered; then the second profile of the best cluster, etc. As can be concluded from the results obtained in the following sections, the choice of the ranking method has a substantial impact on the obtained performances. These two techniques are not exhaustive and should be challenged and combined with other ranking techniques in future research.

Algorithm 2.2 Cluster selection technique (CST)

1: An empty list of selected profiles LSP is created.
2: **if** the list LSC is empty **then**
3: The algorithm stops and the list of selected profiles LSP is returned.
4: **else**
5: Select the first cluster k of LSC.
6: **while** LPC^k is not empty **do**
7: Add the first combination of LPC^k to LSP.
8: Remove the first combination of LPC^k from LPC^k.
9: **end while**
10: Remove k from LSC and go back to line 2.
11: **end if**

Step four and five are contributions of this chapter and allow for the identification of profiles which have certain characteristics. The input

Algorithm 2.3 Level selection technique (LST)

1: An empty list of selected profiles LSP is created.
2: **if** the list LSC is empty **then**
3: The algorithm stops and the list of selected profiles LSP is returned.
4: **else**
5: Select the first cluster k of LSC.
6: Add the first combination of LPC^k to LSP.
7: Remove the first combination of LPC^k from LPC^k.
8: **if** LPC^k is not empty **then**
9: Rank k at the last position of LSC.
10: **else**
11: Remove k from LSC.
12: **end if**
13: Go to line 2.
14: **end if**

for these steps is the output of a SOM analysis and SD extraction. The final result is a list LSP of combinations of dimensions whose groups belong to UG. These combinations are the profiles containing the required targeted dimensions and are ranked in LSP according to importance as expressed in the chosen ranking technique.

2.3.2 Performance measures

The performance measure used to evaluate the performance of the generated profiles is expressed as the ratio between the degree of matching of the profiles generated in the previous section with a testing dataset TN and the degree of matching with TN of randomly generated profiles used as benchmark. Different performance measures are needed in order to express this gain.

A matching function $\theta(n_i, p)$ is used to express the similarity between a given input vector n_i, element of TN, and a profile p, element

of the list of selected profiles LSP:

$$\theta\left(n_i, p\right) = 1 \Leftrightarrow \text{ all the dimensions of p are equal to 1 in } n_i, \quad (2.12)$$

or

$$\theta\left(n_i, p\right) = 0 \Leftrightarrow \text{ at least one of the dimensions of p is equal to 0 in } n_i. \quad (2.13)$$

A second matching function $\lambda\left(n_i, sLSP\right)$ is defined and expresses whether or not at least one profile p belonging to a subset $sLSP$ of LSP matches with an input vector n_i:

$$\lambda\left(n_i, sLSP\right) = 1 \Leftrightarrow \exists p \in sLSP : \theta\left(n_i, p\right) = 1, \quad (2.14)$$

or

$$\lambda\left(n_i, sLSP\right) = 0 \Leftrightarrow \neg\exists p \in sLSP : \theta\left(n_i, p\right) = 1. \quad (2.15)$$

A matching ratio $\alpha\left(TN, sLSP\right)$ is then defined and returns the proportion of input vectors in TN matching with at least one profile p in $sLSP$:

$$\alpha\left(TN, sLSP\right) = \frac{\sum_{n_i \in TN} \lambda\left(n_i, sLSP\right)}{|TN|}. \quad (2.16)$$

The random performance $\chi\left(p\right)$ is defined as the probability of matching when using a randomly generated profile having 1 dimension in each group of UG for which p has a dimension:

$$\chi\left(p\right) = \prod_{d \in p} \frac{1}{|g(d)|}, \quad (2.17)$$

with d a dimension of p, and $g(d)$ the group to which d belongs. Using the random performance function $\chi\left(p\right)$, a second random performance function $\beta\left(sLSP\right)$ is defined as:

$$\beta(sLSP) = \sum_{p \in sLSP} \chi\left(p\right). \quad (2.18)$$

Finally, the gain $\pi(sLSP)$ obtained when using a given subset $sLSP$ on a testing dataset TN can be computed as:

$$\pi(TN, sLSP) = \frac{\alpha\left(TN, sLSP\right)}{\beta(sLSP)}. \quad (2.19)$$

2.4 Market segmentation in the concert industry: an application of the SOM-based profile generator

This section presents a case study carried out in collaboration with Ticketmatic, one of the leading ticketing software companies in Europe. To apply the SOM-based profile generator, online ticket sales data of one concert organizer collected during the period 2007-2010 was used. The data consisted of 63.000 records gathering information about tickets sold as summarized in Table 2.1.

The application programming interface (API) provided by the *last.fm* website[2] was used to gather tags about artists involved in the different concerts. These tags are provided by the *last.fm*'s users and are gathered in a database accessible via the API.

The preprocessing consisted of the selection of all records having values for all the attributes of Table 2.1 combined with a basic outlier detection procedure using the mean and standard deviation of the geographic coordinates to identify geographically isolated customers. A further preprocessing step was to prune all records related to concerts of which the artists were not tagged in the *last.fm* API.

2.4.1 Application of the profiling method

The objective was to profile the customers of the available dataset in order to predict the profiles of the customers potentially interested in a specific future concert, which will be called *The Concert* in the remainder of this chapter. To do so, the preprocessed dataset was divided in two subdatasets. The first one was composed of all the records related to tickets sold for *The Concert* and was used as test dataset while the second one was composed of all records related to tickets sold for other concerts and was used as training dataset. Starting with the preprocessed data about the sold tickets described in Table 2.1, three categories of indices were developed for all customers of both

[2]http://www.lastfm.fr/api

Variable name	Description	Type
ticketID	The ID of the ticket.	integer
creationDate	The date of the purchase.	date
customerID	The ID of the customer related to the ticket.	integer
basketID	The ID of the basket related to the ticket	integer
totalAmountBasket	The price of the basket the ticket is related to.	float
concertName	The name of the concert the ticket gives access to.	string
geoLong	The longitude from which the customer purchased the ticket.	float
geoLat	The lattitude from which the customer purchased the ticket.	float
birthdate	The birthdate of the customer.	date
gender	The gender of the customer.	char

Table 2.1: Summary of the information about tickets sold.

subdatasets. The three RFM variables were constructed for each customer, giving the possibility to rank them based on a score ranging from 1 to 5. An index *total RFM* was then computed by summing the three values of the RFM variables, leading to a score between 3 and 15 for each customer. Four categories of customers were defined based on their respective *total RFM* indices using the following intervals: [3..5], [6..8], [9..11] and [12..15]. The birthdate of the different customers was used to generate an index capturing the age. Five categories were defined using the following intervals in years: [18..25], [26..35], [36..50], [51..65] and [66..]. The gender of the customers was used to build two extra categories. The information captured under the variables *geolong* and *geolat*, using the IP addresses of the booking computers, gave the possibility to obtain an index representing the geographic distance separating the customer from the concert infrastructure. Categories were defined using following intervals in km: [0..5], [6..10], [11.15], [16..25], [26..50] and [51..]. In order to define an index capturing the interest of a given customer for a tag, an artist or a concert, the *last.fm* API data was used in combination with data of the previous concerts involving the given customer. Considering that a concert consists of a series of artists characterized by a series of tags, it is then possible to rank the customers according to their score for a given concert as described in Algorithm 2.4. Note that this interest-based variable is based solely on tags of previously attended concerts, and not on information of *The Concert* itself. In the application, the score for a given concert, *The Concert*, was used as last index and, combined with the other indices of the previous categories, led to a total of 18 dimensions as summarized in Table 2.2.

A 10x12 SOM was trained, using as input vectors the normalized values in the range of 0 to 1 of the 18 dimensions for each customer of the training dataset. A k-means clustering was performed on the neurons of the generated SOM starting with k equal to 10 and selecting the best k using the Davies-Bouldin index[3] value as decision criterion. The Davies-Bouldin index led to the decision to choose a

[3]The interested reader is referred to [23] for more information about the Davies-Bouldin index.

Algorithm 2.4 Score of a given customer for a given concert

1: Define the sets C, R, A and T as respectively the sets of all the customers, concerts, artists and tags.
2: Select a concert r from R.
3: Select a customer c from C the score of which must be calculated for the concert r.
4: **for all** $t \in T$ **do**
5: Define a set A^t composed of all the artists in A having t as one of their tags.
6: Define a set R^t composed of all the concerts in R having minimum one artist in A^t.
7: **end for**
8: Define a set R^c composed of all the concerts in R customer c attended.
9: Define a set A^r composed of all the artists in A related to r.
10: **for all** $a \in A^r$ **do**
11: Define a set T^a composed of all the tags t from T related to a.
12: **end for**
13: Compute the score of the customer c for the concert r as: $score(c, r) = \sum_{a \in A^r} \sum_{t \in T^a} \frac{\left|R^t \cap R^c\right|}{|R^c|}$
14: **return** $score(c, r)$

Dimension	Index category	Original variable	Range	Index name	Value
D1	Demographic	Sex	M	Sex Man	$\{0, 1\}$
D2	Demographic	Sex	F	Sex Woman	$\{0, 1\}$
D3	Demographic	Age	18..25	Age 18-25	$\{0, 1\}$
D4	Demographic	Age	25..35	Age 25-35	$\{0, 1\}$
D5	Demographic	Age	35..50	Age 35-50	$\{0, 1\}$
D6	Demographic	Age	50..56	Age 50-56	$\{0, 1\}$
D7	Demographic	Age	65..	Age 65-more	$\{0, 1\}$
D8	Demographic	Distance	0..5	Distance 0-5	$\{0, 1\}$
D9	Demographic	Distance	5..10	Distance 5-10	$\{0, 1\}$
D10	Demographic	Distance	10..15	Distance 10-15	$\{0, 1\}$
D11	Demographic	Distance	15..25	Distance 15-25	$\{0, 1\}$
D12	Demographic	Distance	25..50	Distance 25-50	$\{0, 1\}$
D13	Demographic	Distance	50..	Distance 50-more	$\{0, 1\}$
D14	RFM	Total rfm	3..5	Total rfm 1	$\{0, 1\}$
D15	RFM	Total rfm	6..8	Total rfm 2	$\{0, 1\}$
D16	RFM	Total rfm	9..11	Total rfm 3	$\{0, 1\}$
D17	RFM	Total rfm	12..15	Total rfm 4	$\{0, 1\}$
D18	Interest-based	The Concert	0..	The Concert	[0..1]

Table 2.2: Summary of the indices generated in the first step of the SOM-based profile generator.

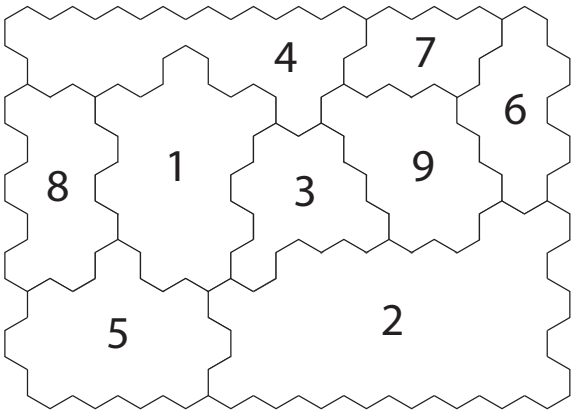


Figure 2.2: Visualization of the clustering of the 10x12 SOM leading to nine clusters.

k of 9 as shown in Figure 2.2. Next, a salient dimensions analysis was performed using the adapted method presented in Section 2.3.1. Table 2.3 shows the results of the difference factor computation for the different clusters and dimensions, to be compared with the mean difference factor of each cluster as explained in Section 2.3.1.

Using Formulas 2.10 and 2.11 of Section 2.3.1, the $sds(k, v)$ values were computed for each dimension and for each cluster as shown in Table 2.3 where a bold number represents a $sds(k, v)$ equal to one and a number in normal script a $sds(k, v)$ equal to zero. Then, Algorithm 2.1 was applied using the results of the previous steps. A detailed sequence of its application is to be found in what follows:

1. $PD = \{D_1, D_2, D_3, D_4, D_5, D_6, D_7, D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}, D_{14}, D_{15}, D_{16}, D_{17}, D_{18}\}$
2. $G = \{\{D_1, D_2\}, \{D_3, D_4, D_5, D_6, D_7\}, \{D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}\}, \{D_{14}, D_{15}, D_{16}, D_{17}\}, \{D_{18}\}\}$
3. $T = \{D_{18}\}$
4. $TG = \{D_{18}\}$
5. $UG = \{\{D_1, D_2\}, \{D_3, D_4, D_5, D_6, D_7\}, \{D_8, D_9, D_{10}, D_{11}, D_{12}, D_{13}\}, \{D_{14}, D_{15}, D_{16}, D_{17}\}\}$
6. $SC = \{3, 5, 8\}$
7. $LSC = (3, 5, 8)$

	Clusters								
Dim	1	2	3	4	5	6	7	8	9
D1	**0,73**	-0,93	**0,49**	**0,79**	-0,94	**0,67**	**0,70**	**0,52**	0,57
D2	-0,93	**3,74**	-0,71	-1,00	**2,03**	-0,95	-1,00	-0,73	-0,78
D3	-0,72	-0,40	-0,87	**0,38**	**1,77**	-0,54	**0,57**	-0,30	**1,07**
D4	**1,46**	-0,28	**1,26**	-0,55	**0,14**	-1,00	-0,99	**1,08**	-0,72
D5	-0,88	**1,22**	-0,66	**0,37**	-1,00	**1,97**	**1,24**	-1,00	-0,31
D6	-0,62	-0,57	-0,88	**0,71**	-0,11	**0,27**	-0,15	-0,22	**3,29**
D7	-0,74	-0,83	-0,61	**0,49**	-0,21	-0,21	**0,45**	-0,82	**7,14**
D8	-0,05	0,13	**0,28**	-0,64	0,01	**0,16**	-0,52	**0,48**	0,13
D9	-0,31	**0,53**	-0,26	0,01	-0,32	**0,48**	-0,12	-0,44	0,08
D10	-0,21	0,07	-0,19	**0,17**	-0,06	**0,18**	0,01	**-0,09**	0,09
D11	-0,03	0,01	**-0,03**	0,04	-0,07	-0,23	**0,23**	**0,20**	-0,01
D12	**0,09**	-0,14	**-0,01**	**0,65**	**0,09**	-0,20	**0,37**	-0,61	-0,10
D13	**0,33**	-0,34	-0,23	**0,21**	**0,16**	-0,12	**0,27**	**0,37**	-0,19
D14	**2,14**	0,07	-0,53	-0,59	-0,57	-0,38	**1,96**	-0,86	-0,30
D15	-0,61	-0,44	-0,74	**1,51**	**1,58**	-1,00	-0,74	**1,78**	-0,78
D16	-0,74	**0,60**	**1,89**	-0,91	-1,00	**1,89**	-0,66	-1,00	**1,63**
D17	**0,81**	**0,85**	-0,39	-0,85	-0,69	**0,89**	0,05	-0,94	0,10
D18	-0,13	0,10	**-0,05**	-0,02	**0,09**	0,00	-0,03	**-0,02**	-0,05
μ_{df}	-0,02	0,19	-0,12	0,04	0,05	0,10	0,09	-0,14	0,60

Table 2.3: Table of the difference factors for each dimension and for each cluster, with bold numbers indicating positive salient dimensions signs.

8. $score(3) = -0.051$
   $score(5) = 0.093$
   $score(8) = -0.016$

   $LSC = (5, 8, 3)$

9. (a) $PC^3 = \{\{D_1, D_4, D_8, D_{16}\}, \{D_1, D_4, D_{11}, D_{16}\}, \{D_1, D_4, D_{12}, D_{16}\}\}$
   (b) $LPC^3 = (\{D_1, D_4, D_8, D_{16}\}, \{D_1, D_4, D_{11}, D_{16}\}, \{D_1, D_4, D_{12}, D_{16}\})$
   (c) $score(\{D_1, D_4, D_8, D_{16}\}) = 3,920$
   $score(\{D_1, D_4, D_{11}, D_{16}\}) = 3,607$
   $score(\{D_1, D_4, D_{12}, D_{16}\}) = 3,626$

   $LPC^3 = (\{D_1, D_4, D_8, D_{16}\}, \{D_1, D_4, D_{12}, D_{16}\}, \{D_1, D_4, D_{11}, D_{16}\})$

   (a) $PC^5 = \{\{D_2, D_3, D_{12}, D_{15}\}, \{D_2, D_3, D_{13}, D_{15}\}, \{D_2, D_4, D_{12}, D_{15}\}, \{D_2, D_4, D_{13}, D_{15}\}\}$
   (b) $LPC^5 = (\{D_2, D_3, D_{12}, D_{15}\}, \{D_2, D_3, D_{13}, D_{15}\}, \{D_2, D_4, D_{12}, D_{15}\}, \{D_2, D_4, D_{13}, D_{15}\})$
   (c) $score(\{D_2, D_3, D_{12}, D_{15}\}) = 5,470$
   $score(\{D_2, D_3, D_{13}, D_{15}\}) = 5,539$
   $score(\{D_2, D_4, D_{12}, D_{15}\}) = 3,833$
   $score(\{D_2, D_4, D_{13}, D_{15}\}) = 3,903$

   $LPC^5 = (\{D_2, D_3, D_{13}, D_{15}\}, \{D_2, D_3, D_{12}, D_{15}\}, \{D_2, D_4, D_{13}, D_{15}\}, \{D_2, D_4, D_{12}, D_{15}\})$

   (a) $PC^8 = \{\{D_1, D_4, D_8, D_{15}\}, \{D_1, D_4, D_{10}, D_{15}\}, \{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\}\}$
   (b) $LPC^8 = (\{D_1, D_4, D_8, D_{15}\}, \{D_1, D_4, D_{10}, D_{15}\}, \{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\})$
   (c) $score(\{D_1, D_4, D_8, D_{15}\}) = 3,856$
   $score(\{D_1, D_4, D_{10}, D_{15}\}) = 3,287$
   $score(\{D_1, D_4, D_{11}, D_{15}\}) = 3,580$
   $score(\{D_1, D_4, D_{13}, D_{15}\}) = 3,752$

   $LPC^8 = (\{D_1, D_4, D_8, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\}, \{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{10}, D_{15}\})$

After this step, the three lists LPC^3, LPC^5 and LPC^8 contained the generated profiles targeting the dimensions d belonging to T. So far, the profiles of the customers interested in *The Concert* had been generated and ranked within each cluster and the clusters had been ordered. The final step consisted of selecting the profiles in order to generate a priority list LSP as described in Section 2.3.1. The interested reader can apply both ranking techniques on the output of the previous step and compare the results with the following resulting priority lists of selected profiles LSPs, generated using respectively CST and LST.

1. Cluster Selection Technique:

   $LSP = (\{D_2, D_3, D_{13}, D_{15}\}, \{D_2, D_3, D_{12}, D_{15}\},$
   $\{D_2, D_4, D_{13}, D_{15}\}, \{D_2, D_4, D_{12}, D_{15}\}, \{D_1, D_4, D_8, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\},$

$\{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{10}, D_{15}\}, \{D_1, D_4, D_8, D_{16}\}, \{D_1, D_4, D_{12}, D_{16}\},$
$\{D_1, D_4, D_{11}, D_{16}\})$

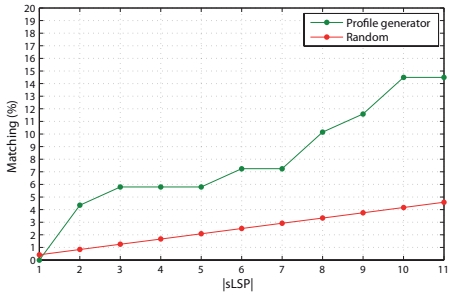
2. Level Selection Technique:

$LSP = (\{D_2, D_3, D_{13}, D_{15}\}, \{D_1, D_4, D_8, D_{15}\},$
$\{D_1, D_4, D_8, D_{16}\}, \{D_2, D_3, D_{12}, D_{15}\}, \{D_1, D_4, D_{13}, D_{15}\}, \{D_1, D_4, D_{12}, D_{16}\},$
$\{D_2, D_4, D_{13}, D_{15}\}, \{D_1, D_4, D_{11}, D_{15}\}, \{D_1, D_4, D_{11}, D_{16}\}, \{D_2, D_4, D_{12}, D_{15}\},$
$\{D_1, D_4, D_{10}, D_{15}\})$

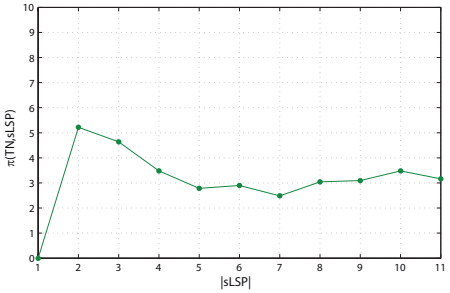
2.4.2 Performance of the generated profiles

The aim of this section is to evaluate the profiles generated by the method developed in this chapter. As introduced in Section 2.4, the profiles were generated using the dataset not incorporating tickets sold for the concert we targeted in order to test the algorithm on an independent test set. The second dataset will now be used to evaluate the quality of the generated profiles. The generation of indices is performed on the subdataset gathering the records of tickets sold for *The Concert* in the way described in Section 2.4. The 18th dimension concerning the interest for the concert is not considered in this section for obvious reasons leading to a test set TN of input vectors with 17 dimensions corresponding to all the customers related to the tickets sold for *The Concert*.

The performance of the profile generator can be compared with the performance of a random tool using the defined measures of performance. Figure 2.3.1 shows the matching function $\alpha(TN, sLSP)$ and the random performance function $\beta(sLSP)$ for different subsets $sLSP$. The number of profiles in $\beta(sLSP)$ is given on the X axis, whereas the Y axis represents the difference between $\alpha(TN, sLSP)$ and $\beta(sLSP)$, giving the added value of the profile generator with regard to a random prediction. The ratio of both $\alpha(TN, sLSP)$ and $\beta(sLSP)$ on the other hand gives the gain $\pi(TN, sLSP)$, expressing how many times the profile generator is better, or worse, than the random profile generator. Figure 2.3.2 is a graph of the gain $\pi(TN, sLSP)$ for the different numbers of selected profiles in $sLSP$. A value greater

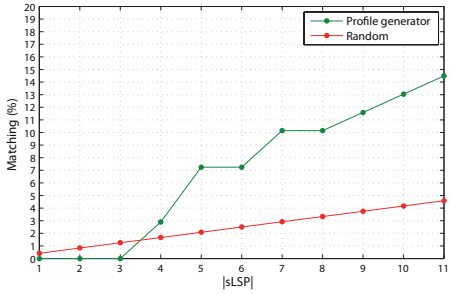


2.3.1: Performance for The Concert using the LST

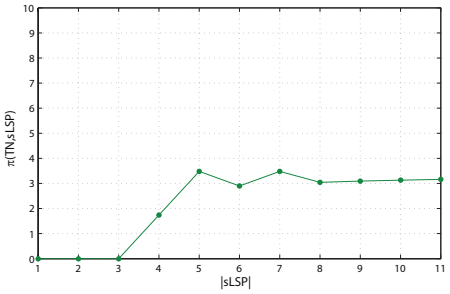


2.3.2: Gain for The Concert using the LST

Figure 2.3: Performance and gain for The Concert using the LST.



2.4.1: Performance for The Concert using the CST

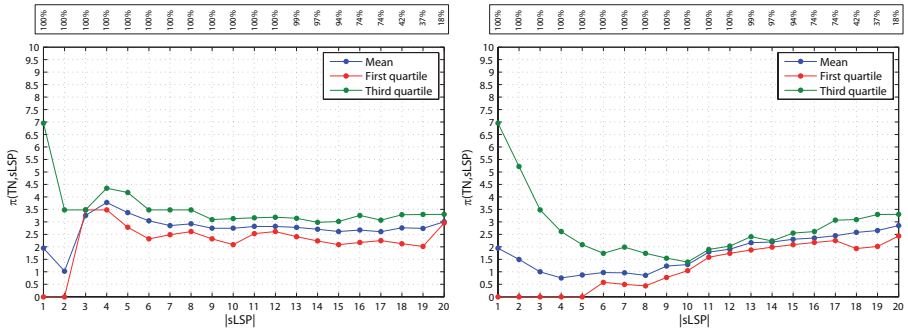


2.4.2: Gain for The Concert using the CST

Figure 2.4: Performance and gain for The Concert using CST.

than 1 implies an improvement of the prediction power when the profile generator is used. Figures 2.3.1 and 2.3.2 show that, using the LST, two profiles are needed to outperform a random generation of the profiles.

Figures 2.4.1 and 2.4.2 show the same functions as in Figure 2.3.1 and 2.3.2 when the profile generator is used combined with CST. It is obvious that the order of selection of the profiles influences the performance as can be seen when comparing Figures 2.3.2 and 2.4.2. When CST is used, four profiles are needed to outperform the random method instead of the two needed when using LST.



2.5.1: Gain for The Concert using the LST

2.5.2: Gain for The Concert using the CST

Figure 2.5: Gain for The Concert using LST and CST.

The previous graphs showed the performance of the profiles generated using the SOM-based profile generator while targeting *The Concert*. However, a certain variability of the results is to be captured because of the clustering technique used in our application, the k-means clustering. The number of clusters k and the form of the clusters have indeed a huge impact on the generated profiles and the k-means technique provides no guarantee concerning the stability of the clusters generated. The point is thus to know whether the obtained results are average results or singularities. In order to give an answer to this question, the method was applied 100 times using the same data and targeting the same concert. The results are shown on Figures 2.5.1 and 2.5.2, for LST and CST respectively. The mean gain obtained for a subset of profiles, the size of which is given on the X axis, and the first and third quartile values are plotted. So far, an important consideration is that some clustering iterations lead to different numbers of profiles. In order to capture the fact that some statistics are not representative for the 100 iterations, this representativity is expressed as a percentage of the executions of the method having generated at least a number $|sLSP|$ of profiles, and is indicated on the top of each subfigure for each number of profiles $|sLSP|$. For this representativity reason, the statistics up to a $|sLSP|$ of 20 profiles have been plotted. A discussion on the impact of factors such as the selection technique

used or the amount of available data is to be found in the next section.

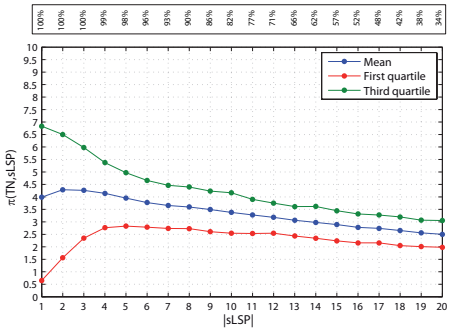
2.5 Discussion

This section will focus on two main points. First of all, the impact of the amount of available data and the selection technique used will be discussed in Section 2.5.1 and will lead to more insight about the proposed method. Finally, Section 2.5.2 will introduce potentially interesting topics for future research with regard to the proposed method.

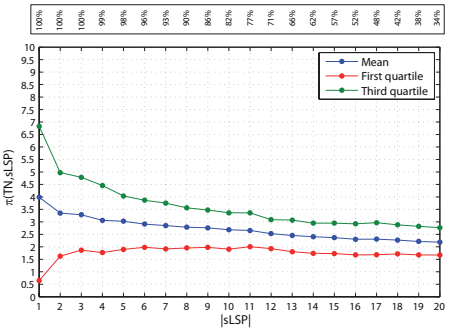
2.5.1 Impact of the parameters

The objective of this section is to answer the two following questions: (1) Is LST better or worse than CST, and (2) does more data lead to better results? To do so, an experiment has been set up using the data described in Section 2.4. The tested factors are the amount of data, consisting of the full dataset, half the dataset, or a fourth of the dataset, and the selection technique used being either LST or CST. The combination of these factors leads to six different experiments involving all the concerts having minimum 100 tickets in the preprocessed dataset in order to have an acceptable amount of data in the test set, leading to 107 concerts meeting the requirements. For each of these experiments, the SOM-based profile generator was applied 100 times for each of the 107 concerts taking as targeted dimension the interest for the given concert as applied in Section 2.4.1. The 100 iterations of the method are needed in order to capture the variability introduced by the k-means technique as mentioned in Section 2.4.2. Figure 2.5.1 shows the results of the six experiments where, for each subfigure representing one of the experiments, the relevant statistics are averaged over the 107 concerts.

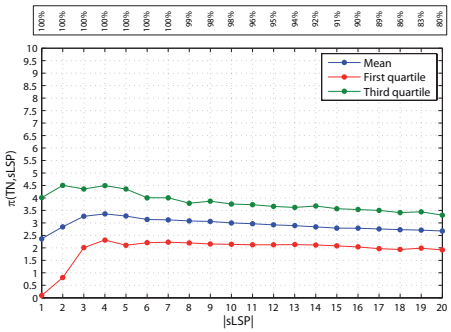
Figure 6 summarizes the 64.200 executions of the method needed to perform the experiment, and enables to answer the two questions introduced in this section. Comparing Figures 2.6.1, 2.6.3, and 2.6.5 with 2.6.2, 2.6.4, and 2.6.6 respectively, a clear outperformance by



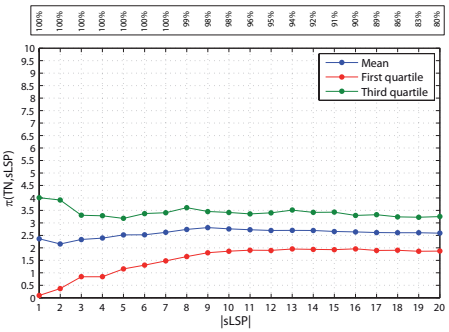
2.6.1: Gain using the full dataset and LST



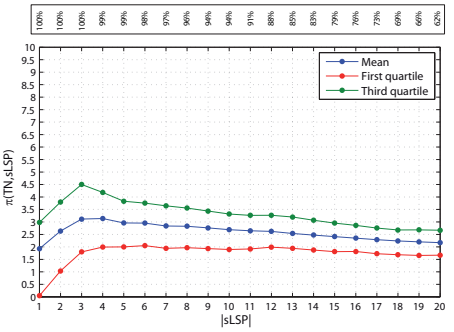
2.6.2: Gain using the full dataset and CST



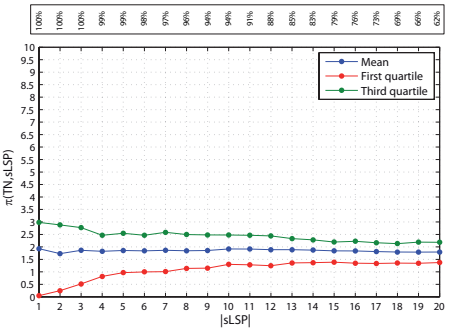
2.6.3: Gain using half the dataset and LST



2.6.4: Gain using half the dataset and CST



2.6.5: Gain using a fourth of the dataset and LST



2.6.6: Gain using a fourth of the dataset and CST

Figure 2.6: Output of the six experiments testing the factors *Amount of available data* and *Ranking technique*.

LST can be seen. It must be noted that the values for the three statistics for the first subset $sLSP$ and the last one (not plotted here) are the same whether LST or CST is used because the start and the finish points of both techniques are the same. The concave curvature of the mean gain, when LST is used, leads to better results than the more convex curvature resulting from the usage of CST. This conclusion is verified for the three levels of the tested factor *Amount of data*, leading to a preference for LST, thus answering the first question of the preceding paragraph. Concerning the impact on the performance of the available amount of data, comparing Figures 2.6.1 and 2.6.2 with Figures 2.6.3 and 2.6.4 respectively, leads to the conclusion that using the full dataset results in better gains for both LST and CST with regard to using half of the dataset. Moreover, comparing Figures 2.6.3 and 2.6.4 with Figures 2.6.5 and 2.6.6 respectively, leads to the conclusion that using half the dataset results in better gains for both LST and CST with regard to the usage of a fourth of the dataset. Both these conclusions can then be generalized by the following: using more data as input for the SOM-based profile generator leads to better profiles with regard to the gain they provide, hereby answering the second question of the previous paragraph.

2.5.2 Further research

In this section, the different steps of the method proposed in Section 2.3.1 will be challenged, leading to different tracks for further research.

The first step of the method, the generation of indices, has been introduced in Section 2.3.1 and extended in Section 2.4 for the purpose of the application. A wider research topic could be the analysis of the impact of the curse of dimensionality on the performance of the proposed method. As mentioned in Section 2.3.1, categorical variables are preferred because of the information provided by the salient dimensions analysis. However, the granularity of the categories should be studied in order to propose a method leading to a definition of an optimal setup for the usage of the proposed SOM-based profile

generator.

The second step of the method, the SOM training, is introduced in this chapter as a black box. A more extensive analysis of the impact on the generated profiles of SOM parameters such as the number of neurons, the shape of the SOM, the chosen learning rate and the used neighboring function, should lead to a better insight and a definition of best practices with regard to training the SOM as a step of the proposed method.

The third step, clustering and salient dimensions analysis, is introduced in Section 2.3.1 and the impact of the clustering technique used is discussed in Section 2.4.2 and Section 2.5.1 where the variability introduced by the clustering technique is analyzed. It should now be clear for the reader that the impact of the clustering technique is not to be neglected and further research could focus on the granularity of the clusters generated and the way of generating them. In fact, the same experiment as the one presented in Section 2.4 has been conducted using hierarchical ascending clustering (HAC) instead of k-means clustering. Although HAC is often used in the literature when SOMs are clustered (e.g. [24]), a comparison with the k-means using the same data and parameters showed a better performance of the k-means. These results are not shown here because of lack of space but further research should focus on the definition of decision criteria for the choice of the clustering method to use in the third step of the proposed framework. Moreover, other techniques than SOM could be used in the second step of the method as mentioned in [22] and could lead to new results and insights. The salient dimensions analysis presented in [21] has been adapted in Section 2.3.1 in order to capture more information while suppressing the arbitrary fixing of the values of z_1 and z_2. Further research could identify another way to determine the sensibility of the salient dimensions analysis based on a more formal and statistical approach.

The fourth step, the generation of profiles, is presented in Section 2.3.1 as Algorithm 2.1. Evolutionary algorithms could be studied in combination with the generated profiles in order to increase the generated population of profiles.

Finally, the fifth step, the ranking of the profiles, should offer a multitude of tracks for further research. A lot of priority rules are indeed conceivable, including LST and CST, and could increase the performance of the proposed method, depending on the application domain and the chosen parameters in the previous steps of the method.

2.6 Conclusion

The main contribution of this chapter to the literature is the development of a generic method for profile generation which is applicable to all cases where the SOM technology is used. The method enables to formalize intuitive feelings and insights resulting from the combination of a SOM analysis and the extraction of salient dimensions.

The SOM-based profile generator proposed in this chapter can be used to go further than a classical SOM analysis. As illustrated, the interest of the method resides in its capacity to generate value-adding profiles targeting given dimensions using the SOM technology and the extraction of salient dimensions. However, a SOM-based analysis, which itself provides valuable output given its visualization power, is not to be replaced but reinforced by the generated profiles.

The performance of the proposed method has been illustrated by an application in the concert industry, showing a real added value while identifying factors, such as the available amount of data or the ranking technique used, being potential sources of improvement. The results concerning the importance of the available amount of data should be an incentive for companies, which are aiming at building customer profiles, to aggregate their data. By translating the gain in prediction power of the generated profiles in terms of money, it is possible to assess the value of the available data and create new business models. Moreover, it should be clear for the reader that the two ranking techniques proposed in this chapter are a starting point for future improvement of the method. It can already be concluded that LST outperforms CST, leading to the insight that other techniques could increase the power of the proposed method.

Chapter 3

A new knowledge-based constrained clustering approach: theory and application in direct marketing

3.1 Introduction

Data mining techniques and tools have been responsible for many of artificial intelligence's recent successes (e.g. [25], [26] and [27]). Amongst these techniques, clustering has always been an exploratory but critical task in the knowledge discovery process and has been applied in nearly all domains in which the grouping of similar objects makes sense (e.g. [28], [29] and [30]). Ranging from the most simple techniques, such as the k-means algorithm (e.g. [31] and [16]), to the most advanced approaches, such as kernel methods ([32]) and spectral approaches ([33]), clustering techniques have received interest from both the scientific and the business community. The users of such techniques are sometimes in possession of background knowledge and would like to include it into the clustering exercise. The

usage of constraints in order to integrate this knowledge into the clustering task, i.e. constrained clustering, is a current active research topic leading to different approaches and techniques (see e.g [34], [35] and [36]). Amongst the different constraints' levels, instance-level constraints, based on pairwise information, such as the famous *must-link* and *cannot-link* ([34]), are widely discussed in the literature ([37], [38], [39], [40], etc.).

Although this kind of constraints is quite known, consider the myriad of papers dealing with semi-supervised clustering, other interesting levels have emerged in the literature and are of great interest when dealing with knowledge integration. In [41], the authors propose a variant of the k-means algorithm with cluster-level constraints, ensuring that no empty clusters or clusters with very few data points are obtained. In [42], an algorithm constrained by feature order preferences is presented, in which attribute-level constraints are used as part of the to-be-optimized objective function. Attribute-level constraints are also present in [36] where *must-link* and *cannot-link* are adapted by creating constraints on the attributes' values. Moreover hybrid approaches integrating different levels have been proposed, see e.g. the approaches of [43] in which both instance- and attribute-level constraints are used in order to guide the clustering task.

Another important dimension concerns the degree to which the constraints have to be satisfied, leading to the concepts of hard and soft constraints. On the one hand, hard constraints are constraints that are required to be fully satisfied. For example, the COP-KMEANS algorithm, presented in [34], requires the assignment of each point to a cluster such that instance-level constraints are not violated. If no such cluster exists, the algorithm fails. The same reasoning is used in [36] in which the mlx-k-Medoids algorithm is presented which fails if no cluster not violating a set of attribute-level constraints can be found. [41] proposed a cluster-level constrained version of the k-means algorithm, which has to satisfy k hard constraints imposing that a cluster k has to have at least τ_k data points. On the other hand, soft constraints are used to guide the algorithm while accepting a partial violation (satisfaction) of the constraints. For example, [44] introduces the notion of

soft-constraints in a soft version of COP-KMEANS, SKOP-MEANS, in which a strength factor α is used to indicate the reliability of a constraint. The objective function is penalized if constraints are violated proportionally to the value of their respective α, allowing a violation of the constraints while trying to minimize it. The same approach is used in [45] in which the authors apply soft-constraints in mixture clustering by using fuzzy constraints and a strength factor γ and optimize an objective function minimizing the constraints' violations. In [42] and [35], parameters λ and w are respectively used in order to fix the significance of the added constraints. Finally, in [46], the authors are using pairwise judgments of similarity and dissimilarity in a soft way by trying to find a partitioning of the vertices into clusters so that the number of violations is minimized.

In this chapter, a new approach based on business considerations is proposed in order to incorporate business knowledge into the clustering task in an easy and efficient way. The goal of this approach is to focus on the perceived value of the partitioning resulting from the clustering task and not only on the statistical aspects of it (see e.g. [35]). This approach is based on the fact that business people, experts or not, have some insights regarding the importance of the variables before actually starting the analysis. If this insight is limited, an unconstrained approach has to be used, which has already been widely discussed in the literature. However, if this insight is considered sufficient, new approaches are needed in order to consider the a priori knowledge as a critical input for the clustering task. By incorporating this knowledge, the comprehensibility and the perceived value of the clustering will increase. From a more technical point of view, we propose a straightforward approach to transform background knowledge about features' importance into a metric that is used, as an example, to constrain the Self-Organizing Map algorithm in a soft way, leading to a soft-constrained attribute-level clustering approach based on metric learning. In a related work, [42] propose a solution to a problem which looks quite similar to the one this chapter is tackling but which, in fact, is totally different. [42] propose a formulation of a clustering objective function penalizing the violation of feature order preferences

making use of background knowledge about the importance of the features in order to create soft attribute-level constraints by parameterizing a weighted distortion measure. This objective function is further incorporated into a prototype-based clustering algorithm in which an iterative approach is used to converge to an accurate partitioning of the data points. Although the approach of [42] and the one proposed in this chapter make use of feature order preference, the purposes of both methods are different. In [42], the approach and the algorithm lead to a better capturing of the ground truth if some background knowledge about the importance of the variables is available. This notion of importance is objective and is purely data-driven. Therefore, during their experiments, the "true" number of clusters is provided and simulated feature order preferences are generated using the ground truth class information. This idea of ground truth information is also exploited in [47], where a semi-supervised clustering method for incorporating instance-level and attribute-level information is proposed. In their work, attribute-level constraints are in the form of order preferences, generated using ground truth class information which enables the calculation of rough estimates of the optimal attribute weights leading to the order preferences. For their experiments, 6 UCI data sets with known class information are used to evaluate the proposed methods. In contrast, the proposed chapter introduces a subjective, goal-driven notion of importance. Indeed, the soft attribute-level constraints of this chapter are based on feature order preferences that reflect the importance of the variables as perceived by the analyst, hence introducing a bias guiding the algorithm and providing the analyst with a powerful exploratory knowledge-based tool.

The remainder of this chapter is structured as follows. Section 3.2 presents the approach for the prioritization of the variables in the clustering task. In Section 3.3, a 5-step methodology implementing the proposed prioritization approach is described. In Section 3.4, the theory is illustrated in a direct marketing context by a comparison between the results of the proposed approach and the traditional approach. The main conclusions are summarized in Section 3.5. Note that this chapter makes use of three techniques introduced in the pre-

vious chapter (see Sections 2.2.2 and 2.3), namely the self-organizing maps, the k-means algorithm and the extraction of salient dimensions.

3.2 Prioritization approach

This section deals with the prioritization approach which is the main contribution of this chapter. It consists of the formalization of the fact that a clustering task, such as segmentation, should not be entirely unsupervised if interesting insights into the importance of the variables exist. In other words, if the analyst is able to assign priorities to the variables he is using, he should be able to do so. The following sections are discussing approaches enabling the integration of this knowledge. In Section 3.2.1, an approach to integrate the knowledge using clustering techniques based on an adapted weighted distance metric is proposed. In Section 3.2.2 a method to transform the knowledge from priorities to weights is presented.

3.2.1 Incorporating business knowledge

The context required to use the knowledge in the clustering task can be described as follows. Define a set $\mathcal{N}$ gathering n input vectors n_i composed of the d quantitative dimensions d_j of the set $\mathcal{D}$, resulting from a data preparation step. In the next step, define a vector w, where w_{d_j} is the weight of the j^{th} dimension of $\mathcal{D}$. The weight w_{d_j} represents the importance of the variable for the analysis as perceived by the business expert. A higher value of w_{d_j} implies, assuming other weights are fixed, a higher importance for the analysis. Finally, apply any clustering technique that uses the Euclidian distance as similarity or dissimilarity measure and replace the traditional Euclidian distance calculation by the following adapted distance between an input vector n_i and another input vector n_m:

$$dist(n_i, n_m) = \sqrt{\sum_{j=1}^{d} w_{d_j}(n_{id_j} - n_{md_j})^2}, \tag{3.1}$$

with w_{d_j} representing the importance level of the j^{th} dimension of $\mathcal{D}$ and n_{id_j} and n_{md_j} the values of the j^{th} dimension of the input vectors n_i and n_m respectively.

The idea behind Equation 3.1 and the proposed approach can be illustrated using the two situations shown in Figure 3.1. Consider the 4 points represented in Figure 3.1.1 which are the corners of a square with a unit length. It is given that the x-axis represents a binary variable in the range $[0, 1]$ while the y-axis represents a continuous variable in the range [0..1]. Applying the k-means algorithm presented in Section 2.2.2.2 with two clusters and a random seed initialization of the centroids, 6 unique partitions can be obtained using the 4 points A, B, C and D of Figure 3.1.1 as can be seen in Table 3.2.1. The well known Davies-Bouldin index ([23]) is calculated as follows:

$$index = \frac{1}{c}\sum_{i=1}^{c}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right), \tag{3.2}$$

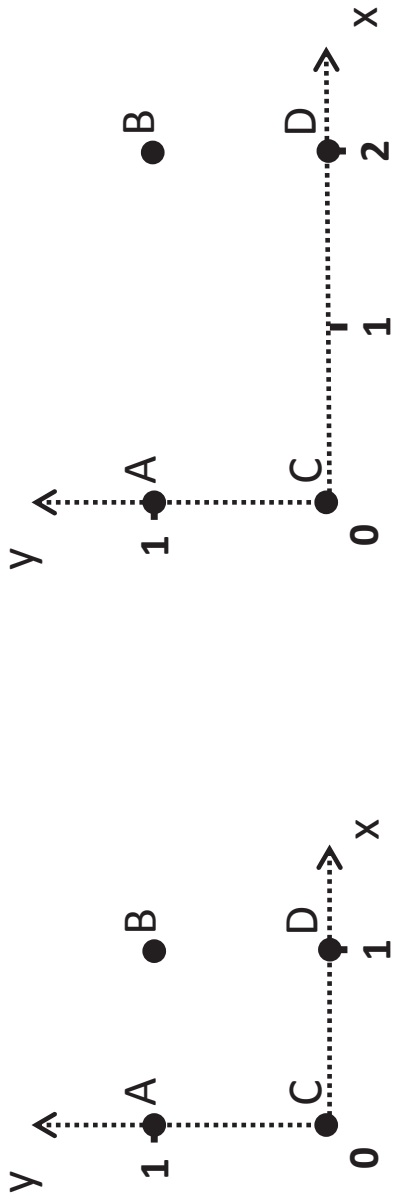
where c is the number of clusters, c_y is the centroid of cluster y, σ_y is the average distance of all elements of cluster y and $d(c_i, c_j)$ is the distance between the centroid of cluster i and cluster j. Since a good partitioning corresponds to a situation where the intra-cluster distances are low and the inter-cluster distances are high, the lower the Davies-Bouldin index, the better the obtained partitioning. Using this index as evaluation metric for the 6 partitions previously obtained, the two best partitions are partitions 1 and 2, as their clusters are dense and far from each other. Of the 2 pairs of centroids representing the two best partitions, only one is meaningful considering the knowledge about the variables represented by the x- and y-axis. Indeed, only a partition grouping the point A with C and the point B with D leads to centroids having coordinates respecting the ranges of the original variables. The idea of the proposed approach is to incorporate the business knowledge into the algorithm in order to guide it to a solution which will be potentially better. One way to do it in the context presented in Figure 3.1.1 is to transform the knowledge in priorities such that a variable with higher priority should be more structural in the resulting partition. Indeed, the variable x, a binary variable, can

be assigned a higher priority than the variable y, a continuous variable, since the algorithm should create partitions mainly structured by the binary variable, hence leading to more meaningful results. This can be achieved by artificially increasing the scale of the variable x as can be seen in Figure 3.1.2. In this context, applying the k-means algorithm with two clusters and a random seed initialization of the centroids will lead to 2 unique partitions. Taking the best partition using the Davies-Bouldin index approach, the partition grouping the point A with C and the point B with D is obtained and leads to meaningful centroids given the available knowledge.

Coming back to Equation 3.1, a dissimilarity between two input vectors will be proportional to the weights and the ranges of the dissimilar attributes so that it becomes easy to include business knowledge by tuning the Euclidian distance with higher or smaller weights. The next section will discuss an approach to fix those weights when the goal is to perform unsupervised learning with an a priori knowledge about the importance of the variables. The approach leads to clustering techniques able to generate clusters mainly structured by the variables with a higher priority for the analyst, which offers new perspectives into the data set at hand.

3.2.2 From priorities to weights

In this section, a general approach for fixing the weights of Equation 3.1 is presented. The purpose of this is to obtain clustering techniques able to generate clusters mainly structured by the variables with a higher priority for the analyst by using Equation 3.1 with the weights resulting from this approach instead of the classical Euclidian distance. Moreover, a specific case of this approach is proposed in a setup involving only categorized variables represented by dummies, leading to a straightforward method which is used in Section 3.4 when training Self-Organizing maps.



3.1.1: Two variables with the same scale.

3.1.2: Two variables with different scales.

Figure 3.1: Two examples with different scales for the dimension x.

Partition	Seeds	c_1	c_2	coord c_1	coord c_2
1	**A, B**	**A, C**	**B, D**	**(0, 0.5)**	**(1, 0.5)**
2	**A, C**	**A, B**	**C, D**	**(0.5, 1)**	**(0.5, 0)**
3	A, D	A, B	C, D	(0.5, 1)	(0.5, 0)
4	A, D	A, C	B, D	(0, 0.5)	(1, 0.5)
5	**A, D**	**A, B, C**	**D**	**(1/3, 2/3)**	**(1, 0)**
6	**A, D**	**A**	**B, C, D**	**(0, 1)**	**(2/3, 1/3)**
7	B, C	A, B	C, D	(0.5, 1)	(0.5, 0)
8	B, C	B, D	A, C	(0, 0.5)	(0.5, 0)
9	**B, C**	**A, B, D**	**C**	**(2/3, 2/3)**	**(0, 0)**
10	**B, C**	**B**	**A, C, D**	**(1, 1)**	**(1/3, 1/3)**
11	B, D	A, B	C, D	(0.5, 1)	(0.5, 0)
12	C, D	A, C	B, D	(0, 0.5)	(1, 0.5)

3.2.1: Partitions obtained using the points of Figure 3.1.1

Partition	Seeds	c_1	c_2	coord c_1	coord c_2
1	**A, B**	**A, C**	**B, D**	**(0, 0.5)**	**(2, 0.5)**
2	**A, C**	**A, B**	**C, D**	**(1, 1)**	**(1, 0)**
3	A, D	A, C	B, D	(0, 0.5)	(2, 0.5)
4	B, C	A, C	B, D	(0, 0.5)	(2, 0.5)
5	B, D	A, B	C, D	(1, 1)	(1, 0)
6	C, D	A, C	B, D	(0, 0.5)	(2, 0.5)

3.2.2: Partitions obtained using the points of Figure 3.1.2

Figure 3.2: Tables showing the different partitions that can be obtained using the k-means algorithm with 2 clusters and random seeds initialization using as input the 4 points of Figure 3.1.1 and Figure 3.1.2, respectively. Each line of the tables represents a partitioning of the 4 points. The first column of the tables, *Seeds*, represents the two points used as initial centroids. The second column, c_1, and the third column, c_2, represent, respectively, the points associated to the first and the second cluster after the k-means algorithm has been applied. The fourth and fifth column represent the coordinates in the (x, y) space of the centroids of c_1 and c_2, respectively. The lines represented in bold are the unique partitions. The way to read the first line of the Table 3.2.1 is as follows: applying the k-means algorithm with two clusters on the 4 points of Figure 3.1.1 using the points A and B as initial centroids, one output of the algorithm can be the two clusters c_1 and c_2. The points A and C are related to c_1 and the points B and D are related to c_2. The coordinates of the centroid of c_1 are (0, 0.5) and the coordinates of the centroid of c_2 are (1, 0.5).

3.2.2.1 General approach

The general approach requires the definition of d quantitative variables d_j from $\mathcal{D}$ showing some variability and a priority vector p of size d that captures the priorities assigned to the different dimensions of $\mathcal{D}$, with p_{d_j} representing the priority of the j^{th} dimension of $\mathcal{D}$. A dimension d_j with $p_{d_j} = 1$ is considered as a dimension with the highest priority and has then a higher priority than a dimension d_o with $p_{d_o} = 2$, etc. In order to derive the weights w_{d_j}, different definitions are necessary and proposed in what follows. The contribution Δ^{d_j} of a dimension d_j to the Euclidian distance between two vectors n_i and n_m,

$$dist(n_i, n_m) = \sqrt{\sum_{j=1}^{d} (n_{id_j} - n_{md_j})^2}, \tag{3.3}$$

is defined as:

$$\Delta^{d_j}_{(n_i,n_m)} = (n_{id_j} - n_{md_j})^2. \tag{3.4}$$

The maximal contribution of a variable d_j, $\Delta^{d_j}_{max}$, is defined as:

$$\Delta^{d_j}_{max} = \max_{n_i,n_m \in \mathcal{N}} \Delta^{d_j}_{(n_i,n_m)}. \tag{3.5}$$

The maximal Euclidian distance $dist_{max}$ is defined as:

$$dist_{max} = \sqrt{\sum_{j=1}^{d} \Delta^{d_j}_{max}}. \tag{3.6}$$

The weighted maximal contribution ψ_{d_j} is defined as:

$$\psi_{d_j} = w_{d_j} \Delta^{d_j}_{max}. \tag{3.7}$$

The maximal weighted Euclidian distance $dist^w_{max}$ is defined as:

$$dist^w_{max} = \sqrt{\sum_{j=1}^{d} w_{d_j} \Delta^{d_j}_{max}} = \sqrt{\sum_{j=1}^{d} \psi_{d_j}}. \tag{3.8}$$

Given those definitions, the goal of this approach is to find the variables' weights such that the weighted maximal contribution of a variable is proportional to the sum of the weighted maximal contributions of the variables with lower priorities. In order to formalize this idea, a set $\mathcal{LP}^{d_j}$ is created which contains the dimensions d_i with lower priorities than d_j:

$$d_i \in \mathcal{LP}^{d_j}, \forall d_i : p_{d_i} > p_{d_j}, \tag{3.9}$$

and the weights w_{d_j} have to satisfy the following inequality:

$$\alpha\psi_{d_j} > \sum_{d_i \in \mathcal{LP}^{d_j}} \psi_{d_i} \tag{3.10}$$

with α being a positive parameter inversely proportional to the desired intensity of the prioritization. This parameter is used as a strength factor influencing the magnitude of the constraints like the α in [44], the τ_{ij} in [45], the λ_1 and λ_2 in [43] or the w in [35].

3.2.2.2 Specific approach

A specific case of this approach is proposed in a setup involving only categorized variables represented by dummies, leading to a straightforward method applied in Section 3.4. In order to meet the requirements of this specific context, a qualitative variable with t different values should be transformed into t dummy variables. A value of 0 or 1 reflects whether or not the input vector is characterized by the value represented by the dummy variable, so that only one of the t dummy variables can be equal to 1. Concerning the quantitative variables, a categorization is possible using intervals represented by dummy variables as done for qualitative variables. Only one of the dummies obtained by the categorization can be equal to 1 if the intervals are not overlapping. Once this data preparation is performed, define a set $\mathcal{G}$ of g non overlapping groups g_k of dimensions of $\mathcal{D}$ so that there is no dimension of $\mathcal{D}$ not belonging to one group of $\mathcal{G}$ and no dimension of $\mathcal{D}$ belonging to two different groups of $\mathcal{G}$. The function $g(d_j)$ returns

the group g_k such that $d_j \in g_k$. The notion of group is introduced in order to capture the fact that only one of the dummies obtained by the categorization of a variable can be equal to 1 if the categories are not overlapping. This information is used in what follows in order to fix the weights assigned to the different variables. An example of such a group could be the different dummies resulting from the categorization of a quantitative variable. For example, consider the initial variable *Age*, where the categories, such as [18..25], [26..35], [36..50], [51..65], and [66..], are the dimensions of the group represented by dummies. In order to complete the definition of a group, which is a real-world subdivision of the dimensions, the same priority should be assigned to all the dimensions of the same group, so that

$$\forall d_j \in g_k, p_{g_k} = p_{d_j}, \tag{3.11}$$

with g_k representing the k^{th} group of $\mathcal{G}$ and p_{g_k} the priority assigned to it. Completing this context, a set $\mathcal{LP}^{g_k}$ is defined and gathers the groups g_l such that:

$$g_l \in \mathcal{LP}^{g_k}, \forall g_l : p_{g_l} > p_{g_k}. \tag{3.12}$$

Finally, given this context and the definitions of Section 3.2.2.1, the weights w_{d_j} can be obtained using the following equation:

$$w_{d_j} = 1 + \frac{|\mathcal{LP}^{g(d_j)}|}{\alpha}, \tag{3.13}$$

with $|\mathcal{LP}^{g(d_j)}|$ being the number of groups having a lower priority than the group of the variable d_j. This equation provides a solution which, given the requirements of the proposed specific context are fulfilled, respects the inequality of Equation 3.10. The way to calculate the weights used in Equation 3.1 in the proposed context is summarized in Algorithm 3.1.

Algorithm 3.1 Fixing the weights

1: Given a set of variables $\mathcal{P}$.
2: Define a set of d dummy variables $\mathcal{D}$ by transforming the variables of $\mathcal{P}$.
3: Define a set of g groups $\mathcal{G}$ partitioning the variables of $\mathcal{D}$.
4: Define a vector p of size d by assigning priorities to the different variables of $\mathcal{D}$.
5: **for** $k = 1 \rightarrow g$ **do**
6: Define a set $\mathcal{LP}^{g_k}$ such that $g_l \in \mathcal{LP}^{g_k}, \forall g_l : p_{g_l} > p_{g_k}$.
7: **end for**
8: Fix the value of the parameter α.
9: Define a vector w with w_{d_j} representing the weight of the j^{th} dimension of $\mathcal{D}$.
10: **for** $j = 1 \rightarrow d$ **do**
11: $w_{d_j} = 1 + \frac{|\mathcal{LP}^{g(d_j)}|}{\alpha}$
12: **end for**
13: **return** w.

3.3 Methodology implementing the prioritization approach

In this section, a 5-step methodology implementing the prioritization approach proposed in Section 3.2 is described assuming the specific context of Section 3.2.2.2. This methodology is further applied in Section 3.4 and compared with a methodology which is not incorporating the available business knowledge. Figure 3.3 shows the five steps which are explained in the next sub-sections.

3.3.1 Data preparation

The first step of the methodology is a data preparation step required to meet the specific context necessary to calculate the weights as proposed in Section 3.2.2.2 with Algorithm 3.1. The different variables are transformed into dummies and groups are defined based on this

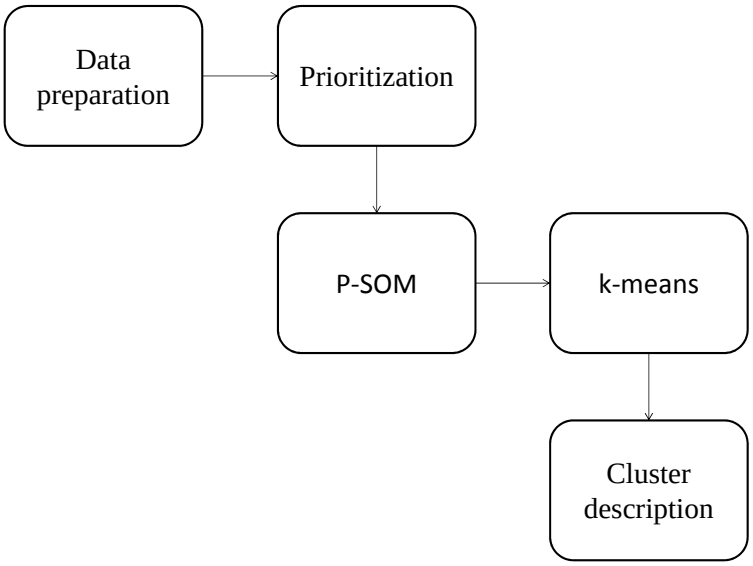


Figure 3.3: Figure showing the 5-step methodology.

transformation as discussed in Section 3.2.2.2, leading to a set $\mathcal{D}$ of d dummies and a set $\mathcal{G}$ of g non-overlapping groups.

3.3.2 Prioritization

In the second step, priorities are assigned to the different variables as explained in Section 3.2.2.1. The parameter α is fixed and the weights of the different variables are calculated using Algorithm 3.1 leading to a vector of weights that can be used in the next step of the methodology. Important to note is the fact that the priorities are an input from the analyst and reflect the subjective importance of the variables. The parameter α is to be fixed and will impose the magnitude of the prioritization. The lower the value of it, the higher are the weights and the higher is the difference of impact between variables having different priorities.

3.3.3 Prioritized Self-Organizing Maps

In the third step, an adapted version of the SOM algorithm presented in Section 2.2.2.1 is proposed in order to implement the prioritization approach by using Equation 3.1 during the identification of the BMUs with the weights obtained in step 2. The Equation 2.1 is hence replaced by the following equation:

$$\sqrt{\sum_{j=1}^{d} w_{d_j}(n_{id_j} - m_{cd_j})^2} = \min_r\{\sqrt{\sum_{j=1}^{d} w_{d_j}(n_{id_j} - m_{rd_j})^2}\}, \quad (3.14)$$

with w_{d_j} the weight of the j^{th} dimension, n_{id_j} the j^{th} dimension of input vector n_i and m_{cd_j} the j^{th} dimension of the neuron m_c. This new approach for identifying the BMUs leads to a new algorithm, Prioritized SOM (P-SOM), which is summarized in Algorithm 3.2 and motivated in what follows. By applying the classical SOM algorithm in a setup involving only dummy variables, the obtained neurons will have values in the range [0..1]. Some variables may be difficult to interpret if the values of the neurons are close to 0.5. This situation can happen when those variables do not structure the data. While it is not a problem for a blind exploration, it is not desirable when an analyst is aware of an a priori prioritization among the variables. Indeed, if a variable is perceived as important by the analyst, the algorithm should be able to incorporate this knowledge and guide the learning process in order to create a relative structure given the existing knowledge. By applying Algorithm 3.2, this knowledge is incorporated in the learning process, leading to neurons mainly structured by the variables with higher priorities, hence improving the perceived quality of the resulting maps. Indeed, the same phenomenon as the one outlined in Section 3.2.1 with the k-means algorithm will lead to neurons with extremer values for the variables with higher priorities.

3.3.4 k-means

In the fourth step, the k-means algorithm is applied to the neurons trained in the previous step. Although the P-SOM algorithm offers

Algorithm 3.2 P-SOM

1: Given a set of input vectors n_i and a set of output neurons m_r.
2: Fix the weights w_{d_j} of the different dimensions of the input vectors.
3: Initialize the neurons.
4: **repeat**
5: Select an input vector n_i.
6: Find the BMU m_c: $\sqrt{\sum_{j=1}^{d} w_{d_j}(n_{id_j} - m_{cd_j})^2} = \min_r\{\sqrt{\sum_{j=1}^{d} w_{d_j}(n_{id_j} - m_{rd_j})^2}\}$.
7: Update the nodes weights: $m_r(t+1) = m_r(t)+\alpha(t)h_{cr}(t)[n_i(t)-m_r(t)]$.
8: **until** Stopping criterion is satisfied

advanced visualization facilities by reducing the dimensionality to two-dimensional maps, a formal analysis of the resulting neurons offers advanced insights into the structure of the data (see e.g. [22] and [48]). The output of this step is a partitioning of the neurons into groups sharing some characteristics. Note that this clustering technique will capture the prioritized structure.

3.3.5 Cluster description

In the final step, the obtained clusters are described by extracting their salient dimensions as explained in Sections 2.2.2 and 2.3. Thanks to the characterization of the clusters, it is possible for the analyst to name and assess the clustering of the neurons. Note that variables with higher priorities should be at the origin of the clusters' structures such that the salient character of a variable should be impacted by its priority.

3.4 Application

The application involves the clustering of a data set originating from the concert industry which is provided by Ticketmatic, a leading ticketing company in Europe. The data set gathers attributes about clients of one concert organizer which are summarized in Table 3.1 and discussed in what follows. Different business considerations motivated the use of the methodology proposed in the previous section. Firstly, the business involved wanted to use a technique allowing them to use their knowledge to guide the clustering algorithm and analyze different perspectives of the data. Some a-priori knowledge about important variables was available and could thus be transformed into priorities. Secondly, feature selection algorithms were rejected by the business because all input variables were considered valuable from a business perspective, discarding a purely statistical unsupervised feature selection strategy. Thirdly, since the analysts were marketing experts, visualization for exploration and reporting was considered as a key feature of the to-be-used technique. These different business considerations were guiding for the design and the application of the proposed methodology. In Section 3.4.1, a comparison between a traditional two-step clustering approach and the prioritized methodology proposed in Section 3.3 is illustrated, using the available data. To do so, a real business context is described and a solution to a relevant decision problem is proposed using both approaches. In Section 3.4.2, the impact of the parameter α of Equation 3.13 on the clustering quality and on the relative importance of the variables in the resulting clustering structure is discussed, based on the results of extensive experiments.

3.4.1 Two approaches

The goal of the application is to perform an analysis of the data at hand in order to capture the profile of the customers that may be interested in a future concert, called *The Concert* in the remainder of this chapter. The different steps of the application are shown in Figure 3.4 and are discussed in what follows. Note that steps 1, 4 and

Dimension	Index category	Original variable	Index name	Value	Group
d_1	Demographic	Gender	Gender Man	$\{0,1\}$	g_1
d_2	Demographic	Gender	Gender Woman	$\{0,1\}$	g_1
d_3	Demographic	Age	Age 18-25	$\{0,1\}$	g_2
d_4	Demographic	Age	Age 25-35	$\{0,1\}$	g_2
d_5	Demographic	Age	Age 35-50	$\{0,1\}$	g_2
d_6	Demographic	Age	Age 50-56	$\{0,1\}$	g_2
d_7	Demographic	Age	Age 65-more	$\{0,1\}$	g_2
d_8	Demographic	Distance	Distance 0-5	$\{0,1\}$	g_3
d_9	Demographic	Distance	Distance 5-10	$\{0,1\}$	g_3
d_{10}	Demographic	Distance	Distance 10-15	$\{0,1\}$	g_3
d_{11}	Demographic	Distance	Distance 15-25	$\{0,1\}$	g_3
d_{12}	Demographic	Distance	Distance 25-50	$\{0,1\}$	g_3
d_{13}	Demographic	Distance	Distance 50+	$\{0,1\}$	g_3
d_{14}	RFM	Total rfm	Total rfm 1	$\{0,1\}$	g_4
d_{15}	RFM	Total rfm	Total rfm 2	$\{0,1\}$	g_4
d_{16}	RFM	Total rfm	Total rfm 3	$\{0,1\}$	g_4
d_{17}	RFM	Total rfm	Total rfm 4	$\{0,1\}$	g_4
d_{18}	RFM	Total rfm	Total rfm 5	$\{0,1\}$	g_4
d_{19}	Interest-based	The Concert	The Concert 1	$\{0,1\}$	g_5
d_{20}	Interest-based	The Concert	The Concert 2	$\{0,1\}$	g_5
d_{21}	Interest-based	The Concert	The Concert 3	$\{0,1\}$	g_5
d_{22}	Interest-based	The Concert	The Concert 4	$\{0,1\}$	g_5
d_{23}	Interest-based	The Concert	The Concert 5	$\{0,1\}$	g_5

Table 3.1: Summary of the variables used in the application.

5 of Figure 3.4 are similar for both approaches in order to allow a comparison of the results.

The first step of the application concerns the data preparation according to the criteria presented in Section 3.2.2.2. An index *total RFM* is computed by summing the three values of the RFM variables (Recency, Frequency and Monetary) calculated for each client, leading to a score between 3 and 15 for each of them. The interested reader is referred to [49] for an extensive discussion of the RFM framework, used in this chapter to assess the value of a customer to the company based on his past behavior. Five categories of customers were defined based on their respective *total RFM* indices, using the following intervals: [3..5], [6..8], [9..11], [12..13] and [14..15]. The birthdate of the different customers was used to generate an index capturing the age. Five categories were defined using the following intervals in years: [18..25], [26..35], [36..50], [51..65] and [66..]. The gender of the customers was used to build two extra categories. Using the IP addresses of the booking computers, an index representing the geographic distance separating the customer from the concert location was created. Categories were defined using the following intervals in km: [0..5], [6..10], [11..15], [16..25], [26..50] and [51..]. In order to define an index capturing the interest of a given customer for a tag, an artist or a concert, the *last.fm* API data was used in combination with data of the previous concerts involving the given customer. Considering that a concert consists of a series of artists characterized by a series of tags, it is then possible to rank the customers according to their score for a given concert as described in [48]. Note that this interest-based variable is based solely on tags of previously attended concerts, and not on information about the customers who really attended *The Concert*. Five categories are then obtained by categorizing this interest variable, from *The Concert 1*, representing customers with a low interest, to *The Concert 5*, representing those with a high interest.

As a second step of the application, the specific prioritization approach presented in Section 3.2.2.2 is performed. Groups are formed as follows: $g_1 = (d_1, d_2)$, $g_2 = (d_3, d_4, d_5, d_6, d_7)$, $g_3 = (d_8, d_9, d_{10}, d_{11}, d_{12}, d_{13})$, $g_4 = (d_{14}, d_{15}, d_{16}, d_{17}, d_{18})$ and $g_5 =$

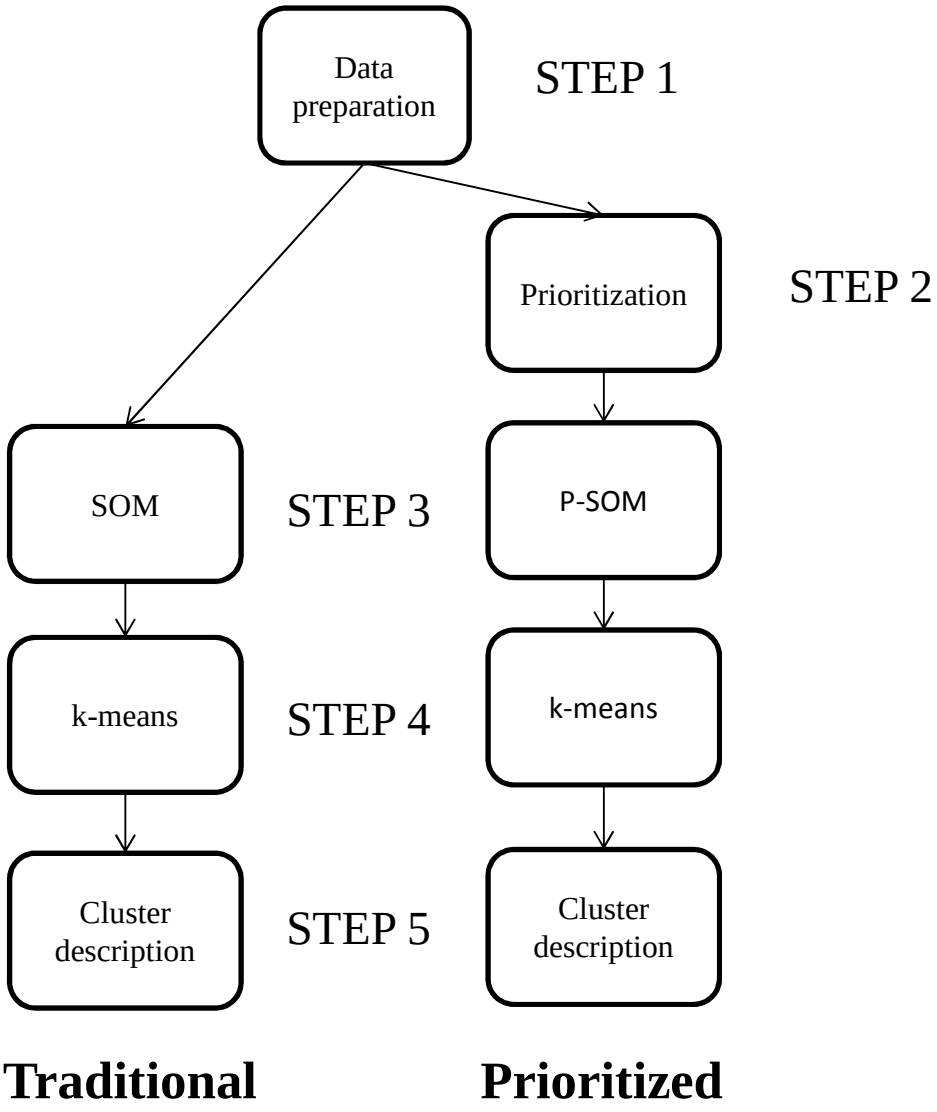


Figure 3.4: Representation of the 5 steps of the application and the steps providing the practitioner with interesting analysis.

$(d_{19}, d_{20}, d_{21}, d_{22}, d_{23})$. Note that these groups gather the dummies related to the same original variables. A priority is assigned to the different variables respecting the definitions of Section 3.2.2. In this case, the business is interested in analyzing the profiles of the potential participants of future concerts and considers the five variables related to the interest for *The Concert* as more important than the other variables. This business knowledge is difficult to integrate in the clustering task with traditional clustering techniques, but the approach proposed allows for an easy translation in terms of priorities and weights. Using the logic presented in Section 3.2.2, a priority of 1 is assigned to the variables *The Concert 1*, *The Concert 2*, *The Concert 3*, *The Concert 4* and *The Concert 5* and a priority of 2 is assigned to the other 18 variables, meaning that they are considered as less important than the variables capturing the interest for *The Concert*. The weights are then calculated using Equation 3.13 with α equal to 0.2 resulting in a weight equal to 21 for the variables of group 5 and a weight equal to 1 for the other variables. The details of the calculation of the weights of the dimensions d_1 and d_{19} are illustrated in what follows.

1. The dimension d_1 has been assigned a priority of 2 and is part of group 1 so that $p_{d_1} = 2$ and $g(d_1) = g_1$. $\mathcal{LP}^{g(d_1)}$ is empty because 2 is the lowest priority in this case such that $\mathcal{LP}^{g(d_1)} = \emptyset$. Finally, with α equal to 0.2, the weight of the dimension d_1, w_{d_1}, is calculated as $w_{d_1} = 1 + \frac{|\mathcal{LP}^{g(d_1)}|}{\alpha} = 1 + \frac{0}{0.2} = 1$.

2. The dimension d_{19} has been assigned a priority of 1 and is part of group 5 such that $p_{d_{19}} = 1$ and $g(d_{19}) = g_5$. $\mathcal{LP}^{g(d_{19})}$ is equal to (g_1, g_2, g_3, g_4). Finally, with α equal to 0.2, the weight of the dimension d_{19}, $w_{d_{19}}$, is calculated as $w_{d_{19}} = 1 + \frac{|\mathcal{LP}^{g(d_{19})}|}{\alpha} = 1 + \frac{4}{0.2} = 21$.

For the next steps of the application, a distinction will be made between the traditional and the prioritized approach. The traditional approach consists of applying the SOM-algorithm introduced in Section 2.2.2.1 whereas the prioritized approach includes the available

business knowledge with the P-SOM algorithm as described in Algorithm 3.3.3.

In the third step, a 10×12 SOM is trained (batch training) using both traditional and prioritized approaches. By doing this, a large set of prototypes is created. According to best practices ([22] and [17]), this number should be larger than the expected number of clusters. This expected number being unknown, business expectations are then used to have an idea of the maximum number of clusters they are willing to handle. In this case, because 10 clusters are expected and the first clustering step is followed by a second one, it has been opted to choose a number substantially higher than 10. According to the same best practices ([17]), a rectangular shape of the output map is preferred, which explains the choice of a 10×12 SOM instead of e.g. a 10×10 SOM. The visualization power of the SOM is illustrated in Figures 3.5 and 3.6, which show the component planes for the traditional approach and for the prioritized approach respectively. The dark red and dark blue neurons of a component plane represent, respectively, the neurons with relatively high and low values for the variable represented by the component plane. Figure 3.5 gives the analyst the possibility to analyze the data at hand in a general way and provides him with valuable patterns. An example of such a pattern is the fact that, based on Figures 3.5.14 and 3.5.13, the analyst can see a correlation between the variables *Distance 50+* and *Total rfm 1* which is a valuable pattern indicating the fact that people coming from far are not perceived as clients with a high customer lifetime value. This is a logic pattern that may convince the analyst to spend more effort on people located near to the concert place. Although this pattern is interesting and the fact that an expert in SOM would be able to find a lot of other patterns using these component planes, the limits of the unsupervised clustering task appear if an analyst is interested in some of the variables and would like to focus on them. Some advanced techniques, such as e.g. the one proposed in [48], can deal with this problem by using statistical approaches, which gives the possibility to explore the data by specifying targeted dimensions. However, this goal can also be achieved by using the proposed prioritized approach which

overcomes those limitations by providing the analyst with a way to focus on specific variables by defining priorities for the analysis. Using Figure 3.6, which focuses on the variables reflecting the interest of the clients for a future concert, the analyst is able to perform a specific exploration of the data, formalizing his knowledge and effectively guiding the algorithm. Richer and cleaner patterns appear when focusing on the variables of group 5 as can be concluded from Figure 3.6. An example of a more specific analysis can be the analysis of the profile of the clients having the most interest in the future concert. Those clients are summarized by the neurons having a high value for the variable *The Concert 5*, and, based on the other sub-figures of Figure 3.6, one can easily conclude that those clients are relatively young people (see Figures 3.6.9, 3.6.12 and 3.6.13), have a low RFM value (see Figures 3.6.14, 3.6.15 , 3.6.16, 3.6.17and 3.6.18) and are used to travel to attend a concert (see Figures 3.6.3, 3.6.6, 3.6.7 and 3.6.8). Given that, in this context, the analyst is interested in the understanding of the profiles capturing the interest for *The Concert*, the reader can easily assess the usefulness of both approaches.

In the fourth step of the experiment, a second clustering step is performed in order to capture the structure of the SOMs obtained using both traditional and prioritized approaches. The k-means algorithm is thus applied twice: a first time to the neurons obtained by the traditional approach and a second time to the neurons obtained by the prioritized approach. The maximal number of clusters is fixed to 10 and the number of epochs to 50 so that a stable clustering is obtained. The Davies-Bouldin index is used in order to select the best clustering for each approach. Figure 3.7.1 shows the 9 clusters $c_k : k \in [1..9]$ obtained when applying the k-means to the neurons resulting from the traditional approach whereas Figure 3.7.2 shows the 8 clusters $c_k : k \in [1..8]$ obtained when the prioritized approach is used.

Finally, the extraction of the salient dimensions of the different clusters obtained in the previous step is performed with the parameters z_1 and z_2 respectively equal to 10000 and 0 so that the sensitivity is maximal ([48]). The characteristics of those clusters can thus be analyzed and are summarized in Tables 3.2 and 3.3. Both tables show

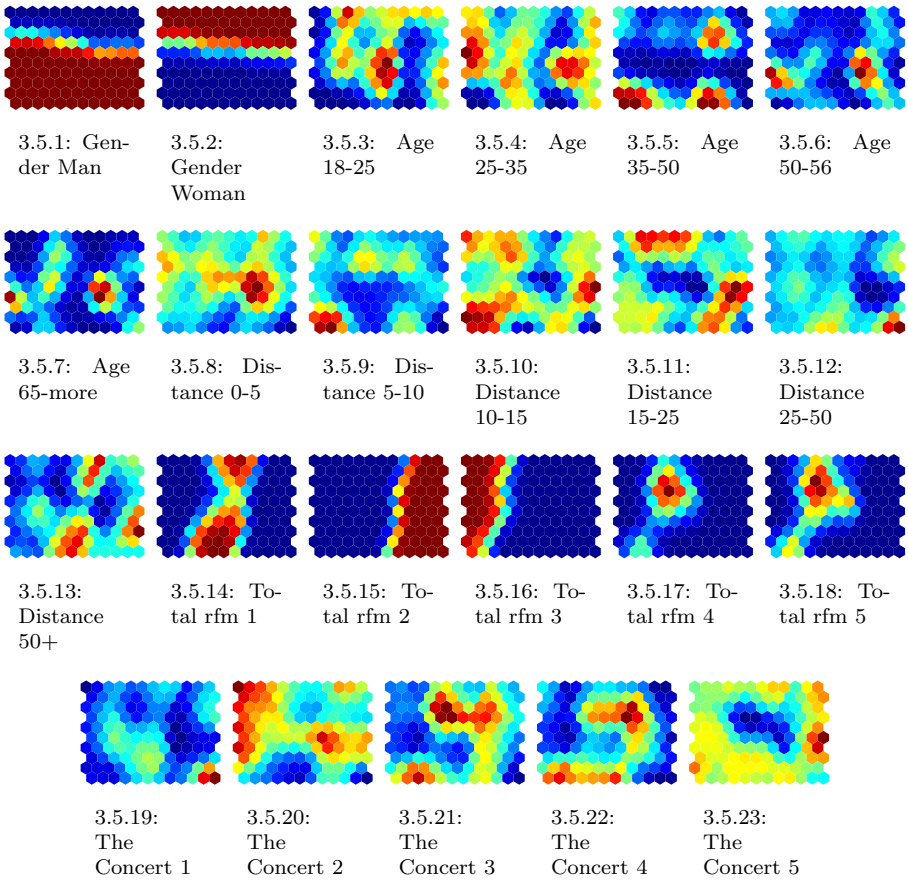


Figure 3.5: SOM output obtained without using a prioritization of the variables.

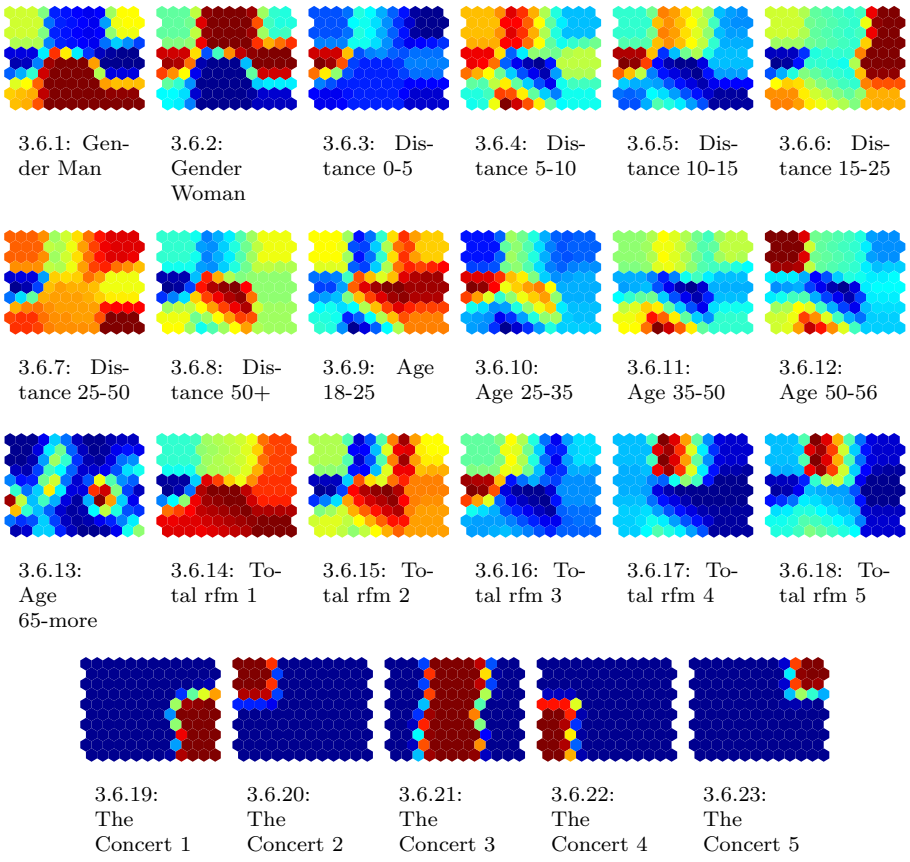
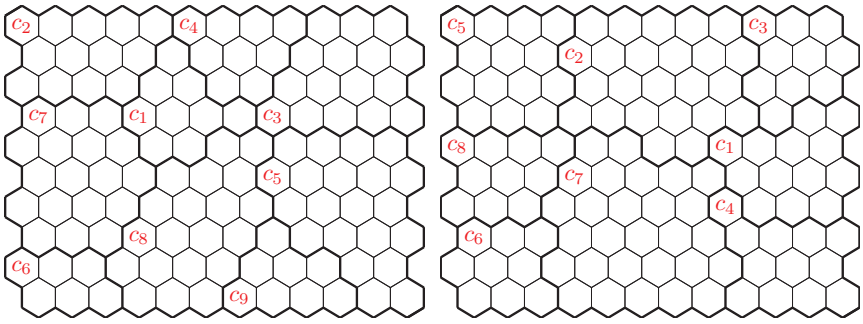


Figure 3.6: SOM output obtained by using a prioritization of the variables by giving more importance to the variables *The Concert 1*, *The Concert 2*, *The Concert 3*, *The Concert 4* and *The Concert 5*.



3.7.1: Traditional approach. 3.7.2: Prioritized approach.

Figure 3.7: Clustering resulting from the application of the k-means to the neurons generated using the traditional approach and the prioritized approach respectively.

the dimensions characterizing the different clusters in descending order of their salientness using the difference factors calculated with Equation 2.8. For example, based on Table 3.2, it can be said that the third cluster obtained with the traditional approach, c_3, has as main characteristic the dimension d_{15}, as second main characteristic the dimension d_2, etc. This means that the centroid of cluster c_3 represents customers characterized by a low RFM value ($TotalRFM2$), a female gender ($GenderWoman$), etc. Although this information is interesting in order to explore the data without any specific goal, the clustering task often hides pre-defined objectives that are difficult to satisfy using a classical unsupervised approach. In this case, such an objective consists of the understanding of the customers' interest in *The Concert*, which is quite difficult given the results obtained and shown in Table 3.2. This goal can be achieved by exploring the data while focusing on the variables of group 5. The reader can see the position of those dimensions, namely the dimensions d_{19}, d_{20}, d_{21}, d_{22} and d_{23}, represented in bold in Tables 3.2 and 3.3. It can be seen that this final step has led to 9 clusters offering a general view on the one hand, and to 8 clusters offering a targeted view on the other hand. Given the precise task the business is involved in, the first partition-

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8	c_9
d_{17}	d_{16}	d_{15}	d_{14}	d_{15}	d_5	d_{16}	d_{14}	d_5
d_{18}	d_2	d_2	d_2	d_7	d_{16}	d_7	d_1	d_{15}
d_2	$\mathbf{d_{20}}$	d_5	d_3	d_6	d_{18}	d_1	d_3	d_1
d_7	d_{10}	$\mathbf{d_{21}}$	d_{13}	d_1	d_{17}	d_6	$\mathbf{d_{19}}$	d_{13}
$\mathbf{d_{21}}$	d_{18}	d_8	d_{18}	d_4	d_9	d_4	d_{13}	d_{11}
d_9	d_{11}	$\mathbf{d_{23}}$	d_{11}	d_8	d_1	$\mathbf{d_{20}}$	d_{18}	d_9
d_8	d_3	d_{13}	d_{17}	$\mathbf{d_{20}}$	d_{10}	$\mathbf{d_{23}}$	d_{12}	d_{12}
$\mathbf{d_{22}}$	d_4	d_9	$\mathbf{d_{22}}$	d_{11}	$\mathbf{d_{22}}$	$\mathbf{d_{19}}$	d_{17}	$\mathbf{d_{23}}$
d_4	d_9	$\mathbf{d_{22}}$	$\mathbf{d_{21}}$	d_{13}	$\mathbf{d_{21}}$	d_{12}	$\mathbf{d_{22}}$	$\mathbf{d_{21}}$
d_3	d_{12}	d_6	$\mathbf{d_{23}}$	$\mathbf{d_{23}}$	$\mathbf{d_{23}}$	d_{10}	d_4	$\mathbf{d_{22}}$

Table 3.2: Dimensions characterizing the 9 clusters obtained with the traditional approach ranked in decreasing order of their difference factors. The dimensions in bold are the dimensions considered as more important by the business.

ing is difficult to use while the second is really suitable. It can indeed easily be concluded, based on Table 3.3, that the customers potentially interested in the future concert (summarized by the cluster c_3), are coming from far, are relatively young people and are perceived as clients with a low value for the company. This might be an incentive for the concert organizer to think about other music groups, assuming that his main objective is to maximize his profit, until he finds a suitable artist which matches his profitable customer base's interest.

3.4.2 Impact of the parameter α

In this section, the impact of the parameter α is discussed and illustrated using the prioritized setup described in Section 3.4. This parameter is used as a strength factor inversely proportional to the amplitude of the desired prioritization. The smaller the value of α, the higher the difference of weights between variables having different

c_1	c_2	c_3	c_4	c_5	c_6	c_7	c_8
$\mathbf{d_{19}}$	$\mathbf{d_{21}}$	$\mathbf{d_{23}}$	$\mathbf{d_{19}}$	$\mathbf{d_{20}}$	$\mathbf{d_{22}}$	$\mathbf{d_{21}}$	$\mathbf{d_{22}}$
d_2	d_{18}	d_{12}	d_1	d_6	d_5	d_1	d_{16}
d_3	d_{17}	d_{11}	d_7	d_7	d_1	d_{13}	d_8
d_{15}	d_2	d_5	d_{14}	d_{16}	d_{18}	d_{14}	d_2
d_{11}	d_{10}	d_{13}	d_{12}	d_9	d_7	d_{15}	d_{10}
d_{14}	d_5	d_{14}	d_3	d_5	d_6	d_{18}	d_4
d_9	d_9	d_3	d_{15}	d_{17}	d_{14}	d_4	d_9
d_{13}	d_{16}	d_1	d_{11}	d_{18}	d_{17}	d_7	d_{17}
d_8	d_7	d_{15}	d_{13}	d_{12}	d_{12}	d_3	d_6
d_{10}	d_8	d_{16}	d_4	d_{10}	d_{13}	d_{12}	d_5

Table 3.3: Dimensions characterizing the 8 clusters obtained with the prioritized approach ranked in decreasing order of their difference factors. The dimensions in bold are the dimensions considered as more important by the business.

priorities. By reducing this parameter, the soft constraints used during the clustering algorithm are strengthened, hence leading to partitions mainly structured by the variables of higher priorities while changing the resulting partitioning of the data points. By artificially constraining the clustering algorithm, it is expected that the subjective and the objective quality of the obtained partitioning will be impacted. In the remainder of this chapter, two metrics will be used to measure this. For the first measure, the Davies-Bouldin index is chosen to capture the objective, data-driven, quality of the clustering. The lower the value of this index, the better the separation of the clusters and the within-cluster density. The second measure, the subjective measure, captures the relative importance of the variables of a given priority with regards to the variables with a lower priority. The relative importance of a set

of variables with priority p_i, ϕ_{p_i} is calculated as

$$\phi_{p_i} = \frac{1}{k} \sum_{s=1}^{k} \frac{\sum_{d_i \in \mathcal{DP}^{p_i}} |df(s, d_i)|}{\sum_{d_j \in \mathcal{LP}^{d_i}} |df(s, d_j)|}, \tag{3.15}$$

with k the number of clusters, $\mathcal{DP}^{p_i}$ the set of variables with priority p_i, $df(s, d_i)$ the difference factor of cluster s for dimension d_j and $\mathcal{LP}^{d_i}$ the set of dimensions with a lower priority than d_i. Both quality measures are used to monitor the impact of the parameter α using the same setup as in Section 3.4.1 and are discussed in what follows.

As a first step of the experiment, a benchmark is created by applying the classical two-step clustering approach involving the training of a 10×12 SOM on the available data followed by the k-means algorithm with a maximal number of clusters equal to 10 and 50 epochs. The best clustering is selected using the Davies-Bouldin index, leading to a benchmark partitioning that is used to analyze the impact of the prioritization. Note that the characteristics of the obtained clusters are studied by extracting the salient dimensions. The prioritized approach is applied on the same data set while assigning a higher priority to the variables of g_5 as in Section 3.4.1. Starting with the training of the 10×12 P-SOM followed by the k-means algorithm with a maximal number of clusters equal to 10 and 50 epochs, the best clustering is selected using the Davies-Bouldin index and is characterized by extracting the salient dimensions of its clusters. The amplitude of the prioritization is iteratively increased by reducing the parameter α used during the calculation of the weights. The initial value of α is set to $\frac{1}{\alpha} = 0.01$ and is reduced at iteration i using the following schema: $\frac{1}{\alpha} = 0.01 + (1 - i)0.05$ during 200 iterations leading to values of α ranging from 100 to $\frac{1}{9.96}$. Figure 3.8 shows on the y-axis the relative importance, ϕ_{p_1}, of the variables representing the interest of the customers for a future concert. The P-SOM is iteratively applied, focusing more and more on these variables by augmenting the value of $\frac{1}{\alpha}$, represented on the x-axis. It can be seen on this figure that the parameter α has the expected impact on the clustering structure considering the relative importance of the variables with higher priorities. Indeed, the

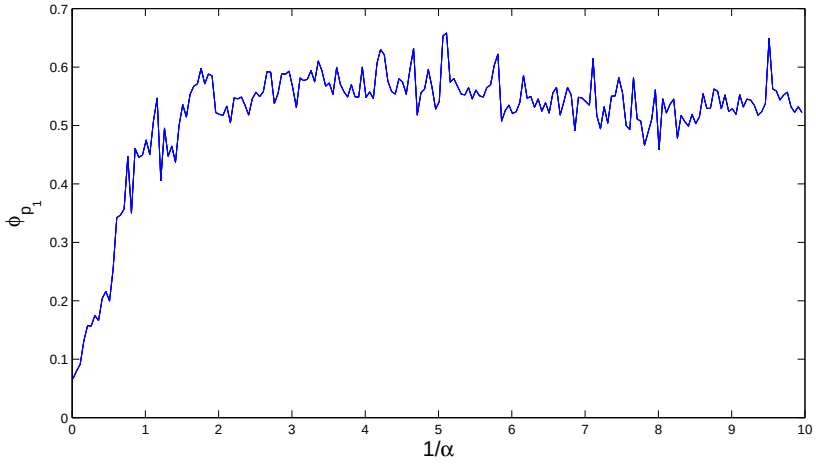


Figure 3.8: Representation of the subjective quality for different values of $\frac{1}{\alpha}$.

lower the value of α, the more the variables with higher priorities are structuring the clustering output. Starting with values around 0.06 for the unconstrained approach, the relative value increases inversely proportionally to the value of α until converging to values around 0.55, as can be seen on Figure 3.8. Figure 3.9 shows on the y-axis the ratio between the Davies-Bouldin index of the unconstrained clustering, DB_{np}, and the index of the constrained clustering, DB_p. The lower the value of the Davies-Bouldin index, the better the obtained clustering, hence a ratio lower than 1 indicates a decrease in the clustering quality when constraining the clustering algorithm while a ratio higher than 1 indicates an increase in the quality. As expected, a relative decrease in the clustering quality is identified when artificial constraints are added to the clustering task. However, the results of Figure 3.9, with a mean value of 0.9433 and some peaks above 1, put into perspective the decrease in the quality of the clustering output. Indeed, while the subjective quality is significantly improved by reducing α, the objective quality is only slightly diminished, leading to an accept-

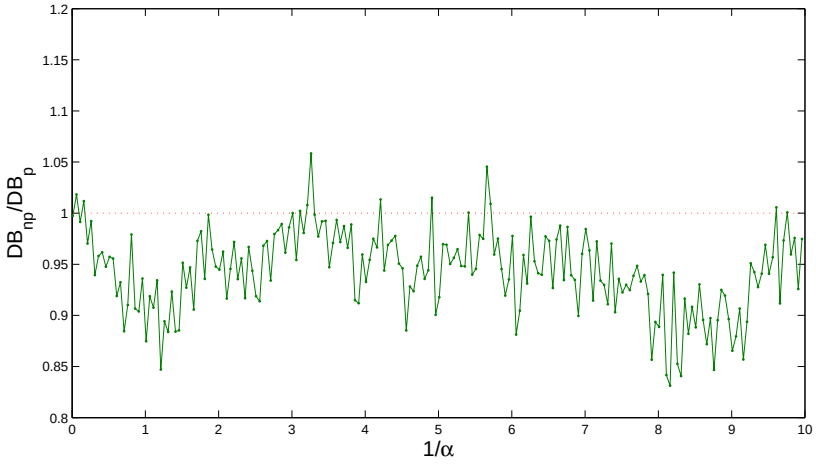


Figure 3.9: Representation of the objective quality for different values of $\frac{1}{\alpha}$.

able constrained clustering output from both objective and subjective points of view.

3.5 Conclusion

In this chapter, a new knowledge-based constrained clustering approach is proposed. This approach is based on business considerations and provides analysts with a formal way to prioritize the variables at hand by transforming the background knowledge successively into priorities, weights and soft attribute level constraints. A two-step prioritization approach is proposed based on an algorithm aiming at fixing weights based on priorities reflecting the importance of the variables. The obtained weights are then used to constrain the clustering problem by parameterizing a distance metric. A general approach for fixing the weights is proposed together with a specific approach suitable in a context involving only dummy variables. A methodology implementing

this approach is also proposed and leads to the creation of a new algorithm, the P-SOM, integrating the power of the SOM algorithm while offering prioritization facilities. This methodology is then applied in detail in a direct marketing context together with a traditional two-step clustering approach. Both methodologies are compared in terms of suitability for answering a pertinent business question. Finally, the impact of a strength parameter is discussed and illustrated based on experiments involving the available data. The impact is monitored using both subjective and objective metrics leading to the conclusion that the proposed methodology provides the analyst with the possibility to focus on some variables, hence guiding the algorithm to mainly structure the partitioning based on the available business knowledge. Although an expected decrease in the objective clustering quality is observed when artificially constraining the clustering algorithm, the experiments show that this decrease is relatively slight when compared to the increase of the subjective quality.

Chapter 4

Business knowledge based segmentation of online banking customers

4.1 Introduction

With the currently expanding internet driven services landscape, the investment in online channels represents a strategic choice for nowadays banks. The increased accessibility of these services has led to the growth of the online customer base. Moreover, it also raised the challenge of matching the marketing campaigns to the appropriate customers in order to provide an experience tailored to the specific needs of each segment. Mostly performed in an unsupervised way, data mining segmentation techniques have proven to be highly efficient in discovering those homogeneous segments of customers, based on their demographic, geographic, behavioral or psychographic characteristics. Table 4.1 presents an overview of the variables and techniques used in segmenting online customers extracted from a representative subset of the literature.

As it can be easily noticed, demographic variables have been predominantly chosen and, besides the fact that they are easily obtainable, this is mostly because of their confirmed influence on the results

Paper	Segmentation Variables					Segmentation approach
	Banking	Demographic	Geographic	Psychographic	Behavioral	
[50]	x	x		x		k-means
[51]	x	x		x		Cluster analysis
[52]		x			x	Self-Organizing Maps + k-means
[53]	x	x			x	k-means, Self-Organizing Maps, fuzzy k-means
[54]		x			x	Decision Tree
[55]		x		x	x	k-means and factor analysis
[56]	x	x		x	x	Discriminant analysis
[57]	x				x	Heuristic approach
[48]		x	x		x	Self-Organizing Maps + k-means
[58]		x	x		x	Self-Organizing maps + k-means + business knowledge

Table 4.1: Online (banking) segmentation variables and techniques in literature.

(e.g. [48], [50], [52]). Concerning the geographical variables, they are not a very popular choice except for the cases in which the population is obviously segmented by heterogeneous geographical attributes. Variables belonging to the third category, psychographic, are more difficult to obtain and therefore they have been less reflected in studies. However, their added value has been strongly defended as they offer a deeper understanding of the customers' behavior (e.g. [48]). Behavioral variables are, as demographic ones, almost omnipresent in previously conducted studies and this is mainly due to the fact that these variables have become widely available due to the recent progress in Information Technology. Overall, we can notice that customers can be characterized by these perspectives, each building onto the other ones in order to complete the analyst's insight: the demographic and the geographical variables capture who the customer is, the behavioral variables capture what the customer is doing while the psychographic variables capture the reasons why the customers are behaving a certain way. Concerning the analysis of the data, most studies of Table 4.1 have relied on partitive algorithms such as k-means which, especially in combination with Self-Organizing Maps (SOM) have produced satisfactory results ([53], [52], [48]). Different works related to the banking and financial sectors have been using the SOM for its two main functions, namely vector quantization and vector projection. In [59], self-organizing maps are used for clustering and visualization of bankruptcy trajectory. In [60], financial efficiency and social impact of microfinance institutions are explored using self-organizing maps while [61] combine it with support vector regression to visualize and evaluate corporate financial structures. From an approach perspective, the customer segmentation techniques could be classified into two main categories, namely quantitative and qualitative. Quantitative techniques focus on segmenting the customer base by feeding the input data to one or more algorithms performing the segmentation in an unsupervised manner and outputting their findings as a result. On the other hand, qualitative approaches allow the segmentation task to take into account the business knowledge possessed by the analyst. They can either be based on fully heuristic methods

([57]) or by enhancing quantitative methods with qualitative information therefore performing a constrained clustering approach ([58]). The aim of this chapter is to segment the online customer base of one of the major international banks using a quantitative as well as a qualitative approach and analyze the results provided by both of them. For the purpose of the application, a two-step clustering approach will be considered for both the quantitative and qualitative approaches. The quantitative approach consists of the application of the classical SOM algorithm followed by the k-means algorithm while the qualitative approach consists of the application of the P-SOM algorithm followed by the k-means algorithm. The SOM, P-SOM and k-means algorithms and the two-step clustering strategy have been already introduced in the previous chapters (see Sections 2.2.2, 2.3 and 3.2). A detailed description of the application is reported in Section 4.2 and is followed by a conclusion in Section 4.3.

4.2 Application

In this section, a quantitative and a qualitative segmentation of the customer base of one of the major international banks, called *TheBank* in the remainder of the chapter, is performed. It leads to different segments providing the analyst with two valuable perspectives on the data at hand. The goal of the analysis is to get insights into the behavior of the customers of *TheBank* concerning their usage of the internet banking application of *TheBank*, called *OnlineApp* in the remainder of the chapter. An important starting point for this study has been defining the *OnlineApp* user, who is represented by the customer that owns a *OnlineApp* contract and that has logged into the online *OnlineApp* application at least once in the past six months. A list of demographic, geographic, behavioral and psychographic characteristics of the *OnlineApp* user has been created, which, due to unavailability or unreliability has been reduced to geo-demographic and behavioral variables. Furthermore, with regard to the online aspect of the customer, the behavioral variables have been separated into general and

OnlineApp related variables. In order to handle the constantly changing environment, the nature of the segmentation should be a yearly repetitive one and therefore the final dataset consists of one year of data, combined from various internal sources. Preprocessing activities included duplicates, missing values, inconsistencies and outliers detection and correction. In case any of these kinds of issues have been found, the general applied rule has been to first attempt to identify and correct the reason behind the issue and, in case no reason was found, to remove the associated observations or variables. Concerning the outliers, the Inter Quartile Range (IQR) method has been used for detection as it is itself less sensitive to outliers than methods such as, for example, the Z-Score standardization (Larose, 2005). Except for extreme ones, the majority of identified outliers have been kept in the dataset, mainly because they would be combined with non-outliers into categories thus diminishing them as extremities. As required by the data mining approach, the data has been further on transformed by categorizing the variables so that it is normalized to a range of {0,1}. Concerning the discrete variables, a category has been created for each individual value or, whenever necessary, various values have been combined to form one category (e.g. the variable maritalStatus). Concerning the continuous variables, in order to avoid as much as possible information loss and to be meaningful for the business, the best cutoff-points have been chosen based on the distribution of the values combined with business knowledge. Table 4.2 summarizes the variables used for the segmentation, the corresponding categories and the associated dummy variables (dimensions). A group that will be used later on to assign priorities has been assigned to each dummy variable.

4.2.1 Quantitative segmentation

Using as input vectors the observations characterized by the 85 dummy variables listed in Table 4.2, a 20x25 SOM has been trained, leading to 500 neurons summarizing the trends in the customer database. Figure 4.1 consolidates the 85 component planes that correspond to

Type	Variable	Categories	Dummies	Group
Demographic	age	(...30] [31..50] [51..60] [61...)	D1-¿D4	g1
	gender	{Male} {Female}	D5-¿D6	g2
	maritalStatus	{Single and Cohabitant } {Widower} {Divorced, Separated Physically, Separated legally} {Married, Legal cohabitant}	D7-¿10	g3
	language	*{FR}{NL}{DE}{EN}*	D11-¿D14	g4
Geographic	region	{Brussels} {International} {Flanders} {Wallonia}	D15-¿D18	g5
Behavioral	clientType	Professional Private	D19-¿D20	g6
	riskClass	None to small risk Medium to high risk	D21-¿D22	g7
	salaryDomiciliation	YesNo	D23-¿D24	g8
	savingAndInvestmentProductsNo	[0..0] [1..1] [2..2] [3..5] [6...)	D25-¿D29	g9
	generalBankingProductsNo	[0..3] [4..6] [7...)	D30-¿D32	g10
	lendingProductsNo	[0..0] [1..1] [2...)	D33-¿D35	g11
	insuranceProductsNo	[0..0] [1..2] [3...)	D36-¿D38	g12
	currentAccountAverage	[0..0] (0..500] (500..2500] (2500..5000] (5000...)	D39-¿D43	g13
	savingsAccountAverage	[0..0] (0..500] (500..2500] (2500..5000] (5000...)	D44-¿D48	g14
	customerValue	(..0) [0..50] (50..500] (500..1500] (1500...)	D49-¿D53	g15
***Online App* related behavioral**	OASeniority	[0..3] [4..8] [9...)	D54-¿D56	g16
	sessionsNo	[0..1] (1..2] (2..4] (4...)	D57-¿D60	g17
	lastOASession	{6} {5} {4} {3} {2} {1} months ago	D61-¿D66	g18
	GIT	{Yes}{No}	D67-¿D68	g19
	mobileApp	{Yes}{No}	D69-¿D70	g20
	monthlyTransactionsAmount	[0..0] (0..500] (500..1500] (1500..3000] (3000...)	D71-¿D75	g21
	monthlyTransactionsNo	[0..1] (1..5] (5..10] (10...)	D76-¿D79	g22
	OATransactionsRatio	[0..0] (0..25] (25..50] (50...)	D80-¿D83	g23
	OAPurchasedProductsRatio	[0..0] (0..100]	D84-¿D85	g24

Table 4.2: Summary of the variables and the associated categories, dummy variables and groups.

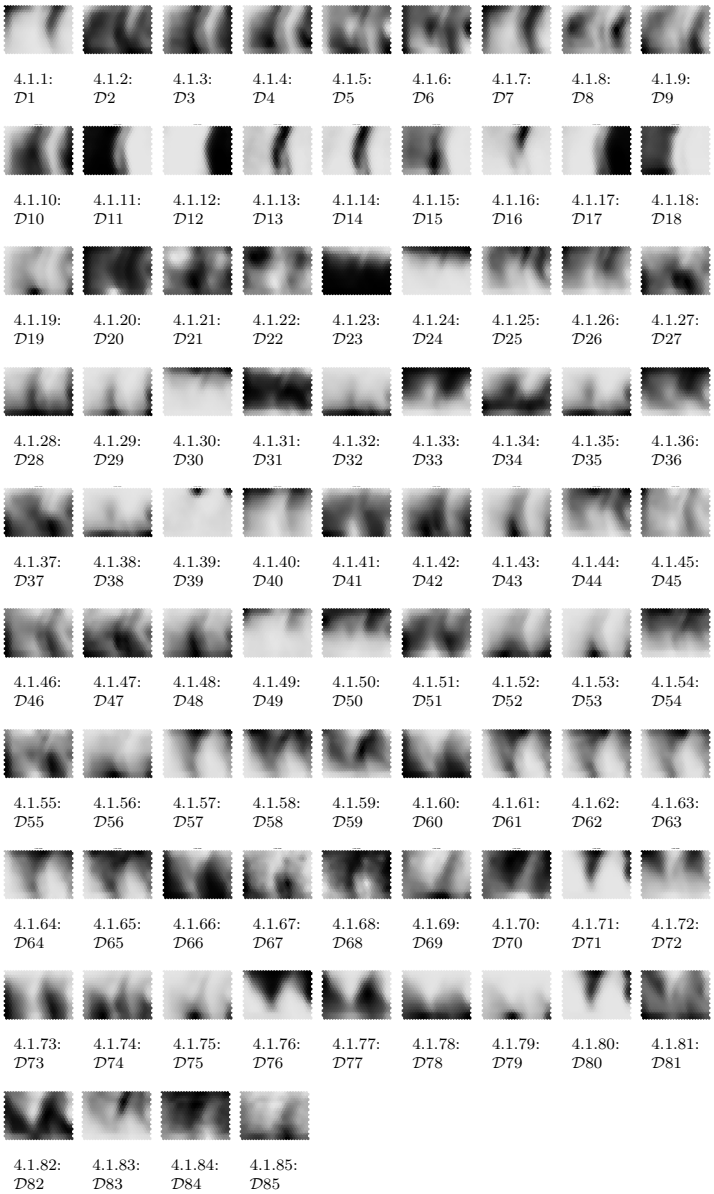


Figure 4.1: SOM output obtained by using the classical SOM algorithm.

the 85 input dimensions. Each component plane represents a projection of the neurons in the output space with a color code based on the corresponding dimension. Low values for that dimension are associated with light shades while high values are associated with dark shades. As a result of the projection, neurons close to each other in the output space should be close to each other in the original space, enabling a visual clustering analysis. Furthermore, similar color patterns found in the same area over two or more component planes indicate a correlation between the values of the corresponding dimensions which can be used to further understand the data at hand. The below are example observations that can be made by visually inspecting the SOM. First of all, concerning the demographic variables, there is a clear correlation between the dimensions D2, D3, D4 and D10, representing the customers aged between 31-50, 51-60 or higher than 61 and with a marital status of married or legal cohabitant which indeed seems to correspond with reality. Similarly, the dimension D1 (age less than 30) is correlated with D7 (marital status single or simply cohabitant). Furthermore, the patterns on these two groups of dimensions (D2, D3, D4 and D10 versus D1 and D7) are opposite, leading to the conclusion that young customers, which are mostly single or simply living together are behaving differently than middle-aged or senior customers which are mostly in an official relationship. The same conclusion can be drawn by looking at the strong correlation between the dimensions D11 (French speaking) and D18 (living in Wallonia) as opposite to the correlated dimensions D12 (Flemish speaking) and D17 (living in Flanders). With regards to the behavioral variables, a pattern can be observed across the variables related to the numbers of the various types of products, account averages and customer value. By analyzing the color patterns over the associated component planes it can be noticed that low values are projected on top of the map gradually moving horizontally towards high values projected on the bottom of the map (e.g. generalBankingProductsNo has 3 categories: [0..3] (D30), with high values clustered on top of the map, [4..7] (D31), with high values clustered in the middle of the map and [7...) (D32), with high values clustered on the bottom of the map). Furthermore,

by correlating with the salaryDomiciliation variable which is false for the neurons on top of the map and true for the neurons on the bottom of the map we can draw the conclusion that a direct relationship exists between the number of products, funds, customer value and whether the customer's salary is sent to his account by default or not. Finally, the *OnlineApp* related behavioral variables OASeniority, sessionsNo, lastOASession, monthlyTransactionsAmount, monthlyTransactionsNo and OAPurchasedProductsRatio indicate a correlation with the general behavioral variables. Concerning the lastOASession variable, it seems that customers whose last session was more than 2 months ago are similar to each other and dissimilar to the ones logging in in the past month. Judging by the correlation between the number of products, the amounts transferred and residing in their accounts and the recency and frequency of their *OnlineApp* sessions, we can conclude that customers that own a high number of products and funds are more active on *OnlineApp* than the ones with lower numbers of products and funds.

Although a lot of information can already be gained by visually exploring the SOM output of Figure 4.1, further insight into the customer behavior can be obtained by grouping the similar behaviors into more general clusters. This is achieved by running the k-means clustering algorithm onto the output of the SOM, hence clustering the neurons. As a balance between a not too low and not too high number of clusters, which would be either not enough or too complex for the business, the k-means algorithm has been executed for a number of 5 clusters, resulting in the clusters depicted in Figure 4.2. When relating to the SOM output of Figure 4.1, it can already be noticed that these clusters are divided according to the language and regional borders (clusters c_2, c_3 and c_5 versus clusters c_1 and c_4) and also by the amount of products, funds and the recency and frequency of the *OnlineApp* sessions (clusters c_2, c_3 and c_4 versus cluster c_1 and c_5). Table 4.3 summarizes the cluster characteristics identified based on the centroid weights. In order to obtain a deeper view, they have been classified into primary characteristics, associated with dimensions corresponding to a weight higher than 0,7 and secondary characteristics,

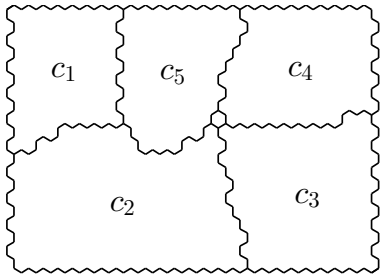


Figure 4.2: Representation of the 5 clusters obtained using the k-means algorithm on top of the prioritized SOM.

associated with dimensions corresponding to a weight between 0,5 and 0,7.

It can be observed that some dimensions appear as characteristics for most clusters (e.g. GIT: No, OAPurchasedProductsRatio: [0..0], mobileApp: No, riskClass: None to low, clientType: Private, salary-Domiciliation: Yes and lastOASession : 1 Month ago). Due to the fact that these categories describe most of the customers (e.g. 95,75% of all customers belong to the category GIT: No) they are predominant in all clusters. Besides these overlapping characteristics, cluster c_1 contains mostly Flemish speaking male customers, aged between 31 and 50 living in Flanders. They have a seniority of 4 to 8 years and carry out an average of 1 to 5 transactions by month during an average of over 4 sessions. Cluster c_2 is characterized by mostly single (or simply living together) French speaking customers living in Wallonia. They are mostly men with a current account average of 0 to 500EUR and are not owning a lot of products (0 lending and insurance, 1 savings and investments and 4 to 6 general banking products). Concerning the *OnlineApp* aspect, they have a seniority of 0 to 3 years and carry out transactions of an average amount of 0 to 500EUR per month. Similar to cluster c_2, the third cluster contains married or legal cohabitant French speaking male customers living in Wallonia and owning no lending or insurance products and 4 to 6 general banking products. Cluster c_4 contains mostly Flemish speaking male customers living in Flanders and not owning many products (0 lending and insurance and

Cluster (% size)	Type	*OnlineApp* behavioral	Demographic	Behavioral
c_1 (24,99%)	Primary	GIT: No OAPurchasedProductsRatio: [0..0] mobileApp: No lastOASession: 1 Month ago	region: Flanders language: NL	salaryDomiciliation: Yes riskClass: None to low clientType: Private
	Secondary	sessionsNo: (4...) monthlyTransactionsNo: (1..5] OASeniority: [4..8]	gender: Male age: [31..50]	generalBankingProductsNo: [4..6] customerValue: (50..500]
c_2 (16,54%)	Primary	OAPurchasedProductsRatio: [0..0] GIT: No mobileApp: No lastOASession: 1 Month ago	language: FR	clientType: Private riskClass: None to low lendingProductsNo: [0..0] insuranceProductsNo: [0..0]
	Secondary	monthlyTransactionsAmount: (0..500] OASeniority: [0..3]	maritalStatus: Single and Cohabitant region: Wallonia gender: Male	generalBankingProductsNo: [4..6] salaryDomiciliation: Yes savingAndInvestment ProductsNo: [1..1] currentAccountAverage: (0..500]
c_3 (12,65%)	Primary	OAPurchasedProductsRatio: [0..0] GIT: No mobileApp: No monthlyTransactionsNo: [0..1]	language: FR	riskClass: None to low clientType: Private
	Secondary	lastOASession: 1 Month ago monthlyTransactionsAmount: [0..0] OATransactionsRatio: [0..0]	maritalStatus: Married / Legal Cohabitant gender: Male	lendingProductsNo: [0..0] insuranceProductsNo: [0..0] generalBankingProductsNo: [4..6] salaryDomiciliation: Yes
c_4 (15,15%)	Primary	GIT: No OAPurchasedProductsRatio: [0..0] mobileApp: No	region: Flanders language: NL	clientType: Private riskClass: None to low lendingProductsNo: [0..0] insuranceProductsNo: [0..0]
	Secondary	monthlyTransactionsNo: [0..1] lastOASession: 1 Month ago OASeniority: [0..3] monthlyTransactionsAmount: (0..500]	gender: Male maritalStatus : Single and Cohabitant	generalBankingProductsNo: [4..6] salaryDomiciliation: No savingsAccountAverage: [0..0]
c_5 (30,67%)	Primary	OAPurchasedProductsRatio: [0..0] GIT: No mobileApp: No lastOASession: 1 Month ago	language: FR	salaryDomiciliation: Yes riskClass: None to low clientType: Private
	Secondary	sessionsNo: (4...)	gender: Male maritalStatus: Married / Legal Cohabitant region: Wallonia age: [31..50]	generalBankingProductsNo: [4..6] customerValue: (50..500]

Table 4.3: Summary of the cluster characteristics obtained using the weights of the centroids obtained by applying the k-means algorithm on top of the traditional SOM algorithm.

4 to 6 general banking products). Their salary is not sent to the *TheBank* account by default and they have no savings. On average they carry out maximum 1 transaction per month and they own a *OnlineApp* contract for maximum 3 years. Cluster c_5 portrays the French speaking male customers living in Wallonia with an age of 31 to 50. They have on average over 4 sessions a month and have a customer value of 50 to 500EUR.

Although the centroid weights offer a good view of the characteristics of all clusters, as it can be noticed, many of the clusters have similar characteristics. In order to subtract those characteristics that set apart each cluster from the others, the salient dimensions of the 5 clusters have been extracted. Having calculated the mean and standard deviation of difference factors for each cluster, two types of salient dimensions have been chosen: primary, those dimensions whose difference factors are higher than 1 standard deviation from the mean and secondary, those dimensions whose difference factors are between 0,75 and 1 standard deviation from the mean. By setting these higher values for the z_2 parameter we ensure that we identify those characteristics which are truly distinctive for each cluster and which are summarized in Table 4.4. It can be easily noticed that the extracted salient dimensions are consistent with the SOM findings, making the salient dimensions extraction algorithm a good automated addition to the SOM. Briefly, the 5 identified clusters correspond to: the *OnlineApp* active Flemish speaking customers living in Flanders and owning a high number of products (Cluster c_1), their French speaking counterparts living in Brussels or Wallonia (Cluster c_5) , the young low active Flemish speaking customers from Flanders owning almost no products (Cluster c_4), their French speaking counterparts (Cluster c_2) and the low-active international customers with almost no products (Cluster c_3). Although the clusters identified by applying the quantitative method are divided in a clean and intuitive way, offering great information about the customers at hand, they do not describe the customers from a *OnlineApp* perspective. As it can be observed, the main axes that describe the users are the language, region in which they live and the amounts of products and assets they own. These

Cluster (% size)	Type	*OnlineApp* behavioral	Demographic	Behavioral
c_1 (24,99%)	Primary	monthlyTransactions No: (5-10]	language: NL region: Flanders	savingAndInvestment ProductsNo: [6...) lendingProductsNo: [2...) insuranceProductsNo: [3...)
	Secondary	monthlyTransactions Amount: (500..1500] and (1500..3000] monthlyTransactionsNo: (10...)		generalBanking ProductsNo: [7...)
c_2 (16,54%)	Primary	monthlyTransactions Amount: (0..500]	age: (...30] maritalStatus: Single and Cohabitant language: FR region: Brussels, Wallonia	riskClass: Medium to high risk savingAndInvestment ProductsNo: [1..1] lendingProductsNo: [0..0] currentAccountAverage: (0..500] customerValue: (...0) and [0..50]
	Secondary	OASeniority: [0..3]		salaryDomiciliation: No generalBanking ProductsNo: [0..3] insuranceProductsNo: [0..0] savingsAccountAverage: (0..500]
c_3 (12,65%)	Primary	sessionsNo: [0..1] monthlyTransactions Amount: [0..0] monthlyTransactions No: [0..1] OATransactionsRatio: [0..0]	language: DE, EN region: International	salaryDomiciliation: No currentAccountAverage: [0..0]
	Secondary	lastOASession: 3-6 months ago		generalBanking ProductsNo:[0..3]
c_4 (15,15%)	Primary	monthlyTransactionsNo: [0..1] OATransactionsRatio: (50...)	age: (...30] language: NL region: Flanders	salaryDomiciliation: No savingAndInvestment ProductsNo: [0..0] generalBanking ProductsNo: [0-3] lendingProductsNo: [0..0] customerValue: (...0) and [0..50]
	Secondary	OASeniority: [0..3] sessionsNo: (1..2] monthlyTransactions Amount: (0..500]		insuranceProductsNo: [0..0] currentAccountAverage: [0..0] and (0..500] savingsAccountAverage: [0..0]
c_5 (30,67%)	Primary	OASeniority: [9...) monthlyTransactions Amount: (1500..3000] and (3000...) monthlyTransactions No: (5..10] and (10...)	language: FR region: Brussels, Wallonia	savingAndInvestment ProductsNo: [6...) generalBanking ProductsNo: [7...) lendingProductsNo: [2...) insuranceProductsNo: [3...) customerValue: (500..1500] and (1500...)
	Secondary			savingAndInvestment ProductsNo: [3..5] lendingProductsNo: [1..1] currentAccountAverage: (5000...) savingsAccountAverage: (5000...)

Table 4.4: Summary of the cluster salient dimensions obtained using the weights of the centroids obtained by applying the k-means algorithm on top of the traditional SOM algorithm.

would be great characteristics when performing a general purpose segmentation, however, for the aim of this exercise, focusing on a specific subset of customers, the *OnlineApp* aspects should be the main characteristics of the clusters.

4.2.2 Qualitative segmentation

As noticed from the results of the non-prioritized quantitative segmentation, the *OnlineApp* aspects have been overwhelmed by the demographic and general behavioral ones. In order to overcome this limitation, the segmentation has been executed with business knowledge input on the priority of the variables. Therefore, a priority of 1 has been assigned to the groups corresponding to the *OnlineApp* related variables (g17 to g24) and a lower priority of 2 has been assigned to the remaining groups (g1 to g16). With these priorities serving as input, the weights associated to all dimensions have been computed and further on used to train a P-SOM. The component planes of the *OnlineApp* prioritized SOM are summarized in Figure 4.3.

As opposed to the non-prioritized SOM which offers a visualization of the raw data, the *OnlineApp* prioritized SOM portrays the same data but from a different viewpoint, that of the customers' *OnlineApp* related behavior. This can be easily noticed from the cleaner component planes associated with the dimensions belonging to the higher priority groups (D54 to D85). Once more, the enhanced visualization power of the SOM proves to be efficient in a first-hand identification of patterns. For example, a similarity can be observed between customers whose last session was between 2 and 6 months ago (D61 to D65) as opposed to the customers with sessions in the past month. This pattern makes sense in the context where, having to pay bills every month for example, active users would fall in the category lastOASession=1 month ago. Another pattern can be observed over the more technology related binary variables such as whether the customer uses the GIT *OnlineApp* feature (D67 to D68), the mobileApp (D69 to D70) or whether he/she purchased any products via the *OnlineApp* channel (D84 to D85). This indicates that customers that opt for one

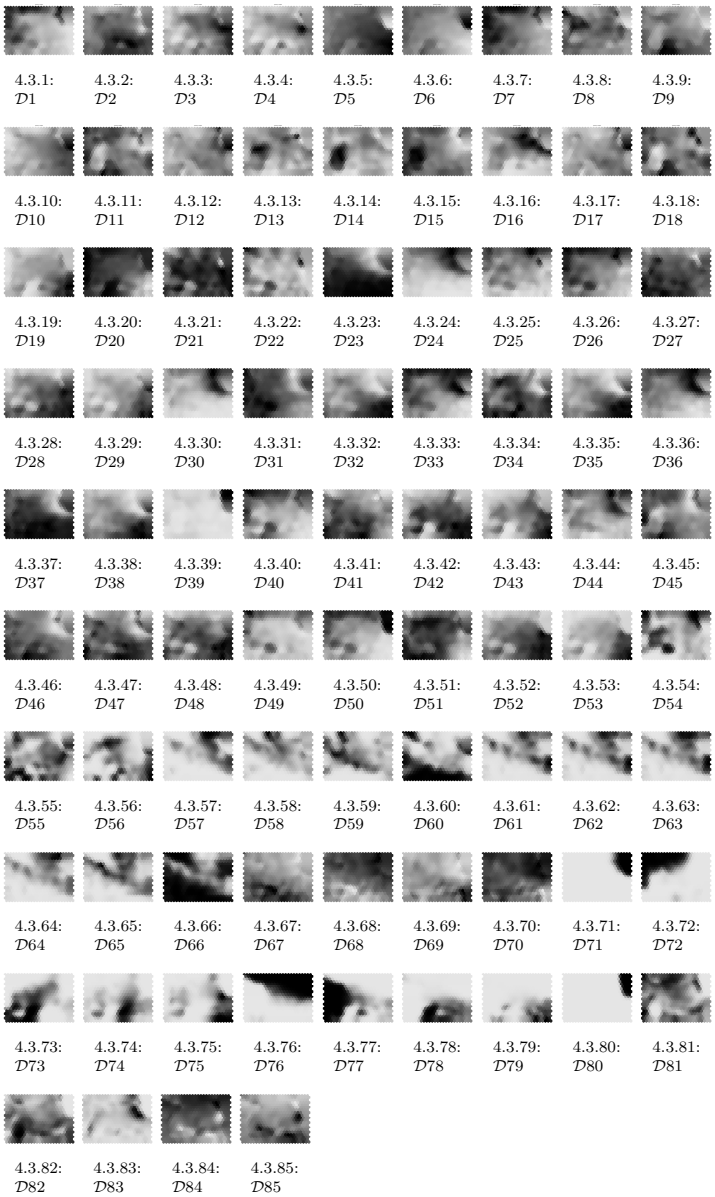


Figure 4.3: SOM output obtained by using the prioritized SOM algorithm.

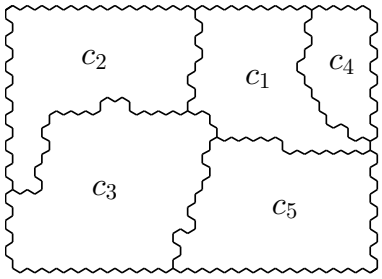


Figure 4.4: Representation of the 5 clusters obtained using the k-means algorithm on top of the prioritized SOM.

of these technology oriented services are likely to opt for the others as well. Although the difference between the high and low values is not clearly defined due to the high number of customers not using any of these services, those customers that use them are clustered towards the bottom of the map. Moreover, the likelihood to adopt these services is consistent with the high number of transactions (D79) which can easily be justified as *OnlineApp* itself is a technology oriented service. This pattern could be useful for example when identifying the primary target group when releasing a new *OnlineApp* feature or technology related service. The prioritized SOM has been further on clustered in order to group the customers with similar behavior into homogeneous groups based on a statistical approach rather than an exploratory one. For a maximum number of 5 clusters, the customers are segmented as presented in Figure 4.4 using the k-means algorithm.

As in the case of non-prioritized approach, the main cluster characteristics have been extracted as primary, with dimensions whose cluster centroid weights for the associated cluster are higher than 0,7, and secondary, with dimensions whose cluster centroid weights for the associated cluster are between 0,5 and 0,7. Same as for the non-prioritized segmentation, several dimensions are characteristic for all clusters (e.g. GIT: No). Nevertheless, it can be noticed that the ratio of *OnlineApp* variables characterizing each cluster has increased due to the higher priorities assigned to them.

Cluster c_1 is characterized by customers that have a maximum

Cluster (% size)	Type	*OnlineApp* behavioral	Demographic	Behavioral
c_1 (8,31%)	Primary	GIT: No OAPurchasedProductsRatio: [0..0] mobileApp: No monthlyTransactionsNo: [0..1]		riskClass: None to low clientType: Private
	Secondary	sessionsNo: [0..1]	gender: Male maritalStatus: Married / Legal Cohabitant language: FR	insuranceProductsNo: [0..0] lendingProductsNo: [0..0] salaryDomiciliation: No
c_2 (24,46%)	Primary	OAPurchasedProductsRatio: [0..0] GIT: No mobileApp: No monthlyTransactionsAmount: (0..500] lastOASession: 1 Month ago		clientType: Private riskClass: None to low
	Secondary	OATransactionsRatio: (0..25] monthlyTransactionsNo: [0..1]	language: FR gender: Male	salaryDomiciliation: Yes generalBankingProductsNo: [4..6] lendingProductsNo: [0..0] insuranceProductsNo: [0..0]
c_3 (16,31%)	Primary	monthlyTransactionsNo: [0..1] GIT: No OAPurchasedProductsRatio: [0..0] mobileApp: No monthlyTransactionsAmount: [0..0] OATransactionsRatio: [0..0] lastOASession: 1 Month ago		riskClass: None to low clientType: Private
	Secondary		maritalStatus: Married / Legal Cohabitant gender: Female language: FR	salaryDomiciliation: Yes lendingProductsNo: [0..0] insuranceProductsNo: [0..0] generalBankingProductsNo: [4..6]
c_4 (29,53%)	Primary	OAPurchasedProductsRatio: [0..0] GIT: No mobileApp: No lastOASession: 1 Month ago monthlyTransactionsAmount: (500..1500]		salaryDomiciliation: Yes riskClass: None to low clientType: Private
	Secondary	monthlyTransactionsNo: (1..5] sessionsNo: (4...) OATransactionsRatio: (25..50]	gender: Male language: FR age: [31..50]	generalBankingProductsNo: [4..6]
c_5 (21,39%)	Primary	GIT: No OAPurchasedProductsRatio: [0..0] mobileApp: No	gender: Male	salaryDomiciliation: Yes riskClass: None to low clientType: Private
	Secondary	lastOASession: 1 Month ago OATransactionsRatio: (25..50] monthlyTransactionsNo: (5..10] sessionsNo: (4...)	maritalStatus: Married / Legal Cohabitant language: FR age: [31..50]	

Table 4.5: Summary of the cluster characteristics obtained using the weights of the centroids obtained by applying the k-means algorithm on top of the prioritized SOM algorithm.

average of 1 session and transaction per month and whose salary is not sent to their *TheBank* account by default. They are mostly French speaking men in an official relationship. The fact that there is no salary domiciliation is a good indication that *TheBank* is not their main bank which is consistent with their low number of products (0 insurance and lending) and the low *OnlineApp* activity. Cluster c_2 portrays the customers that have an average amount in transactions less than 500EUR. With a maximum average of 1 transaction per month, less than 25% of their total number of transactions is carried out through *OnlineApp*. These customers are mostly French speaking male with 0 lending or insurance products and 4 to 6 general banking products. Cluster c_3 is characterized by French speaking females that do not carry any transactions through *OnlineApp*. Similar to cluster c_2, they own 0 lending and insurance products and 4 to 6 general banking products. Customers in cluster c_4 have a higher usage rate as they have an average number of monthly sessions higher than 4 and, while 1 to 5 transactions are carried out on a monthly average, it represents 25 to 50 % of their total number of transactions. The amount of money transferred through transactions is also higher, with a monthly average of 500 to 1500EUR. Furthermore, this cluster is mainly containing French speaking male customers with an age of 31 to 50 that own 4 to 6 general banking products. Finally, cluster c_5 is characterized by a high number of sessions as well and although the ratio of transactions carried out through *OnlineApp* is also 25 to 50 %, the average number of transactions is higher than the one of cluster 4. Once more, the cluster is dominated by French speaking male customers with an age of 31 to 50 years. Certain variables seem to consistently appear as cluster characteristics and, when consolidating the corresponding *OnlineApp* related dimensions specific for each cluster, certain trends emerge, summarized in Table 4.6.

Taking into account that cluster c_1 was the only one characterized by a lack of salary domiciliation, we can associate it with customers whose main bank is not *TheBank* and which have almost no sessions and transactions. Concerning the remaining clusters, if we define the *OnlineApp* usage as a combination of their transactions ratio, number

Cluster	OATransactionsRatio	monthlyTransactionsNo	monthlyTransactionsAmount	sessionsNo
1		[0..1]		[0..1]
2	(0..25]	[0..1]	(0..500]	
3	[0..0]	[0..1]	[0..0]	
4	(25..50]	(1..5]	(500..1500]	(4...)
5	(25..50]	(5..10]		(4...)

Table 4.6: Summary of *OnlineApp* related dimensions specific for each cluster based on the centroids.

and amount in EUR, then cluster c_3 could be perceived as a low usage cluster, followed by cluster c_2, cluster c_4 and finally cluster c_5 with the highest usage. Furthermore, by extracting the salient dimensions as described for the non-prioritized segmentation, those characteristics that truly distinguish each cluster from the others are identified and described in Table 4.7. By comparing the salient dimensions in Table 4.7 to the ones corresponding to the non-prioritized segmentation summarized in Table 4.4, it can be easily noticed that the perspective from which the data is now seen is the one of the customers' *OnlineApp* related behavior, thanks to the P-SOM algorithm.

Although most of the cluster characteristics presented in Table 4.5 and summarized in Table 4.6 are identified as salient as well, new dimensions emerge from Table 4.7 that complete the centralized view of the *OnlineApp* customer clusters as presented in the below Table 4.8, where the salient dimensions are marked in italic.

Cluster c_1 has been now completed with the fact that a subgroup of the customers belonging to it are international. This is consistent with the fact that they do not have their salary sent to the *TheBank* account by default, a low number of general banking products and sessions that date more than 2 months ago. Furthermore, being abroad could explain the fact that most of their transactions, although very few, are carried through *OnlineApp*. Concerning the remaining clusters, the usage trend is now reinforced. Moreover, cluster c_4 contains customers that are more likely to use the GIT functionality and can be therefore perceived as the technology oriented users. In comparison with the non-prioritized quantitative approach, we can clearly see the difference in perspective. Whereas the first method provided general purpose clusters, where the customers have been grouped according to demo-geographic and general behavior, the clusters resulting from the qualitative approach are focused on the *OnlineApp* related behavioral aspects. Furthermore, due to the newly gained knowledge, specific actions can be appointed to each *OnlineApp* cluster, such as marketing campaigns that could increase the usage rate of customers in clusters c_2, c_3, c_4. Being perceived as the most technology oriented cluster, customers in cluster c_4 could be approached with new online or mobile

Cluster (% size)	Type	*OnlineApp* behavioral	Demographic	Behavioral
1 (8,31%)	Primary	sessionsNo: [0..1] lastOASession: 2-6 Months ago monthlyTransactions No: [0..1] OATransactionsRatio: (50...)	region: International	salaryDomiciliation: No generalBanking ProductsNo: [0..3]
	Secondary			
2 (24,46%)	Primary	monthlyTransactions Amount: (0..500]		
	Secondary			
3 (16,31%)	Primary	OATransactionsRatio : [0..0]		
	Secondary			
4 (29,53%)	Primary	monthlyTransactions Amount: (500..1500] monthlyTransactions No: (1..5]		
	Secondary	sessionsNo: (4...) GIT: Yes		
5 (21,39%)	Primary	monthlyTransactions Amount: (1500..3000] monthlyTransactions Amount: (3000...) monthlyTransactions No : (5..10] monthlyTransactions No: (10...)		
	Secondary			

Table 4.7: Summary of the cluster salient dimensions obtained using the weights of the centroids obtained by applying the k-means algorithm on top of the prioritized SOM algorithm.

Cluster	OATransactionsRatio	monthlyTransactionsNo	monthlyTransactionsAmount	sessionsNo	Other
1	*(50...)*	[0..1]		[0..1]	*salaryDomiciliation: No* *generalBankingProductsNo: [0..3]* *region: International* *lastOASession: 2-6 months ago*
2	(0..25]	[0..1]	(0..500]		
3	[0..0]	[0..1]	[0..0]		
4	(25..50]	(1..5]	(500..1500]	(4...)	*GIT: Yes*
5	(25..50]	(5..10] *(10...)*	*(1500..3000] (3000...)*	(4...)	

Table 4.8: Summary of *OnlineApp* related dimensions specific for each cluster based on the centroids and salient dimensions.

services. Depending on its strategy, *TheBank* could choose to either invest or not in the international customers belonging to the first cluster. However, besides cluster c_1 for which the reasons for the specific customers' behavior is rather clear, the remaining clusters only portray their customers' behavior and more information could be gained by introducing psychographic variables as input for the segmentation.

4.3 Conclusion

In this chapter the online customer pool of one of the major international banks has been segmented both by using a quantitative approach and a business knowledge enhanced qualitative approach. The quantitative approach consists of two steps, namely SOM training and cluster characteristics and salient dimensions extraction. On the other hand, the qualitative approach is a generalization of the quantitative one, where an additional step has been added prior to the SOM training, giving the business-savvy analyst the opportunity to guide the algorithm to a certain direction by assigning priorities to the different groups of dimensions. In the current context, as the online behavior of the customer is the aspect that should be captured, the prioritized approach yields more meaningful results as it captures the relevant patterns from a *OnlineApp* perspective, without being overwhelmed by demographic or non-*OnlineApp* related behavior characteristics. Nonetheless, taking into account that in order to obtain a view from a specific viewpoint, the information is manipulated in order to create that specific perspective, the prioritized approach should not be used exclusively but only when the non-prioritized method does not offer suitable information. Furthermore, each step in both approaches builds upon the previous one in order to offer a deeper analytical insight into the dataset. Prioritized or not, the SOM offers a great visual insight into the data and it sets a steady basis that remains consistent throughout the next steps. Although highly visual, the SOM is also highly exploratory and additional insight can be gained by automating the clustering task which will output the desired number of

clusters in an unsupervised manner. The characteristics of each cluster can be further on described by using the centroid weights and a view of the clusters' distinctive characteristics can be obtained by extracting the so-called salient dimensions. Future research could be conducted towards formalizing the categorization method in order to normalize the input space as well as achieving a single cluster labeling method by combining both the cluster characterization based on the centroid weights with the salient dimensions extraction in order to identify those salient dimensions that are also characteristic for a given cluster.

Chapter 5

A dynamic understanding of customer behavior processes based on clustering and sequence mining

5.1 Introduction

Various data mining techniques have been proven to be a valuable approach in the quest for knowledge discovery in data from an exploratory point of view. Clustering techniques, for instance, combined with strong visualization techniques, allow analysts to get fast insights into the data they are confronted with. For these reasons, techniques such as k-means clustering and self-organizing maps have been widely and successfully applied in practice and extensively discussed in the literature ([18]).

When executed at one specific moment in time, however, as it often happens, the aforementioned techniques offer a static picture describing the composition of the data set at hand based on certain patterns derived from the attributes characterizing the instances in this data set (see e.g. [62], [31] and [63]). It would, however, be of great interest for the analyst to be able to understand the dynamics associated with

the items represented in the data base, hence recording a "movie" of the data set instead of static pictures at specific points in time. This concept is denoted "trajectory" or "customer behavior trajectory" in the remainder of this chapter. By describing an object using different attributes, it is possible to obtain a state which describes this object. When repeating this description at different points in time, a sequence of states, or trajectory, is obtained and can be analyzed. In a case where the object of interest is a customer, different attributes linked to her behavior can be captured at a specific point in time and will provide a description of the state of this customer, also called customer behavior. By repeating this description, a customer behavior trajectory is obtained.

In this chapter, which is a journal extension of [64], an approach enabling the exploratory understanding of such dynamics inherent in the capture of customers' data at different points in time is proposed. The contribution of this chapter is twofold. First, a general methodology is proposed and offers a comprehensible way to analyze movements in high dimensional spaces using unsupervised methods and visualization. Although multiple researchers are working on dynamic clustering or trajectory mining, few of them are really interested in the comprehensibility of the results for practitioners, which is one of the main research motivations of this work. Broadly summarized, our novel approach is based on a two-step clustering approach, incorporating both self-organizing maps and k-means that will generate coordinate sequences used as input for a sequence mining technique. The proposed methodology combines these methods to discover prominent customer behavior trajectories in data bases, which together help analysts to understand the behavior process as it is followed by particular groups of customers. Second, the methodology is applied in order to answer a complex business question in a real-life ticketing context. From a business perspective, understanding the dynamics of customer behaviors is a logical next step for companies applying segmentation techniques to understand their customers since, by definition, they may not stay indefinitely in the same segments. Capturing these movements becomes then a crucial objective which can only be achieved if comprehensi-

ble techniques are proposed. With this in mind, the step-wise visual approach proposed in this work aims not only at identifying the movements but also at reporting them in a way comprehensible for end-users. These different considerations show the relevance of this work for both researchers and practitioners. Moreover, thanks to the general methodology proposed in Section 5.3.2, the experiments of Section 5.4 can be easily repeated in other contexts.

The remainder of the chapter is structured as follows, in Section 5.2, an overview of related work is provided. Next, in Section 5.3, the different techniques and approaches used in the remainder of the chapter are introduced from a theoretical perspective. In Section 5.4, an application using real-life data from the concert industry is proposed and illustrates how the different concepts and techniques can be combined in order to answer advanced business questions. Section 5.5 concludes the chapter.

5.2 Related Work

Some related works following the idea of applying a dynamic approach towards exploratory data mining — and hence, clustering — have been introduced in different works. Some relevant examples are introduced in what follows. In [65], the authors propose an approach for the spatialization of multi-temporal, multi-dimensional trajectories using the self-orgnaizing map method and provide the reader with different visualization techniques combined with traditional GIS data structures in order to visualize demographic trajectories. In [66], Markov models are build using a self-organizing map to represent the different states of a process. The methodology is then used as a tool for reliability assessment in a power system network. Finally, in a recent work, [59] proposes a methodology for the clustering and visualization of bankruptcy trajectory using self-organizing map. In this approach, two self-organizing maps are trained. The first network uses vectors characterizing banks as input, offering coordinates used to generate trajectories. A second network is then trained to cluster the obtained

trajectories and visualize them. The main considerations differencing the approach of [59] from the one proposed in this chapter are the use (in this chapter) of sequence mining techniques to generate the frequent trajectories, the introduction of business knowledge to guide the clustering algorithm, the removal of repetitions as focus is put on movements rather than the duration of a particular item remaining in a certain cluster and the introduction of statistical descriptions of the identified movements. As mentioned in the above, our proposed methodology incorporates a modified sequence mining procedure similar as described in [67]. In recent years, a new research field denoted as "process mining" has sprung up, aiming to extract valuable knowledge from event based data repositories ("event logs"), including the extraction of a high-level business process model, capturing thus also an aggregated, frequency based dynamic view over the given input (see [68], [69], [70] and [71]). Contrary to this collection of techniques, however, our described approach is data driven using derived state (rather than event) sequences derived from instance level attributes collected over time, meaning that no full event log data is required to utilize the outlined methodology. In addition, state-of-art clustering techniques are applied, allowing for the identification of different customer groups (based on the behavior patterns discovered in the data) behind the same decision outcome (e.g. opting for a subscription).

5.3 Theoretical Approach

In this section, the different techniques supporting our dynamic approach are discussed. References to related work are also described in this section. For the purpose of the application, a two-step clustering approach is considered, leading to clusters that will be used further on to generate trajectories summarized using a sequence mining approach. The following subsections introduce respectively the generalized sequential pattern algorithm and the proposed methodology. This methodology makes use of concepts and algorithms introduced in the previous chapters.

5.3.1 Contiguous Sequences Identification

We apply the generalized sequential pattern (GSP) algorithm proposed in [72] to extract customer behavior trajectories. The goal of the algorithm is to find contiguous sequential patterns by analyzing a sequence data set. The algorithm starts with a first pass over the data and will store the number of occurrences of each individual item forming the different sequences and knows at the end of this step which items are frequent using a minimum support. The identified frequent items are forming the frequent sequences of size 1. In a next pass, the algorithm will create candidate frequent sequences by combining the frequent sequences of the previous step, the seed sequences. Each candidate sequence has one more item than a seed sequence and its support is obtained during the pass over the data. The algorithm terminates when there are no frequent sequences at the end of a pass, or when there are no candidate sequences generated. The interested reader is referred to [72] for and exhaustive discussion of the strategies linked to the generation and the counting of the candidates.

5.3.2 Proposed Methodology

The main objective of the proposed methodology is to provide the analyst with a technique enabling the identification of frequent trajectories and main trends followed by items represented in a database and showing an evolution through time. The state, consisting of different attributes, of each item is captured at different moments. Through time, some variables describing the states of the different items will vary, hence making the items move in a space characterized by the different attributes. In order to reduce the possible coordinates of the moving items and ease the description of their movements, a two-step clustering approach is used in order to capture the main structure of the data, hence summarizing the different possible states. By following the individual items through time in this main structure, trajectories can then be obtained and represent the evolution of the items using coordinates relative to the main structure. The proposed approach aims at creating and understanding these trajectories using both visual and

statistical techniques.

The first step consists of the application of the P-SOM algorithm, leading to a set a neurons summarizing the structure of the data while introducing some business-knowledge in the exercise. This knowledge, used as input by the algorithm, will guide the clustering task in order to achieve a partitioning showing both objective and subjective qualities. Once the neurons are trained, they represent prototypes of the items characterized by the input data. A projection of the neurons on two-dimensional maps offers then a powerful visualization facility, as will be seen later with Fig. 5.2. Because of the topology preservation feature of the SOM, and hence the P-SOM, neurons sharing similar characteristics will be located close to each other on the maps, called component planes. Patterns can then be identified and linked back to the instances using the mapping with the neurons. Although visualizing the structure of the data set provides the analyst with valuable insights, the next step of the approach consists of the capture of the visual patterns using a second clustering step: the k-means algorithm. By doing this, the output of the P-SOM can be summarized and main trends can be identified in the form of clusters of neurons. The P-SOM enabling the introduction of business knowledge in the analysis, the second clustering step, using the neurons as input, will be indirectly guided by the available knowledge about priorities. As mentioned in Section 2.2.2.2, multiple iterations of the k-means algorithm are necessary in order to obtain stable results. Concerning the number of clusters k, in practice, a maximum number of clusters k^{MAX} is chosen and partitions are created for all k with a value smaller or equal to k^{MAX}. This maximum number of clusters is often (see e.g. Section 5.4) an input from the analyst, or business, involved in the segmentation task or can be considered as a tuning parameter. The best partitioning is then selected using the Davies-Bouldin index ([23]) that can be calculated as:

$$index = \frac{1}{c}\sum_{i=1}^{c}\max_{i\neq j}\left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)}\right), \tag{5.1}$$

where c is the number of clusters, c_y is the centroid of cluster y, σ_y is the average distance of all elements of cluster y and $d(c_i, c_j)$ is the

distance between the centroid of cluster i and cluster j. A low value for the Davies-Bouldin index is to be preferred, enabling a ranking of the different partitions. A low value meaning that the inter-cluster distances are high and the intra-cluster distances are low. Note that if an upper bound as the k^{MAX} is not used, the Davies-Bouldin index cannot be used to select the partition because the optimal clustering will consist of a clustering where each point is considered as its own cluster. The best partitioning is then selected, leading to clusters characterized by centroids that can then be interpreted. Cluster characteristics can be initially derived from the weight vectors of the cluster centroids. Each weight vector has a length equal to the number of dimensions in the input space, where each weight corresponds to a specific dimension. For a cluster obtained using binary variables, a weight associated with a dimension portrays the degree to which that cluster is characterized by that dimension. High values are indicators of a high degree of characterization, as opposed to low values which correspond to a low degree of characterization.

Once the clusters are obtained, a mapping between the items and the neurons and between the neurons and the clusters allows the identification of trajectories followed by the items through time, i.e. items moving from cluster to cluster. At this point, it is very important to understand that this technique aims at understanding movements of items using the clusters as possible positions, in contrast to other dynamic clustering approaches (e.g. [73]) where clusters' movements are analyzed. The coordinates of the sequence of points forming the trajectory of each item can thus be represented by the clusters the item belongs to at different moment in time. Considering the coordinates $x^t_{n_i}$ of an input vector n_i in the original space $\mathcal{D}$ at the period $t : t \in [1..T]$, a function $\alpha(x^t_{n_i})$ returning the BMU m_c corresponding to this input vector and a function $\beta(m_c)$ returning the cluster j corresponding to this neuron, the instance-level trajectory $ITr_{n_i} = \langle x^{t=1}_{n_i}, x^{t=2}_{n_i}, ..., x^{t=T}_{n_i} \rangle$ of the input instance n_i can be transformed to the cluster-level trajectory $CTr_{n_i} = \langle \beta(\alpha(x^{t=1}_{n_i})), \beta(\alpha(x^{t=2}_{n_i})), ..., \beta(\alpha(x^{t=T}_{n_i})) \rangle$. The creation of the cluster-level trajectories is summarized in Algorithm 5.1.

Once the cluster-level trajectories of the different input vectors are

Algorithm 5.1 Create cluster-level trajectories

Require: Input vectors per instance $n_i \in N$, time horizon T.
Ensure: Cluster-level trajectories per input vector n_i: CTr_{n_i}.

1: Capture the coordinates $x^t_{n_i}$ of each input vectors n_i at each time $t : t \in [1..T]$.
2: Create a set d^A gathering the coordinates of all the input vectors at all times $t : t \in [1..T]$.
3: Apply the P-SOM algorithm on d^A.
4: Apply the k-means algorithm on the neurons obtained in the previous line.
5: **for all** $n_i \in N$ **do**
6: **for all** $t : t \in [1..T]$ **do**
7: The best matching unit $m_c = \alpha(x^t_{n_i})$ is obtained using the output of line 3.
8: The corresponding cluster $j = \beta(m_c)$ is obtained using the output of line 4.
9: The cluster-level trajectory of the input vector n_i at time t, $CTr^{t_t}_{n_i}$, is set to j.
10: **end for**
11: **end for**

obtained, two approaches are proposed in this chapter in order to understand and describe the main trends within the trajectories. For the first approach, the CTr_{n_i} of the different input vectors n_i are used as input for the GSP algorithm, leading to the generation of a set of frequent cluster-trajectories CTr_{freq} with a minimum support m_{Sup}. Because the coordinates of the frequent trajectories obtained using the GSP are clusters which coordinates in the original space are known, a mapping with the output of the P-SOM is possible using the mapping between neurons and clusters. The frequent trajectories of CTr_{freq} can thus be visualized on the component planes of the P-SOM, providing a powerful visualization facility as illustrated in Section 5.4 where a subset of the obtained CTr_{freq} is plotted on the maps. Although the visual properties of this approach can contribute to the understanding of the dynamics present in a data set, the second approach proposed in this chapter aims at summarizing the different trajectories CTr_{n_i} using a statistical approach, hence providing a statistical description of the dynamics. Instead of a description of the entire trajectory, this approach focuses on specific segments of a trajectory in order to identify trends. A cluster-level movement, or delta, $\delta^{n_i}_{t_a,t_b}$ of an input vector n_i calculated by comparing the cluster-level coordinates of n_i respectively at times t_a and t_b, with $t_b > t_a$, is defined as:

$$\delta^{n_i}_{t_a,t_b} = c_{CTr^{t_b}_{n_i}} - c_{CTr^{t_a}_{n_i}}, \tag{5.2}$$

with $CTr^{t_t}_{n_i}$ returning the cluster of the input vector n_i at time t_t and with $c_{CTr^{t_t}_{n_i}}$ returning the coordinates of the centroid of this cluster in the original space. Using this definition, multiple sets of deltas can be obtained and used to identify trends. Once a set of deltas δ^H is generated for a subset H of the input vectors, a clustering algorithm (e.g. the k-means algorithm) can be applied on δ^H in order to capture the main trends. The centroids resulting from this last step could then be used to characterize the main trends forming the dynamics of the input vectors.

To sum up and clarify the relationships between the different algorithms, Fig. 5.1 shows a summarizing overview of the main steps and outputs of the proposed methodology. The methodology starts with a

set of items (e.g. customers in Section 5.4) that are summarized using the P-SOM algorithm. The output of this step is a set of neurons which represent prototypes of items. During the second step, the k-means algorithm is applied on the neurons trained using the P-SOM algorithm, leading to a set of clusters. In a third step, the clusters are then used to create trajectories as explained in Algorithm 5.1 (see the lines 5 to 11). At this point, a set of cluster-level trajectories is created by using different algorithms (see the Steps 1, 2 and 3 of Fig. 5.1). In a fourth step, the trajectories are analyzed using two approaches. For the first approach (Step 4a of Fig. 5.1), the GSP is applied and leads to a set of frequent trajectories that are then plotted on the P-SOM maps. In the application of Section 5.4, a variant of the GSP (see Algorithm 5.2) is used to meet the requirements of the application. The only difference compared to the classical GSP is that the input trajectories are first truncated in order to better answer the business question (see line 1 of Algorithm 5.2). The second approach (Step 4b of Fig. 5.1) involves the k-means algorithm applied on delta's calculated using the cluster-level trajectories and leads to clusters capturing the main trends in some specific segments of the trajectories.

5.4 Application

An application of the proposed methodology in a ticketing context is reported in this section. The data consists of records about 67846 unique customers gathered during 66 months by a major event organizer based in the Netherlands. After preprocessing and transformation, 31 binary variables (D1 to D31) are representing the different customers. The 4 first variables represent the average number of days separating the purchase of a ticket and the event related to it; D1, D2, D3 and D4 representing respectively the categories 0, 1 to 6, 7 to 30 and more than 30 days. 4 variables represent the length of the relationship between the customer and the event organizer; D5, D6, D7 and D8 representing respectively a very short, a short, a long and a very long relationship. 5 variables represent the percentage of orders

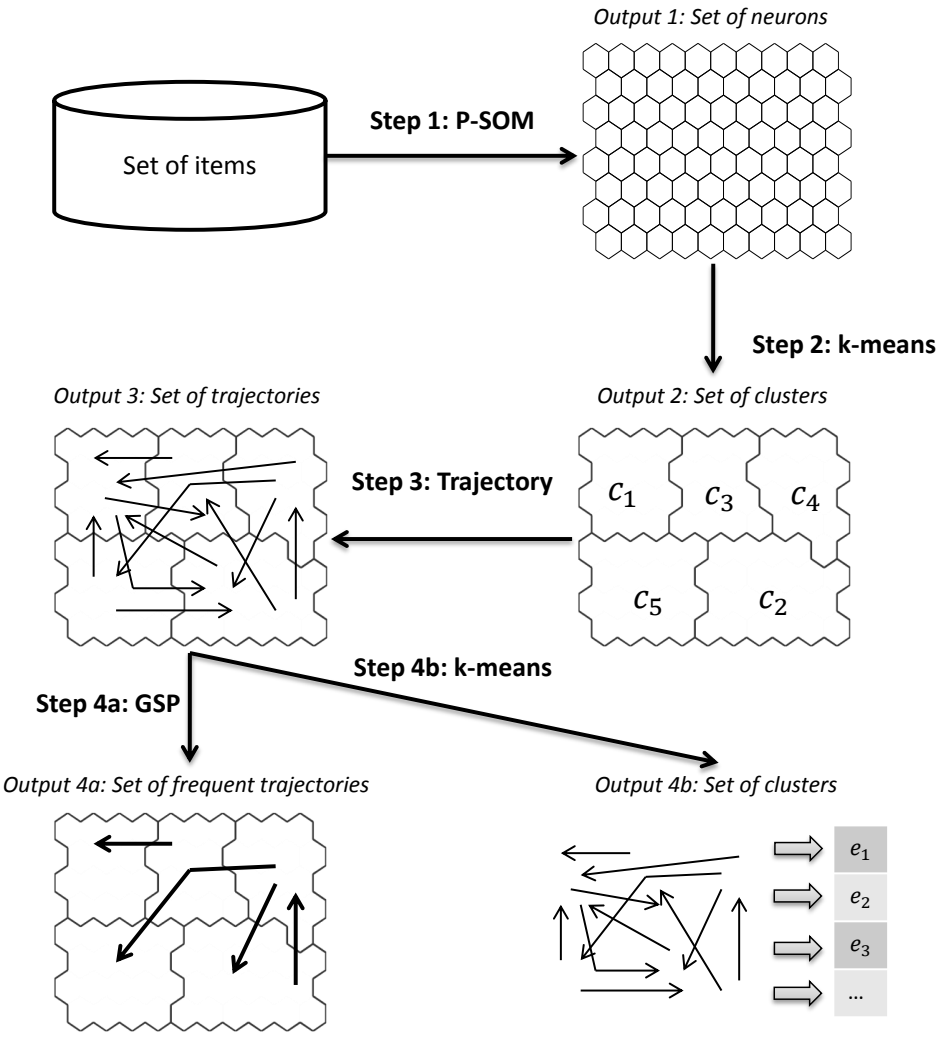


Figure 5.1: Stepwise representation of the developed methodology.

Table 5.1: Summary of the different binary variables used in this application, together with their original variables' names and ranges.

Original variable	**Range**	**Dummy**
dayDelta	0	D1
dayDelta	1 to 6	D2
dayDelta	7 to 30	D3
dayDelta	30 and more	D4
relationshiplength	very short	D5
relationshiplength	short	D6
relationshiplength	long	D7
relationshiplength	very long	D8
webPercentage	0 to 10	D9
webPercentage	11 to 40	D10
webPercentage	41 to 60	D11
webPercentage	61 to 90	D12
webPercentage	91 to 100	D13
numberTicket	1	D14
numberTicket	2	D15
numberTicket	2 and more	D16
customerCat	males	D17
customerCat	females	D18
customerCat	families	D19
customerCat	companies	D20
customerValue	very bad	D21
customerValue	bad	D22
customerValue	good	D23
customerValue	very good	D24
distance	0 to 5	D25
distance	6 to 10	D26
distance	11 to 15	D27
distance	16 to 25	D28
distance	26 to 50	D29
distance	more than 50	D30
subscriber	yes	D31

placed online; D9, D10, D11, D12 and D13 representing respectively 0 to 10, 11 to 40, 41 to 60, 61 to 90 and 91 to 100 percents of the orders. 3 variables represent the average number of tickets purchased for each event; D14, D15 and D16 representing respectively the categories 1, 2 and more than 2 tickets. 4 variables represent the category of the customers; D17, D18, D19 and D20 representing respectively the males, females, families and companies. 4 variables represent the value of the customer for the event organizer; D21, D22, D23 and D24 representing respectively the very bad, the bad, the good and the very good customers. 6 variables represent the distance separating the customer from the event location; D25, D26, D27, D28, D29 and D30 representing respectively the categories 0 to 5, 6 to 10, 11 to 15, 16 to 25, 26 to 50 and more than 50 kilometers. Finally the variable D31 captures whether or not a customer already subscribed. A summary of the different variables is to be found in Table 5.1.

The objective of this application is to understand the decision behavior of customers opting to subscribe for the first time. A person first becomes a prospect of an event organizer and will be approached (using marketing campaigns) until a first purchase is registered. This purchase can consist of a normal ticket or a subscription. Continuing his relation with the organizer, incentives will be used in order to promote the subscription to different product categories. The process ends when the prospect is removed from the data base and not approached anymore. In general, from a marketing point of view, customers with a subscription (subscribers) are associated with the customers with the highest value for the company, as confirmed by the marketing manager of the organizer providing the data used in this chapter. The goal of the organizer is thus not only to keep customers as subscribers as long as possible in order to maximize the total value of their customer base, but also to uncover the dynamics driving the decision behavior of customers opting to subscribe for the first time. Understanding consumer behavior as it dynamically changes through time is put forward as the topic of interest for this chapter. We will hence zoom in to the decision preceding the first subscription, i.e. we aim to uncover what drives customers to “buy first subscription”, as opposed to customers

which do not do so. To do so, the 66 months of data are divided into 66 periods l, with $l \in [1..66]$, respectively finishing at time l^t. At the end of each period, a data set d_l is formed and gathers the 31 variables D1 to D31 of each customer n_i present at that moment l^t in the database, hence capturing the different coordinates in the original space $x_{n_i}^{l^t}$ of each customer. An aggregated data set d^A is formed by consolidating all period data sets d_l in one main data set. In this application, with 66 periods l, the aggregated data set d^A contains 2764859 records. By definition, customers present in the database at l^t will remain in the database at $(l+1)^t$, hence being represented multiple times in d^A. Thanks to the unique identifiers of the different customers, it is thus possible to follow the evolution though time of the values of the 31 variables at a customer-level.

After this preparation step, the coordinates of the different customers in the original space $x_{n_i}^{l^t}$ and the aggregated data set d^A are captured, allowing the application of Algorithm 5.1. In a first step, the P-SOM algorithm is applied to train a network of 25×20 neurons using the aggregated data set d^A as input. The size and the shape of the SOM follow best practices (see [17]). The number of neurons is selected to be higher than the expected number of clusters, which is set to a maximum of 30 in the second clustering step. Concerning the shape of the output map, a rectangular shape is preferred, which explains the choice of a 25×20 SOM instead of e.g. a 20×20 SOM. A higher priority is used for the variable D31 (subscription holders) in order to guide the clustering algorithm and obtain a partitioning mainly structured by this variable. To do so, a priority of 1 is given to D31 and a priority of 2 is given to the other variables. Weights are then calculated using Equation 3.13 with a parameter α set to 1 and considering binary variables issued from same original variable (see Table 5.1) as being part of the same group g_l. In order to evaluate the quality of the neurons, the MQE is calculated as the average distance separating the input vectors from their respective BMUs and is equal to 1.3608. Given that 500 neurons are used to summarize the 2764859 records of d^A, such a low MQE value is the indicator of a good quantization. The output of this step is represented in Fig. 5.2 where the 31

component planes are represented based on the output of the P-SOM algorithm. The color code used for the component planes is as follows. The neurons with the highest value for each specific variable are represented, in the respective component plane, in dark red while neurons with relatively low values are colored in dark blue. The color of the neurons with intermediate values is thus ranging from dark blue (low values) to dark red (high values). Thanks to the focus on the variable D31, the reader can see that this variable is effectively structuring the output, easing the remainder of the analysis.

In a second step, the k-means algorithm is applied on top of the P-SOM output, hence clustering neurons characterized by the 31 variables of Table 5.1. A relatively high maximum number of clusters ($k^{MAX} = 30$) is used in order to allow a fine-grained analysis, leading to the 30 clusters represented in Fig. 5.3 after selecting the best partitioning using the Davies-Bouldin index. To do so, the k-means algorithm is applied multiple times (200 in this experiment) for each value of k smaller than k^{MAX}. The Davis-Bouldin index is then calculated for the different partitions and the best partition (smallest index) is selected. The value at the neuron-level of the Davies-Bouldin index (how well are the neurons represented by the clustering output) for the reported partition is equal to 0.8205 while the customer-level value (how well are the customers represented by the clustering output) is equal to 2.1806. Note that the index at the neuron-level is better (smaller) than the index at the customer-level. This observation is logical because the partition is trained directly based on the values of the neurons and is thus better able to summarize them than it does with the customers. Note also that the choice of the maximum number of clusters can have an impact on the analysis. By increasing the number of clusters used during the second clustering step, the number of possible coordinates will increase. This implies that the length and the diversity of the obtained trajectories will be proportional to the value of k^{MAX}. This parameter can thus be used as a tuning parameter depending on the desired output and business constraints.

By comparing Fig. 5.2 and 5.3 or by analyzing the coordinates of the different clusters' centroids resulting from the k-means, 6 clusters,

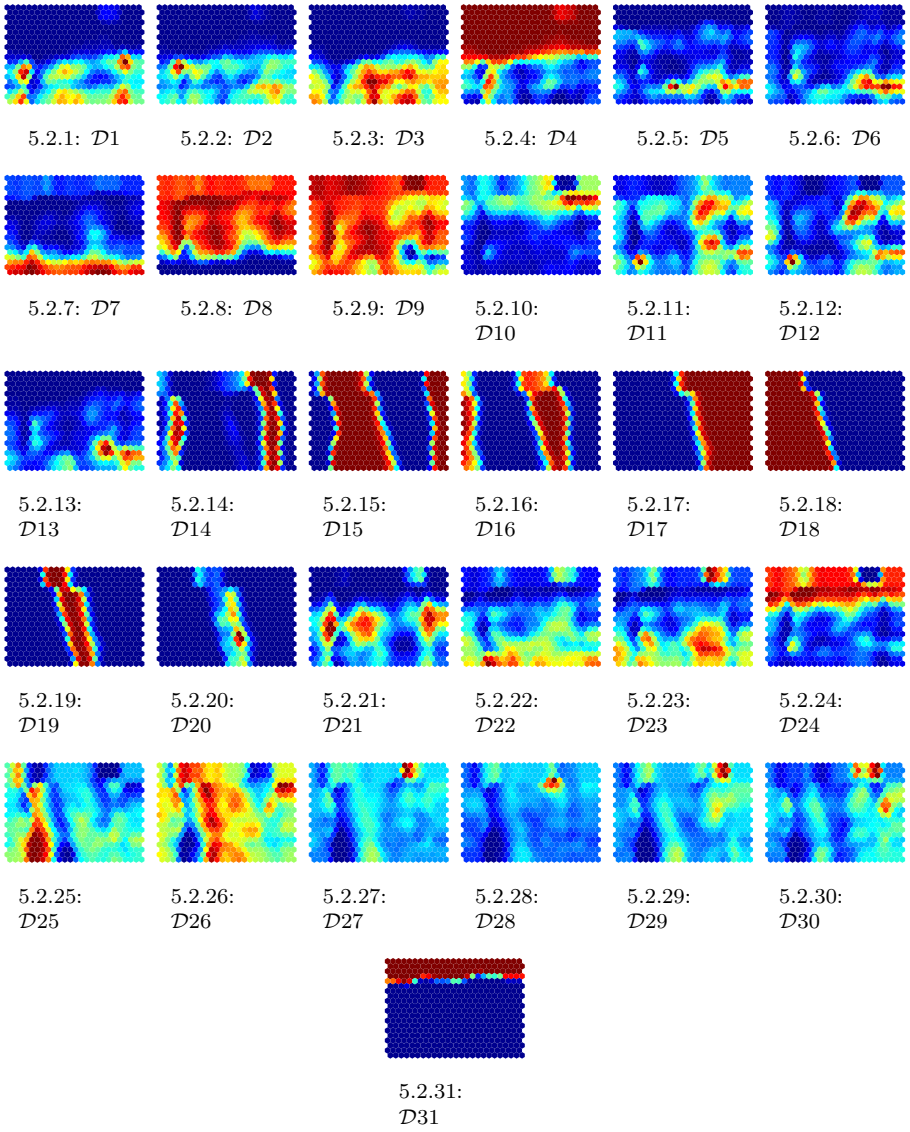


Figure 5.2: SOM output obtained by using the prioritized SOM algorithm, shown for the 31 component planes.

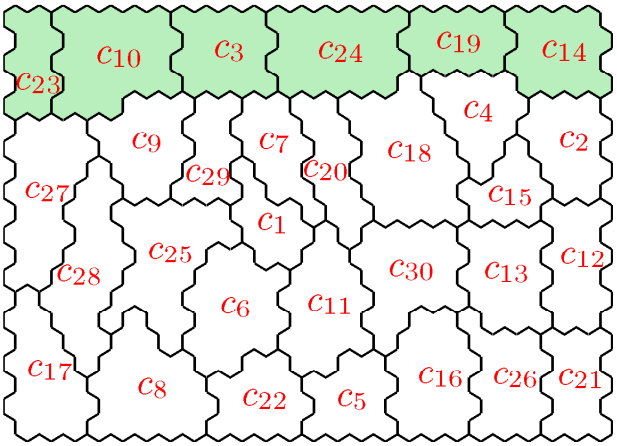


Figure 5.3: Representation of the 30 clusters obtained by applying the k-means on top of the output of the P-SOM algorithm.

namely c_{23}, c_{10}, c_3, c_{24}, c_{19} and c_{14}, are identified as clusters of subscription holders and are shaded in Fig. 5.3. In order to understand the decision behavior of customers opting to subscribe for the first time, the cluster-level trajectories CTr_{n_i} are generated for all customers (3597) who subscribed in one of the 66 periods as explained in Section 5.3.2. To take into consideration the fact that we are interested in the first period of subscription, the constrained cluster-level trajectories $CTr_{n_i}^{D31}$ are defined and consist of the cluster-level trajectories truncated after the first period in which one of the clusters representing the subscription holders (clusters c_{23}, c_{10}, c_3, c_{24}, c_{19} and c_{14}) is encountered. After removing repetitions in $CTr_{n_i}^{D31}$ (as focus is put on movements rather than the duration of a particular item remaining in a certain cluster), meaningful trajectories $CTr_{n_i}^{D31_{noRep}}$ are obtained and will be used in the remaining of this application in order to understand the trends preceding the first subscription.

As explained in Section 5.3.2, two approaches are then proposed in order to capture the dynamics. In a first approach, the different trajectories $CTr_{n_i}^{D31_{noRep}}$ are used as input for the GSP algorithm with a minimum support m_{Sup} set to 10. The set of frequent trajectories

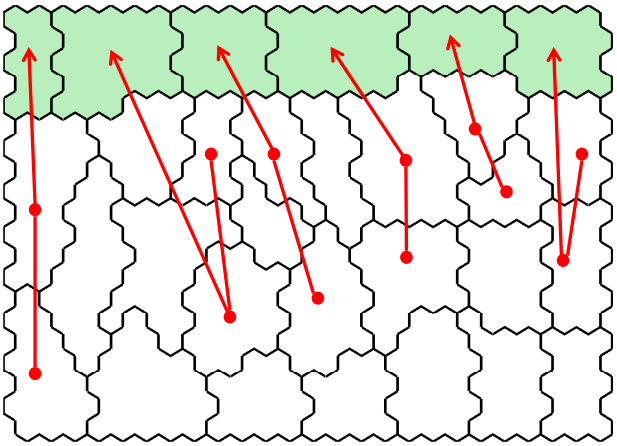


Figure 5.4: Six frequent trajectories leading to the different clusters of subscription holders.

CTr_{freq} obtained can provide the first insights concerning the decision behavior of customers opting to subscribe for the first time. As an illustration, Fig. 5.4 shows, for each cluster of subscription holders, the frequent trajectory of length 3 with the highest support. Given the number and the different lengths of the trajectories, efficient visual exploration can only be achieved by selecting subsets of the generated trajectories. This explains why trajectories of length 3 are, as an illustration, selected to show one of the possible outputs of this visual tool. Algorithm 5.2 formalizes the construction of constrained cluster-level trajectories $CTr_{n_i}^{Dj_{noRep}}$ and the extraction of frequent trajectories with minimum support m_{Sup}.

A first visual analysis of, for example, the trajectory of Fig. 5.4 leading to the cluster c_3 of Fig. 5.3 can be done by referring to the component planes of Fig. 5.2. It can then be said that this trajectory, leading to the first subscription, is associated with an increase in the average number of days separating the purchase of the tickets and the event related to it (see Fig. 5.2, dimensions D3 and D4) and an increase in the customer value (see Fig. 5.2, dimensions D21 to D24). This approach allows for an exploratory understanding which

comprehensibility highly depends on the number of frequent trajectories selected. That is why it as been decided to work with frequent trajectories of length 3 for this application while emphasizing that each business question could lead to a selection of different subset of CTr_{freq}.

As a final step — although other visual explorations could lead to additional insights into the dynamics — a statistical approach is used in order to obtain an idea of the main trends in the steps preceding the first subscription. To do so, the different trajectories $CTr_{n_i}^{D31_{noRep}}$ previously obtained are used to calculate deltas, $\delta_{t_a,t_b}^{n_i}$, with t_b and t_a being respectively the time at which the customer n_i entered for the first time a cluster of subscribers and the time preceding it. The idea here is thus to select a segment of the trajectory of each customer representing the last movement leading to the clusters of interest (namely the clusters c_{23}, c_{10}, c_3, c_{24}, c_{19} and c_{14}). As explained in Section 5.3.2, other segments of the trajectories could be investigated, leading to additional insights that one should carefully interpret, considering the different decisions previously made. The obtained deltas are then clustered using the k-means algorithm in order to understand the main trends. With a k^{MAX} equal to 5, 5 clusters representing the main trends within the deltas are obtained and are further detailed by referring to their respective centroids, summarized in Table 5.2. Note that the value of k^{MAX} is arbitrarily set to 5 as an illustration. More clusters could be created, potentially leading to more trends. However, too much clusters will make the output of the analysis more complex, which was not suitable for the business involved in this application. In order to characterize the clusters, values in Table 5.2 greater or equal to 0.6 (highlighted in Table 5.2) are considered as significant increases. The main cluster, cluster e_2, gathering 1235 deltas out of the 3597, represents customers not showing a significant increase in the value of any variable before reaching one of the clusters of subscription holders (D31). The second cluster, cluster e_1, gathering 772 deltas, represents customers showing a strong increase in the value of variables D4 and D24, hence buying their tickets earlier and becoming customers with a higher value. The third cluster, cluster e_5, gather-

ing 715 deltas, represents customers showing an increase in the value of the variables D4, D8 and D24, hence buying their tickets earlier, becoming customers with a very long relationship with the organizer and becoming customers with a higher value. The fourth cluster, cluster e_4, gathering 457 deltas, represents customers showing an increase in the value of the variables D15, hence being more and more used to buy pairs of tickets. Finally, the fifth cluster, cluster e_3, gathering 418 deltas, represents customers showing an increase in the value of the variables D16 and D23, hence buying more tickets per event and becoming better customers.

These main trends combined with the different visual patterns provide the analyst with an exploratory approach for customers' dynamics. Both techniques should however be considered as complementary. On the one hand, the first approach uses visualization techniques allowing the understanding of full trajectories while showing some limitations concerning the comprehensibility of it with an increasing number or length of the trajectories. On the other hand, the second approach offers a statistical description limited to some segments of the trajectories while overcoming the comprehensibility issue of the first approach.

Some insights obtained during this experiment have practical implications in the understanding of the customer's dynamics. First, as a result of the two-steps clustering, clear patterns can be identified within the customer base. This means that some customers are sharing some characteristics which can help the company to understand and approach them differently. A logical second step is to understand the evolution at the customer-level of these patterns, which is a new exercise for the company involved in this chapter. A first practical implication is the fact that patterns exist concerning the movements between clusters. Before this exercise, it was assumed that customers could be labeled by segmenting them, hence taking a picture. After this experiment, a clear need to understand the dynamics of the customers in order to understand their evolution has been identified. The current behavior of a customer is now enriched by analyzing the trajectory leading him to his current state, which gives again an opportunity

Table 5.2: Values of the 31 dimensions of the centroids of the 5 clusters obtained using the deltas.

	$\mathbf{e_1}$	$\mathbf{e_2}$	$\mathbf{e_3}$	$\mathbf{e_4}$	$\mathbf{e_5}$		$\mathbf{e_1}$	$\mathbf{e_2}$	$\mathbf{e_3}$	$\mathbf{e_4}$	$\mathbf{e_5}$
D1	-0.2	0.0	-0.1	-0.1	-0.2	**D17**	0.0	0.0	0.0	0.0	0.0
D2	-0.1	0.0	-0.1	0.0	-0.2	**D18**	0.0	0.0	0.0	0.0	0.0
D3	-0.3	0.0	-0.2	-0.1	-0.4	**D19**	0.0	0.0	0.0	0.0	0.0
D4	**0.6**	0.0	0.5	0.2	**0.8**	**D20**	0.0	0.0	0.0	0.0	0.0
D5	0.0	0.0	0.0	0.0	0.0	**D21**	-0.5	-0.1	-0.3	-0.2	-0.2
D6	-0.1	-0.1	-0.1	-0.1	-0.2	**D22**	-0.1	0.0	-0.1	0.0	-0.2
D7	0.1	0.1	-0.1	0.1	-0.4	**D23**	-0.1	0.1	-0.1	0.0	-0.2
D8	0.0	0.0	0.3	0.0	**0.7**	**D24**	**0.6**	0.0	**0.6**	0.3	**0.6**
D9	0.0	0.1	0.1	0.1	0.1	**D25**	-0.1	0.0	0.0	-0.1	-0.1
D10	0.0	0.0	0.0	0.0	0.0	**D26**	0.0	0.0	0.0	0.0	0.0
D11	0.0	0.0	0.0	0.0	0.0	**D27**	0.0	0.0	0.0	0.0	0.0
D12	0.0	0.0	0.0	0.0	0.0	**D28**	0.1	0.0	0.0	0.0	0.1
D13	-0.1	0.0	-0.1	0.0	-0.1	**D29**	0.0	0.0	0.0	0.0	0.0
D14	-0.1	0.1	0.2	-0.3	0.0	**D30**	0.0	0.0	0.0	0.0	0.0
D15	0.0	0.0	-0.8	**0.9**	0.2	**D31**	**1.0**	**0.9**	**1.0**	**0.9**	**1.0**
D16	0.1	-0.1	**0.6**	-0.6	-0.1						

to differentiate the approaches with the customers. A second implication of this chapter comes from the advanced business questions that can be explored using the proposed methodology. It is indeed possible to answer complex questions (e.g. dynamics preceding an event) by following the general method while generating additional insights at each step. As one of the objectives of this chapter is to provide the analyst with comprehensible techniques, different visual knowledge-based techniques are proposed and applied in this work. The design of the different techniques has been guided by the feedback of the different users who were able after a short training to gain insights into a high dimensional database by using advanced data mining techniques. Although this chapter reports only some examples used to illustrate the general method, understanding the customer's dynamics is now a key objective in the strategy of the business involved and other experiments are already planned.

5.5 Conclusion

In this chapter, a novel approach enabling the exploratory analysis of the customer's dynamics is proposed. The main aim of this chapter was to provide the analyst with techniques enabling the stepwise exploration of the data by constantly enriching the insights about the movements present in the database. In Section 5.3 the self-organizing maps, the knowledge-based constrained clustering, the k-means algorithm and the generalized sequential pattern algorithm were presented from a theoretical point of view and combined in a generic methodology enabling the understanding of the dynamics of items present in a data set. To achieve this, cluster-level trajectories are created and used as input for two approaches capturing the main trends in these trajectories. The first approach aims at finding frequent trajectories that are then plotted on the SOM, providing a powerful visualization facility. In order to summarize the trends using statistical approaches instead of the visualization, the second approach captures the main trends by focusing on specific segments of the trajectories by calculat-

ing deltas which are further clustered and interpreted. The proposed methodology has been applied making use of real data and advanced business-oriented questions in a ticketing context. The methodology was illustrated and explained into detail while guiding the reader into one way to use the proposed methodology, creating new insights into the dynamics of the customers' behavior preceding the first subscription.

From a research perspective, this chapter contributes to the literature by presenting a general methodology enabling a comprehensible exploration of movements in high dimensional spaces while introducing prior knowledge in the clustering task. The methodology is based on a combination of different algorithms, some of them being well known algorithms, some of them being contributions of this work. Being able to understand and report movements in high dimensional spaces has been identified as a gap in the dynamic clustering literature and is one of the main motivations of this chapter. By proposing two approaches to understand generated trajectories, the first moves towards comprehensible dynamic techniques are made. Moreover, by using a technique allowing the introduction of some prior knowledge in order to guide the clustering algorithm, this chapter contributes to the literature of constrained clustering and illustrates how prior knowledge can be used in an unsupervised setting.

From a business perspective, the application of the proposed methodology in a ticketing context in order to answer a complex business question with a time-dimensional aspect is, to the best of our knowledge, a novel exercise. By making each step of the proposed methodology value-adding and comprehensible, this work allows other practitioners to explore the dynamics of their databases in an unsupervised way. As a result of this experiment, the business involved decided to investigate the daily usage of dynamic techniques in order to approach their customers in an appropriate way. The main insights relate to the validation of the hypothesis that cross-clusters movements exist (can be captured and reported) and the fact that a unique answer to a complex business question may lead to incorrect conclusions. Although the experiments reported in this work are lim-

ited to one application of the proposed methodology, multiple projects in other contexts are already planned.

In future work, possibilities towards applying the proposed approach as a basis for predictive use cases can be investigated. It is indeed important to notice that this chapter aims at exploring the dynamics underlying a data set and at providing the analyst with summarization tools, hence covering the descriptive aspects of the analysis. A next step could then consist of a model making use of the knowledge generated with the proposed approach as input to better predict future states of identified trajectories (both for new data instances based on attributes available from the start or for instances already having visited different clusters), hence including predictive aspects in the analysis. Finally, further research should focus on the creation of new techniques enabling the description of full trajectories and movements in a way comprehensible for both humans and machines in order to balance the subjectivity introduced by purely visual techniques.

Algorithm 5.2 Extract frequent trajectories

Require: Variable of interest Dj, cluster-level trajectories $CTr_{n_i}, \forall n_i \in N$, time horizon T, number of clusters c, cluster(s) of interest $C := \{c_i | 1 \leq c_i \leq c\}$, minimum support m_{Sup}.

Ensure: Constrained cluster-level trajectories $CTr_{n_i}^{Dj_{noRep}}$, set of frequent trajectories CTr_{freq}.

1: Create $CTr_{n_i}^{Dj_{noRep}}$ by truncating CTr_{n_i} after first occurrence of Dj and removing repetitions:
$\forall t \in [1..T] : (\exists t', t' < t \wedge CTr_{n_i}^{t_{t'}} = Dj) \vee (CTr_{n_i}^{t_t} = CTr_{n_i}^{t_{t+1}}), remove\ CTr_{n_i}^{t_t}$
2: $CTr_{freq} := \{\}$
3: Construct initial trajectories of length 1: $Q := \{\langle c \rangle | c \in C\}$.
4: **while** $Q \neq \emptyset$ **do**
5: **for all** $q \in Q$ **do**
6: $Q := Q \setminus q$
7: Calculate support *sup* for subsequence q in $CTr_{n_i}^{Dj_{noRep}}$.
8: **if** $m_{Sup} \leq sup$ **then**
9: Add q as frequent trajectory: $CTr_{freq} := CTr_{freq} \cup \{q\}$.
10: Expand q in longer trajectories and add to queue: $\forall c' \in 1..c, Q := Q \cup \langle c' \rangle \| \langle q \rangle$.
11: (Optional) Prune Q: remove subsumed trajectories.
12: **end if**
13: **end for**
14: **end while**

Note: $\|$ is the sequence concatenation operator, e.g. $\langle a \rangle \| \langle b \rangle = \langle a, b \rangle$. Line 11 describes an optional pruning step where frequent trajectories are removed from Q when another trajectory exists in Q which contains the first trajectory, thus only retaining longer trajectories.

Chapter 6

Identifying next relevant variables for segmentation by using feature selection approaches

6.1 Introduction

Using data mining techniques as a support for operational decisions seems to be present in the agenda of more and more companies willing to monetize their data. While some companies are only starting the journey, some others are already a step further, facing operational challenges related to the post-processing and updating of the generated knowledge. As a well-known and hence frequent data mining application, the segmentation of customers using clustering techniques is also impacted by these considerations [74]. Once segments are obtained, different practical steps can be considered. If the clustering structure based on which the segmentation is made is taken as a fixed structure and if some characteristics of the customers can vary trough time, a logical next step could be to update the positions of the customers, hence updating their memberships to the segments. By capturing these movements relatively to the fixed structure, the dynamics of the cus-

tomers can thus be explored (see e.g. [75]). On the other hand, one could consider the clustering structure as an organic component that can evolve and change as a reaction to some triggers and decisions. For example, a company could decide to update the segmentation of its customers by re-running the segmentation's algorithms at regular interval using updated characteristics of its customers or by adding new customers. Typical decisions are then made concerning how to make the structure evolve and whether or not to increase or decrease of the number of clusters. Analyzing these clusters' movements is studied by a domain called dynamic clustering in which dynamic patterns at the cluster-level are identified through time (see e.g. [76]). Such an analysis focuses on the impact of the evolution of the values of the customers' characteristics on the segmentation.

Contrasting to this, one could be interested in the impact of the evolution of the set of characteristics on the segmentation, which is the topic of this work. More especially, this chapter discusses a particular case of such an evolution of the characteristics in a segmentation context. Considering a customer segmentation based on a set of variables, further referred as the original variables, this work aims at ranking variables from another set of new attributes, further called the candidate variables, based on their relevance for improving the original segmentation.

The business relevance of this work is further illustrated by a case study involving the marketing department of a main event organizer based in Europe. After performing a segmentation of their customers using the variables perceived at that moment as the relevant variables, the business involved integrated the obtained segments in their strategy and are still, at the moment of the writing of this paper, using these segments to guide their efforts. As mentioned above, different steps as dynamic clustering or segmentation maintenance through update are conceivable. This work reports experiments performed while trying to identify next relevant variables that could enrich the segmentation while considering the original segments as prior knowledge. This idea of prior knowledge used as input for a new analysis is the main justification of this research from a practical point of view. Facing the

decision to re-segment from scratch or to enrich their current segmentation, the business involved opted for the later option. Inspired by other works as [58] and [77], a need for a methodology allowing to guide a new segmentation by selecting the appropriate variables using the results of an existing segmentation has been identified and discussed in this work. The original segmentation and the identification of the next relevant variables are described in detail in this work using real data and solving a real problem of the ticketing industry.

The scientific relevance of this work is threefold. First, in order to rank the candidate variables conditionally to the original segmentation, seven different candidate techniques from the literature are described and used as feature selection mechanisms. The results of the experiments are then evaluated using four different evaluation criteria for clustering. Second, a new feature selection algorithm designed for this problem is proposed and compared with the other techniques. Finally, the different steps are reported as a generic methodology that could be applied in other domains, opening new tracks for research.

The remainder of this chapter is structured as follows. In Section 6.2, the theoretical background on unsupervised learning for customer segmentation is summarized, including a description of relevant evaluation metrics for clustering performance. Section 6.3 presents different feature selection techniques that are relevant in this work. The proposed feature selection framework for updating a current clustering model is presented in Section 6.4. In Section 6.5, the proposed methodology is applied in a real case involving a marketing department from the event industry. Section 6.6, concludes this chapter and identifies new tracks for research.

6.2 Data clustering for customer segmentation

Customer segmentation is an approach aiming at grouping similar customers in order to better understand and approach them. Widely discussed in the literature, segmentation exercises have been conducted

in different contexts, allowing researchers to identify critical issues and best practices. In a recent work, [78] identifies five of these main critical methodological issues related to segmentation research and discusses different considerations related to it. The first category of issues mentioned in their work concerns the problem definition related issues, one of the major considerations of which is the selection of segmentation variables and models. In this paper, the aim is to identify new relevant variables using an existing segmentation of a customer base, which is strongly related to the segmentation variables and models selection consideration of [78] and further discussed in the remaining of the manuscript. In this specific section, the clustering approach used as segmentation technique in the application of Section 6.5 is briefly introduced while referring the reader to previous works discussing it into detail. Note that the focus of this paper is more on the identification of new relevant variables for segmentation than on the segmentation technique itself, which explains the limited space used to discuss it while referring the reader to interesting related works. In order to allow for an evaluation of the resulting clustering partitions, an introduction to four evaluation metrics widely discussed in the literature is added to this section.

6.2.1 The two-step clustering approach

The clustering approach used in this work is based on a two-step clustering strategy discussed and applied in works as [60], [48], [21] and [22]. The first step of this strategy consists of reducing the number of data points by training a self-organizing map (SOM) with a high number (lower than the number of input data points while being significantly higher than the expected number of clusters) of neurons (see [17] for more details). By doing this, prototypes of the original data points are obtained and preserve the topology of the data in the original space. Combining both quantization and visualization facilities offered by the SOM algorithm, it is then possible to represent and visualize the structure of the data while capturing the main patterns with a second clustering step. During this second step, the neurons pre-

viously trained are clustered using a classical partitioning algorithm as the k-means algorithm [79]. Since the number of neurons is significantly lower than the original number of data points while preserving the dimensionality, more computationally expensive experiments can be conducted as discussed in [22] and illustrated in Section 6.5. As an output of this two clustering steps, a mapping between the original data and the neurons and between the neurons and the clusters of the second step allow to cluster the original data points and to obtain cluster centers represented in the original space. This strategy is summarized on Figure 6.1 and applied in Section 6.5. The reader is referred to [17] for an exhaustive discussion of the SOM algorithm and to the previously cited works for examples of applications.

6.2.2 Clustering evaluation metrics

When facing a partitioning of data points resulting from a crisp clustering algorithm as the one presented in the previous section, internal clustering validation measures have to be considered in order to evaluate the quality of the obtained output. Different measures are proposed in the literature, a good summary of which is discussed in [80]. Although these evaluation metrics have specific characteristics, two general evaluation criteria can be identified. The first criterion is the *compactness* of a partitioning, which measures how closely related the objects in a cluster are. By reducing the variance within the different clusters of a partitioning, a higher *compactness* is obtained, resulting in a better clustering. The second general criterion is the *separation*, reflecting how well-separated or distinct a cluster is from other clusters. Creating partitions with a high inter-cluster dissimilarity, hence being well-separated, leads to clusters showing unique characteristics, which improves the general quality of the clustering. These two general criteria are key concepts leading the design of multiple clustering algorithms and evaluation metrics. In what follows, four metrics widely used and discussed in the literature are briefly introduced and positioned with regard to the two general criteria of *compactness* and *separation*. These evaluation metrics are further used in Section 6.5 in

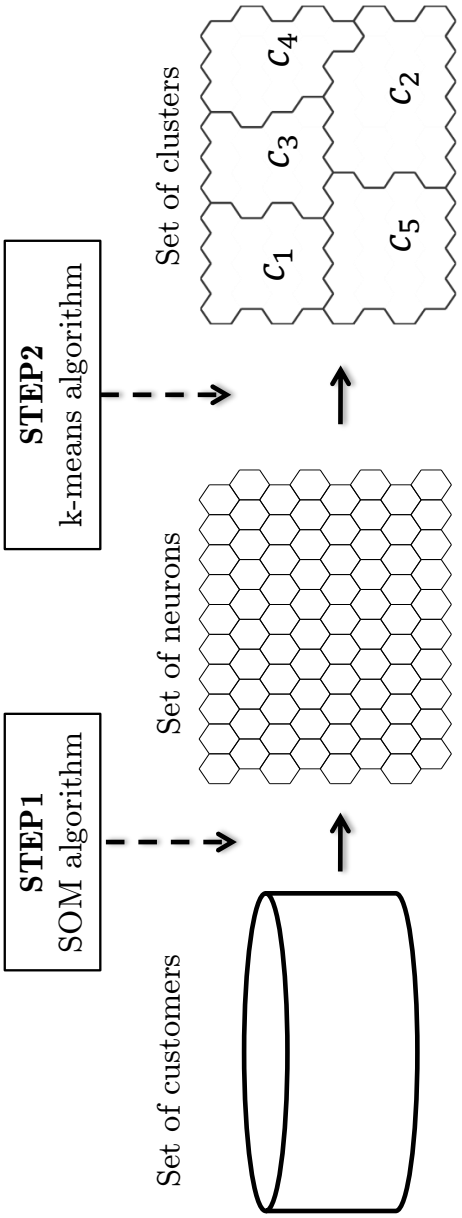


Figure 6.1: The 2-step clustering strategy.

order to evaluate the different partitions resulting from the different feature selection approaches.

The first evaluation metric is the root-mean-square standard deviation (RMSSTD) of a partition $\mathcal{K}$ (a set of clusters $\mathcal{C}_i$, with $i = 1, ..., NC$) calculated as:

$$RMSSTD(\mathcal{K}) = \sqrt{\frac{\sum_i \sum_{x \in \mathcal{C}_i} d(x, c_i)^2}{p \sum_i (n_i - 1)}}, \tag{6.1}$$

with x being one of the n objects of the set $\mathcal{N}$ of all samples. Each element of $\mathcal{N}$ belong to a given cluster $\mathcal{C}_i$ with center c_i, n_i being the number of objects in cluster $\mathcal{C}_i$, p being the number of variables in the model, and $d(x, c_i)$ being the Euclidean distance between an object x and the centroid c_i. This metric evaluates the homogeneity of the formed clusters and is hence focusing on the *compactness* criterion.

The next metric is the R-squared (RS) of a partition $\mathcal{K}$ calculated as:

$$RS(\mathcal{K}) = \sqrt{\frac{\sum_{x \in \mathcal{N}} d(x, c)^2 - \sum_i \sum_{x \in \mathcal{C}_i} d(x, c_i)^2}{\sum_{x \in \mathcal{N}} d(x, c)^2}}. \tag{6.2}$$

By calculating the ratio of the sum of squares between clusters to the total sum of squares of the whole data set, an evaluation of the degree of difference between the clusters is obtained, hence evaluating the *separation* criterion. The next metric, the Calinski-Harabasz index (CH), combines both the *separation* and *compactness* criteria by taking into account the between- and within-cluster sum of squares. The CH of a partition $\mathcal{K}$ is calculated as:

$$CH(\mathcal{K}) = \frac{\frac{\sum_{x \in \mathcal{N}} d(x,c)^2}{NC-1}}{\frac{\sum_i \sum_{x \in \mathcal{C}_i} d(x,c_i)^2}{n-NC}}, \tag{6.3}$$

with c being the center of $\mathcal{N}$. The fourth metric considered in this chapter is the Davies-Bouldin index (DB) of a partition $\mathcal{K}$ calculated

as:

$$DB(\mathcal{K}) = \frac{1}{NC} \sum_{i} \max_{i',i' \neq i} \left(\frac{\frac{1}{n_i} \sum_{x \in \mathcal{C}_i} d(x, c_i) + \frac{1}{n_{i'}} \sum_{x \in \mathcal{C}_{i'}} d(x, c_{i'})}{d(c_i, c_{i'})} \right). \tag{6.4}$$

Each cluster $\mathcal{C}_i$ is assigned the highest similarity between this cluster and all other clusters. The DB index is then the average of these different similarities, a low value meaning that clusters are very distinct, hence focusing on the *separation* criterion.

Although these four metrics only represent a subset of the available metrics (see [80] for additional references), combining them to evaluate a partitioning allows us to obtain a clear idea of the output quality. Note that since the scales of the different metrics are different and difficult to interpret as such, a relative ranking of different clustering results based on the different metrics is preferable. For example, consider two partitions $\mathcal{K}^1$ and $\mathcal{K}^2$ obtained after clustering a data set $\mathcal{N}$. If $RMSSTD(\mathcal{K}^1)$ is greater than $RMSSTD(\mathcal{K}^2)$, $RS(\mathcal{K}^1)$ is smaller than $RS(\mathcal{K}^2)$, $CH(\mathcal{K}^1)$ is smaller than $CH(\mathcal{K}^2)$ and $DB(\mathcal{K}^1)$ is greater than $DB(\mathcal{K}^2)$, we may say that $\mathcal{K}^2$ is a better partitioning than $\mathcal{K}^1$. In some cases, different metrics may lead to different conclusions, which motivates the use of more than one metric to assess the relative quality of a clustering output.

6.3 Feature ranking techniques

In this section we briefly describe the feature ranking approaches used in this work to sort the candidate variables according to their relevance, given a current clustering partition obtained using the original variables. Feature ranking for unsupervised learning aims at ranking these candidate variables based on feature extraction techniques, taking into account the redundancy between the candidate variables. Principal Component Analysis and Generalized Hebbian Algorithm [81] are used to extract one component from the correlation matrix created with all candidate variables, while these variables are ranked

according to their correlation with this component. Supervised feature ranking is performed by assessing each candidate variable's correlation with the labels (in our case, the current partition). This can be approached either in a bivariate fashion (using e.g. methods as Fisher Score, Chi Square, and Information Gain), or using multivariate methods (using e.g. methods as RELIEFF and Random Forest). A brief description of these different techniques is reported in what follows in order to provide the reader with a basic understanding of the main concepts. These techniques are further used in Section 6.5 as benchmark for the proposed approach.

6.3.1 Fisher Score

A commonly used filter method is the Fisher Criterion Score (F) [82], which assesses each feature's importance by computing the correlation between each candidate variable $j \in \mathcal{D}$, where $\mathcal{D}$ is the set of the candidate variables, and the output labels given by the current clustering partition:

$$F(j) = \frac{\sum_{i=1}^{NC} n_i \left(\mu_{i,j} - \mu_j\right)^2}{\sum_{i=1}^{NC} n_i \sigma_{i,j}^2}, \tag{6.5}$$

where μ_j represents the mean of variable j, $\mu_{i,j}$ and $\sigma_{i,j}$ are the mean and standard deviation for variable j on cluster i, respectively, and n_i is the number of elements that belong to cluster $i = 1, ..., NC$.

6.3.2 Chi Square test

The Pearson's χ^2 independence test determines whether the distribution of paired observations on two categorical variables, expressed in a contingency table, are similar to each other [83]. The test is the sum of the squared difference between observed and the expected (or theoretical) frequency that holds under the hypothesis of independence, divided by the expected frequency in all possible levels (numerical

candidate variables are previously binned into R levels):

$$\chi^2(j) = \sum_{r=1}^{R} \sum_{i=1}^{NC} \frac{(n_{i,r} - \mu_{i,r})^2}{\mu_{i,r}}, \tag{6.6}$$

where $\mu_{i,r} = \frac{n_{*i} n_{r*}}{n}$ is the expected fraction of samples for the r-th feature category of variable j and cluster i, and $n_{i,r}$ is the number of examples in the r-th level of variable j that belongs to cluster i.

6.3.3 Information Gain

The information gain corresponds to the change of information entropy by including the variable, with respect to the baseline entropy of the dataset [84]. The information gain of a variable j that has been binned into R levels is defined as:

$$IG(j) = -\sum_{i=1}^{NC} P(i) log\ P(i) + \sum_{r=1}^{R} P(r) \sum_{i=1}^{NC} P(i|r) log\ P(i|r), \tag{6.7}$$

where $P(\cdot)$ is the ratio of the particular category in the dataset: $P(i) = \frac{n_i}{n}$, $P(r) = \frac{n_r}{n}$, with n_r being the number of examples that belong to level r, and $P(i|r) = \frac{n_{i,r}}{n_r}$ being the fraction of samples of level r that belongs to cluster i.

6.3.4 Random Forests

Feature selection can be performed with tree-based ensembles, such as Random Forests, which inherit all nice properties of single tree while providing a more reliable ranking since the relevance measure is averaged over the N trees in the ensemble [85]. Considering a tree that uses information gain, the variable importance can then be defined as the sum over the tree nodes t [86]:

$$M(j) = \sum_{t \neq T} IG(j, t). \tag{6.8}$$

The variable importance at the ensemble-level is then the average $M(j)$ over the N trees in the ensemble:

$$RF(j) = \frac{1}{N} \sum_{k=1}^{N} M_k(j). \tag{6.9}$$

6.3.5 RELIEFF

RELIEFF is an iterative algorithm that estimates the quality of the variables according to how well their values distinguish between objects that are near to each other [87]. For a randomly selected example $\mathbf{x}$, the method searches for its nearest neighbors: one from the same class (*near-hit*) and one for every opposite class (*near-miss*). The algorithm iteratively updates a weight vector W for all attributes by subtracting the distance between the randomly selected example $\mathbf{x}$ and its near-hit, and adding the weighted distances to the near-miss of every opposite class, as follows:

$$W := W - (\mathbf{x} - NH)^2 + \sum_{i \neq class(\mathbf{x})} \frac{P(i)(\mathbf{x} - NM_i)^2}{m}, \tag{6.10}$$

where NH and NM_i respectively denotes the near-hit and near-miss to every opposite cluster i. The rationale behind this approach is that a relevant attribute should differentiate between examples from different classes (clusters in our case) and should have similar values for samples from the same class [87]. These weights are used as a relevance vector for feature ranking after a predefined number of iterations.

6.3.6 Feature Ranking via PCA and GHA

Extraction can be used as a filter technique for unsupervised feature selection. Different methods have been proposed in the literature to perform such extraction of components: Principal Component Analysis (PCA), Kernel PCA, Singular Value Decomposition (SVD), nonlinear PCA, and the Generalized Hebbian Algorithm (GHA); see [81]

for an overview of the respective approaches. These methods can be adapted to rank the original attributes in terms of their influence in the components. In particular, we use the weights associated with each attribute that constructs the first component to rank the variables in terms of relevance, where the higher the weight of an attribute in magnitude the more relevant it is considered.

In this chapter the Principal Component Analysis (PCA) and the Generalized Hebbian Algorithm (GHA) are used for feature ranking. PCA performs feature extraction in such a way that the first principal component has the largest possible variance, resulting in the eigenvector associated to the largest eigenvalue of the covariance matrix. Each succeeding component has the highest variance possible under the constraint that it is orthogonal to the preceding components. A detailed discussion about PCA can be found in [81].

GHA has been presented in [88] and generalizes the learning rule proposed in [89]. Inspired by insights from synaptic plasticity in neuroscience, this rule basically says that the weight between two neurons increases if both neurons are active at the same time. Applying this idea to the extraction of principal components can be formalized as follows. First, the single output neuron of a one-layer feedforward neural network y that represents the first principal component is computed as follows:

$$y = \sum_{j \in \mathcal{D}} w_j x_j, \tag{6.11}$$

where x_j represents input neuron (attribute) j and w_j its respective weight. The following rule is used to update the weights:

$$w_j := w_j + \eta y \left(x_j - y w_j \right), \tag{6.12}$$

where η is a positive learning rate. GHA offers certain advantages over PCA since it determines the components in an iterative way rather than calculating the covariance matrix as is the case in PCA.

6.4 The proposed feature ranking methodology

The proposed methodology is a filter approach that ranks a set of candidate variables $\mathcal{D}$ using the information of an existing set of clusters $\mathcal{K}$. To do so, two of the main concepts related to the intrinsic quality of a partitioning are considered: namely the *compactness* and the *separation* of the resulting partition. The idea here is to sort the variables according to their capacity to form compact clusters that are well separated from other clusters. In order to assess the individual variables instead of the whole partitioning (as traditionally done), we introduce the concept of *compactness* at an attribute-level, $Comp_j$, and the concept of *separation* at an attribute-level, Sep_j; j being a variable of $\mathcal{D}$. In order to calculate these two measures, different intermediary steps are necessary, as explained in what follows.

Recalling the notation presented for the method Fisher Score in the previous section, we define $\mu_{i,j}$ as the mean for candidate variable j on cluster i, where $\mu_{i,j} = \frac{\sum_{x \in \mathcal{C}_i} x_j}{n_i}$. Additionally, we refer to $\mu^C_{i,j}$ as the mean value for variable j of all objects that do not belong to cluster i, i.e. $\mu^C_{i,j} = \frac{\sum_{x \in \mathcal{N} \setminus \mathcal{C}_i} x_j}{|\mathcal{N} \setminus \mathcal{C}_i|}$, where $|\mathcal{N} \setminus \mathcal{C}_i|$ is the cardinality of the set of all elements but excluding those that belongs to cluster i.

Based on previous definitions, we compute the *compactness* of a variable j as:

$$Comp_j = \frac{1}{NC} \sum_{i \in \mathcal{K}} \left(\frac{1}{n_i} \sum_{x \in \mathcal{C}_i} abs(x_j - \mu_{i,j}) \right) \tag{6.13}$$

and the *separation* of a variable j as:

$$Sep_j = \frac{1}{NC} \sum_{i \in \mathcal{K}} abs(\mu_{i,j} - \mu^C_{i,j}), \tag{6.14}$$

where $abs(f)$ is the absolute value of f. The variables in $\mathcal{D}$ are sorted using this two criterion, creating two different rankings. A first ranking, LC, contains the variables in $\mathcal{D}$ sorted according to their *compactness*, starting with the variable with the lowest value. A second list,

LS, contains the variables in $\mathcal{D}$ sorted according to their *separation*, starting with the variable with the highest value. For both lists, ties are solved randomly. As a final step, considering a function $p(L, j)$ returning the position of a variable j in a list $L \in \{LC, LS\}$, the final ranking of the variables of $\mathcal{D}$, $CS(\mathcal{D})$, is a list of the variables of $\mathcal{D}$ sorted according to both their *separation* and *compactness*, starting with the variables with the lowest values of $pos(LC, j) + pos(LS, j)$, solving ties randomly.

The result of this approach is a sorted list of the variables of $\mathcal{D}$ conditionally to the previously obtained clustering $\mathcal{K}$. The first variables of this ranking are variables with a high score for the *compactness* and a low score for the *separation*. This approach is used as a feature selection mechanism in Section 6.5 and compared to the alternative feature ranking approaches introduced in the previous section. A summary of the main steps of this approach is proposed in Algorithm 6.1.

Algorithm 6.1 CS algorithm for feature ranking

Require: a set of observations $\mathcal{N}$, a set of candidate variables $\mathcal{D}$ and a set of clusters $\mathcal{K}$.

Ensure: a sorted list $CS(\mathcal{D})$.

1: Calculate $\mu_{i,j}$ and $\mu^C_{i,j}$ for each cluster i and for each dimension j.
2: Calculate the *compactness*, $Comp_j$, of each dimension j.
3: Calculate the *separation*, Sep_j, of each dimension j.
4: Create a list LC by sorting, in an increasing order, the dimensions of $\mathcal{D}$ according to their values of $Comp_j$.
5: Create a list LS by sorting, in a decreasing order, the dimensions of $\mathcal{D}$ according to their values of Sep_j.
6: Calculate the different positions, $pos(LC, j)$ and $pos(LS, j)$, of the dimensions of $\mathcal{D}$ in the two lists LC and LS.
7: Create a list $CS(\mathcal{D})$ by sorting, in an increasing order, the dimensions of $\mathcal{D}$ according to their values of $pos(LC, j) + pos(LS, j)$.

6.5 Application

In this section, an application in the concert industry is reported and discussed in detail in order to guide the reader through the different steps taken by the business and researchers involved in this project. The business consists of members from the marketing department of a main event organizer based in the Netherlands that collaborated with the authors in order to better understand their customer base and obtain actionable segments that would be integrated in their marketing strategy. After obtaining a segmentation of their customers based on meaningful variables, new variables are created and evaluated in order to be incorporated into the existing segmentation. This section consists of two main sub-sections. In Section 6.5.1, the different steps leading to the original segmentation of the customer base are presented in detail. Section 6.5.2 describes how new relevant variables can be identified in order to complete and improve the segmentation obtained in Section 6.5.1. Note that the different concepts and techniques used in this section are introduced in Sections 6.2, 6.3 and 6.4.

6.5.1 Original segmentation of the customer base

The customers of the company involved in this application, called *EventOrga* in the remainder of this chapter, are interacting with the company through different channels. Thanks to the collaboration of different technology providers, these interactions enable the capture of data that further can be used to understand the customers. In this section, the different steps leading to the creation of segments of customers are described. This segments are currently used by *EventOrga* in order to guide their marketing efforts on a daily basis.

Two main sources of data are used in order to create the variables based on which customer segments are formed. The first data source consists of six years of ticketing data gathering the logs of the transactions made by the different customers. Demographic data about the customers and transactional data at the level of tickets are captured and aggregated at the customer level in order to fuel this database.

Note that a customer, in this context, is the person buying the tickets and is not especially the person attending the concert. However, in this chapter, it is said that a customer attended a concert if he bought a ticket for it and the concert took place in the past. A second data source consists of a product database in which the characteristics of the different products (events in this case) are stored. Because the transactional logs are linked to specific events, both databases can be combined in order to create rich and meaningful variables for the business. The way these variables are calculated is shortly explained in what follows. First the well known RFM variables (*Recency*, *Frequency* and *Monetary*) are calculated [90]. The variable *Recency* is obtained by calculating the delta, in days, between the last concert attended by a customer and the moment of the calculation of this variable. The output of this variable is a number of days. The variable *Frequency* corresponds to the weighted number of concerts attended by a customer. The weights are business input assigned to the different seasons[1] and reflect the fact that a customer with a high frequency of visits during the last season should have a higher value for *Frequency* than a customer with the same frequency of visits but during an older season. The output of this variable is a positive number reflecting the weighted frequency of visits of a customer. The variable *Monetary* is calculated as the average money spent by season. The output is an amount in euro. Three variables (*OnlyFuture*, *FutureEvent* and *Trend*) linked to the RFM variables are then calculated. The variable *OnlyFuture* is equal to 1 if a customer purchased tickets only for future events (the customer is in the database but his associated events are in the future) and is equal to 0 otherwise. The variable *FutureEvent* is equal to 1 if a customer has tickets for at least one future event and is equal to 0 otherwise. The variable *Trend* is obtained by dividing the number of active seasons[2] of a customer by the length of his relationship, expressed in seasons, with *EventOrga*. The next variable, *DayDelta*, captures the average number of days separating the pur-

[1] A season is a period of 12 months starting on the 1^{st} of August.

[2] An active season is a season during which a customer attended at least one event.

Variable	**Range**	**Unit**
Recency	$[0, \infty)$	days
Frequency	$(0, \infty)$	weighted frequency
Monetary	$[0, \infty)$	euro
OnlyFuture	$[0, 1]$	dummy
FutureEvent	$[0, 1]$	dummy
Trend	$(0, 1]$	ratio
DayDelta	$[0, \infty)$	days
Distance	$[0, \infty)$	km
FirstLog	$[0, \infty)$	years
NbTickets	$(0, \infty)$	number of tickets
SubscriptionRatio	$[0, 1]$	ratio
ExternalRatio	$[0, 1]$	ratio

Table 6.1: Summary of the different variables used as input for the original segmentation, together with their ranges and units.

chase of a ticket for an event and the event itself. The output is a number of days. The variable *Distance* captures the geodesic distance in kilometers separating the customer (based on his address) and the event infrastructure. The variable *FirstLog* expresses, in years, the duration of the relationship between the customer and *EventOrga*. The variable *NbTickets* is calculated as the average number of tickets per event, the output being a number of tickets. Finally, two variables representing ratios are calculated. The variable *SubscriptionRatio* is calculated as the number of tickets purchased using a subscription divided by the total number of tickets purchased by a customer while the variable *ExternalRatio* captures the ratio of external concerts[3] to the total number of concerts attended by a customer. The different variables, their ranges and units are reported in Table 6.5.1.

After defining the different variables, a period of time is selected

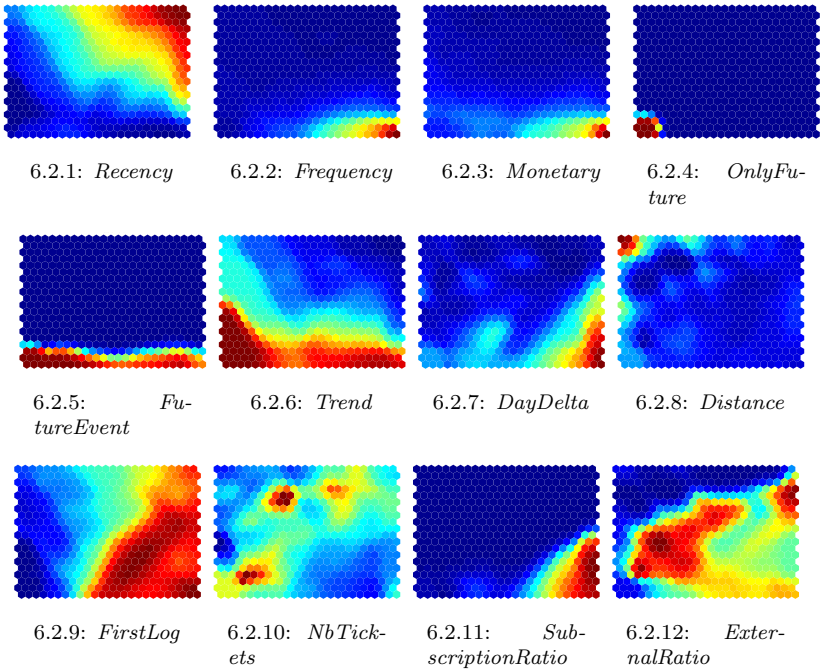
[3]An external concert is a concert programmed by an external organization but taking place in the infrastructure of *EventOrga*.

and data is gathered in order to calculate the values of the different variables for the customers active during this period. For this segmentation, a period of six years has been selected by the business based on operational considerations related to the quality of the data. The different data sources necessary to create the variables of Table 6.5.1 have been collected and cleaned using statistical tools and the business expertise. Note that the removal of outliers in the database has been a very important preprocessing step for this analysis. Although statistical tools can support this process, collecting experts' insight is a crucial step in the preprocessing phase. For example, after analyzing the available data, it has been decided to remove some of the customers because they were representatives from companies and hence showing unusually high RFM values. Deciding whether or not to keep these customers is a subjective but important step in the quest of value-adding segments and depends on business expectations and goals. After preprocessing, a total of 82082 customers are selected. These customers are customers active during the six last years. The variables of Table 6.5.1 are calculated for each of them by only using data of this period, except for the variable *FirstLog* for which older data sources are also considered in order to better reflect the reality. As the objective of this section is to create segments within the customer base (i.e. within the 82082 customers previously obtained), groups have to be identified such that customers within a group are relatively similar to other customers in that group while being relatively dissimilar to customers in other groups. To do so, two clustering steps are performed as described in what follows.

The first clustering step consists of the application of the SOM algorithm in order to summarize the 82082 customers characterized by the variables of Table 6.5.1 after standardization. The goal is to summarize the input data by using a relatively high number of neurons, allowing data summarization, visualization and noise removal. By applying this first step, a first exploration of the data is possible as illustrated in Figure 6.2 where the component planes of a 20×25 SOM are depicted. The choice of the number of neurons and the shape of the maps are following best practices. It is indeed recommended to

use a number of neurons significantly higher than the expected number of clusters while preferring a rectangular shape in order to facilitate the projection [17]. This explains why a 20×25 SOM, instead of e.g. a 10×10 SOM, is chosen. Each component plane of Figure 6.2 represents the values for the different neurons of the respective variable using a color code ranging from dark blue (low values) to dark red (high values). For example, by referring to Figure 6.2.1, the component plane of variable *Recency*, it can be said that neurons at the top right corner of the map are representing neurons, hence prototypes of customers, with relatively high values for the variable *Recency*. By combining multiple component planes, different conclusions can be drawn. For example, by referring to the bottom right corners of Figures 6.2.2 and 6.2.3, it can be said that some neurons, hence a group of customers, have a specific characteristic being relatively high values for the variables *Frequency* and *Monetary*. By identifying such visual patterns, the analyst is able to get first insights at an early stage of the segmentation, allowing him to perform, if necessary, some corrective actions. Although such visual analysis can be useful to make a first exploration of the data by e.g. visually identifying outliers or potential clusters, a second clustering step is necessary to obtain a manageable number of clusters that can be described and reported statistically. Note that this first clustering step allows the analyst to continue the analysis with 500 neurons (20*25) instead of the 82082 customers, which eases both manipulation and computational efforts during the second clustering step as mentioned in the theoretical section. To summarize, the customers are first represented by a high number of neurons that are further clustered.

As a second clustering step, the neurons trained with the SOM algorithm are used as input for the k-means algorithm (note that the weights of the SOM output are already based on standardized variables such that no standardization or normalization is performed for this second clustering step). The maximum number of clusters, k^{MAX}, is set to 20 based on business constraints and first insights obtained during the visual exploration using the component planes. Using the Davies-Bouldin index to select the best partition, the 20 clusters (c_1



6.2.1: *Recency* 6.2.2: *Frequency* 6.2.3: *Monetary* 6.2.4: *OnlyFuture*

6.2.5: *FutureEvent* 6.2.6: *Trend* 6.2.7: *DayDelta* 6.2.8: *Distance*

6.2.9: *FirstLog* 6.2.10: *NbTickets* 6.2.11: *SubscriptionRatio* 6.2.12: *ExternalRatio*

Figure 6.2: The 12 component planes obtained after applying the SOM algorithm on the original variables.

to c_{20}) of Figure 6.3 are selected. The centroids, or centers, of these 20 clusters are reported in Figure 6.4. The color code used reflects the relative values within each variable of the different clusters; the cluster having the highest value for a given variable being colored in dark green while the lowest value is represented in dark red. Note that cluster c_{12} has no value for the variable *Recency*, which makes sense since the value of the variable *OnlyFuture* is equal to one, meaning that this cluster captures customers whose first events are in the future, such that the *Recency* value can not be calculated. The number of customers and the percentage of the customer base it represents are also reported, thanks to a mapping between customers and neurons and between neurons and clusters. It is indeed important to note that neurons are prototypes of customers and clusters are groups of neurons. It is thus possible to link the customers to their BMUs (closest neurons) and these BMUs to their closest clusters, hence linking the clusters back to the customers. The centroids of Figure 6.4 are representing the average values per cluster of the different variables of Table 6.5.1. Although the second clustering step is based on the neurons, the values reported are the averages at the customer-level in order to better reflect the reality. The different outputs of this section are used in the daily practice of *EventOrga*, especially Figure 6.4 which provides them with an advanced and detailed segmentation output based on meaningful variables.

Different next steps can then be considered by *EventOrga* in order to build on this first exercise and better understand their customers. On the one hand, other data mining techniques as dynamic clustering approaches (see e.g. [75] and [76]) or recommender systems can be applied and combined with the output of this segmentation, hence using it as a new knowledge source for future analyses. On the other hand, the current segmentation could also be enriched by considering new data sources while building on the original segmentation as illustrated in the next section.

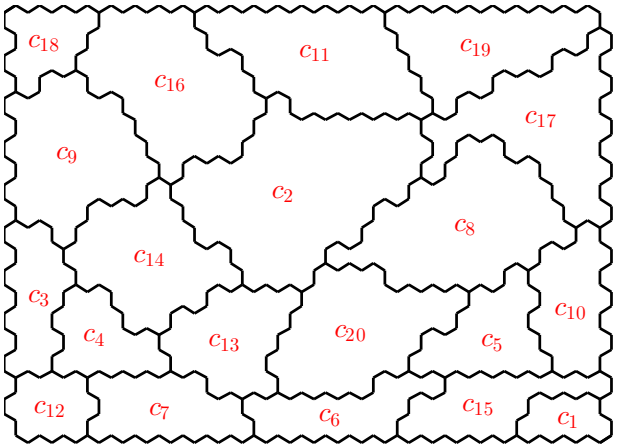


Figure 6.3: The 20 clusters obtained after applying the k-means algorithm on the output of the SOM algorithm.

6.5.2 Identification of new variables

In this section, the techniques proposed in the theoretical sections are applied in order to identify variables to add to the existing segmentation. As it is often the case in practice, a segmentation task is based on a set of variables considered as relevant at the beginning of the analysis. Although these variables may remain relevant in the future, new variables could be later considered to enrich the segmentation of the customers. This situation can be approached in two different ways. A first approach could consist of a new segmentation based on both sets of variables. However, while this approach is typically considered in practice, the method proposed in this chapter aims at ranking the new set of variables based on how it could enrich the original set of variables and improve the segmentation results. The main advantage of this approach is that it allows the practitioner to continue with the original variables while considering the original segmentation as a prior knowledge guiding the selection of the new variables. By doing this, even if new segments of customers are obtained, these segments are aligned with the original segmentation since they are constructed based on the previously generated knowledge. Aligning the second seg-

	Recency	Frequency	monetary	OnlyFuture	FutureEvent	Trend	DayDelta	Distance	FirstLog	NbTickets	SubscriptionRatio	ExternalRatio	# customers	%
c1	61.55	205.57	549.95	0.00	0.98	0.99	259.67	15.89	6.73	2.45	0.79	0.65	2425	2.95
c2	887.39	5.34	26.52	0.00	0.00	0.32	35.14	23.44	2.83	2.42	0.00	0.93	7058	8.60
c3	68.18	7.52	81.58	0.00	0.00	0.99	41.62	41.47	0.37	2.02	0.01	0.03	2494	3.04
c4	105.04	11.91	99.69	0.00	0.00	0.96	32.45	19.10	0.87	3.21	0.02	0.88	1199	1.46
c5	430.53	47.68	136.22	0.00	0.00	0.67	149.35	15.92	5.88	2.15	0.58	0.49	1827	2.23
c6	213.29	68.07	193.21	0.00	0.85	0.87	124.67	17.29	6.52	2.28	0.12	0.53	977	1.19
c7	303.90	27.47	117.78	0.00	1.00	0.88	121.06	21.35	2.02	2.52	0.14	0.33	1610	1.96
c8	859.72	8.15	32.33	0.00	0.00	0.39	48.46	11.87	6.33	2.28	0.01	0.28	5041	6.14
c9	316.91	6.25	40.02	0.00	0.00	0.51	43.02	17.59	1.12	2.23	0.00	0.01	6308	7.68
c10	1020.35	22.60	77.95	0.00	0.00	0.38	283.67	21.93	5.65	2.29	0.88	0.39	4306	5.25
c11	1143.38	3.61	19.59	0.00	0.00	0.26	38.35	30.47	3.44	2.63	0.00	0.01	9045	11.02
c12		7.11	83.80	1.00	1.00	1.00	104.78	36.65	0.34	2.88	0.03	0.10	1881	2.29
c13	423.99	12.01	96.58	0.00	0.00	0.54	200.21	34.01	2.72	2.64	0.01	0.92	1591	1.94
c14	333.62	7.27	45.72	0.00	0.00	0.52	30.72	22.14	1.23	2.34	0.01	0.94	3936	4.80
c15	125.43	103.30	257.54	0.00	0.84	0.91	237.46	17.67	5.74	2.13	0.70	0.49	1840	2.24
c16	691.74	5.02	31.22	0.00	0.00	0.35	42.65	15.15	2.21	3.39	0.00	0.01	7537	9.18
c17	1512.36	2.77	16.16	0.00	0.00	0.22	69.81	20.27	5.31	2.54	0.00	0.84	5860	7.14
c18	599.03	5.79	39.79	0.00	0.00	0.42	51.17	119.67	2.04	2.13	0.01	0.16	4007	4.88
c19	1556.83	1.90	11.98	0.00	0.00	0.19	27.46	17.08	4.84	2.39	0.00	0.00	9043	11.02
c20	360.46	23.21	78.32	0.00	0.00	0.68	50.10	14.25	6.26	2.26	0.02	0.43	4097	4.99

Figure 6.4: The centroids of the 20 clusters.

mentation with the first one while bringing new insights is believed to help the business to continue to follow the initiated strategy. Note that the goal of this section is not to provide the practitioner with a unique best segmentation but to provide a way to enrich a segmentation in a meaningful way. If new insights are generated by making a new segmentation, it is still important to keep the original segmentation in mind and consider the new segmentation as a complement and not as replacement. In order to illustrate the relevance of this problem, the case of the marketing team of *EventOrga* is considered. After a successful data mining exercise, the team obtained the 20 segments of Section 6.5.1 based on the 12 variables of Table 6.5.1 and decided to integrate it into its daily strategy and practices. As more and more data sources are getting available, new variables relevant for the business are identified and calculated after some months, which raises the question of how to integrate them considering the current segmentation. In what follows, different techniques discussed in this chapter are applied in order to propose a solution to this problem by ranking a set of candidate variables conditionally to a segmentation based on an original set of variables. Since, to the best of the knowledge of the authors, this problem has never been approached or discussed in the literature, state-of-art existing techniques (introduced in Section 6.3) traditionally used in other contexts are assessed with regard to their ability to solve this problem. Additionally, a technique (proposed in Section 6.4) designed for this specific problem is applied and compared to the other alternatives. The best strategy to adopt for this application is then selected and the candidate variables are ranked, allowing, inter alia, a new segmentation of the customers. These different steps are reported in detail in what follows, starting with a brief description of the candidate variables.

By making use of data from the product database, the preferences of the customers concerning the day and the genre of the events can be captured and new variables, the candidate variables, can be created. 14 variables capturing the preferences concerning the day of the events are first created. Firstly, the day of the concert is exploited and 7 variables, *MondayCount*, *TuesdayCount*, *WednesdayCount*, *Thurs-*

dayCount, *FridayCount*, *SaturdayCount* and *SundayCount*, are calculated at the customer level and represent the number of concerts, attended by the customer, that took place on the respective days. Based on these 7 variables, 7 other variables, *MondayRatio*, *TuesdayRatio*, *WednesdayRatio*, *ThursdayRatio*, *FridayRatio*, *SaturdayRatio* and *SundayRatio*, are obtained by dividing each of them by the total number of concerts attended by the customer, hence obtaining ratios. Secondly, the genre of the concert is exploited and 7 main genres are identified, namely FamilyConcert, Jazz, JazzWorldMusic, Classical, Popular, RedSofa and WorldMusic. Similar to the day of the concert, the genre of the concert is thus used to calculate 14 variables. The 7 first variables, *FamilyConcertCount*, *JazzCount*, *JazzWorldMusicCount*, *ClassicalCount*, *PopularCount*, *RedSofaCount* and *WorldMusicCount*, are calculated at the customer level and represent the number of concerts, attended by the customer, of the respective genres. 7 other ratio-variables, *FamilyConcertRatio*, *JazzRatio*, *JazzWorldMusicRatio*, *ClassicalRatio*, *PopularRatio*, *RedSofaRatio* and *WorldMusicRatio*, are then created. These 28 variables are reported in Tables 6.3, together with their respective ranges.

At this point, a segmentation based on the variables of Table 6.5.1 of the 82082 customers of *EventOrga* is considered together with the 28 variables of Table 6.3 which are calculated for the 82082 customers, also based on six years of data. Referring to the theoretical section, a set D of candidate variables and a prior segmentation K of the data points of a set N are created. The set D gathers the 28 variables of Table 6.3, the partitioning K is based on the 20 segments obtained in the previous section and the set N represents the 82082 customers. Looking for a set of new variables can then be achieved by ranking the variables of the set D conditionally to the clustering K. Since different techniques are available (the 7 techniques of Section 6.3 and the approach proposed in Section 6.4), multiple rankings of the candidate variables can be obtained as shown on Table 6.3 where the positions in the rankings of the 28 variables are reported for each of the 8 identified ranking approaches. A variable at the position 1 of the ranking obtained by applying a given strategy being the best candidate variable

Variable	**Range**	**Variable**	**Range**
FamilyConcertCount	$[0, \infty)$	*MondayCount*	$[0, \infty)$
JazzCount	$[0, \infty)$	*TuesdayCount*	$[0, \infty)$
JazzWorldMusicCount	$[0, \infty)$	*WednesdayCount*	$[0, \infty)$
ClassicalCount	$[0, \infty)$	*ThursdayCount*	$[0, \infty)$
PopularCount	$[0, \infty)$	*FridayCount*	$[0, \infty)$
RedSofaCount	$[0, \infty)$	*SaturdayCount*	$[0, \infty)$
WorldMusicCount	$[0, \infty)$	*SundayCount*	$[0, \infty)$
FamilyConcertRatio	$[0, 1]$	*MondayRatio*	$[0, 1]$
JazzRatio	$[0, 1]$	*TuesdayRatio*	$[0, 1]$
JazzWorldMusicRatio	$[0, 1]$	*WednesdayRatio*	$[0, 1]$
ClassicalRatio	$[0, 1]$	*ThursdayRatio*	$[0, 1]$
PopularRatio	$[0, 1]$	*FridayRatio*	$[0, 1]$
RedSofaRatio	$[0, 1]$	*SaturdayRatio*	$[0, 1]$
WorldMusicRatio	$[0, 1]$	*SundayRatio*	$[0, 1]$

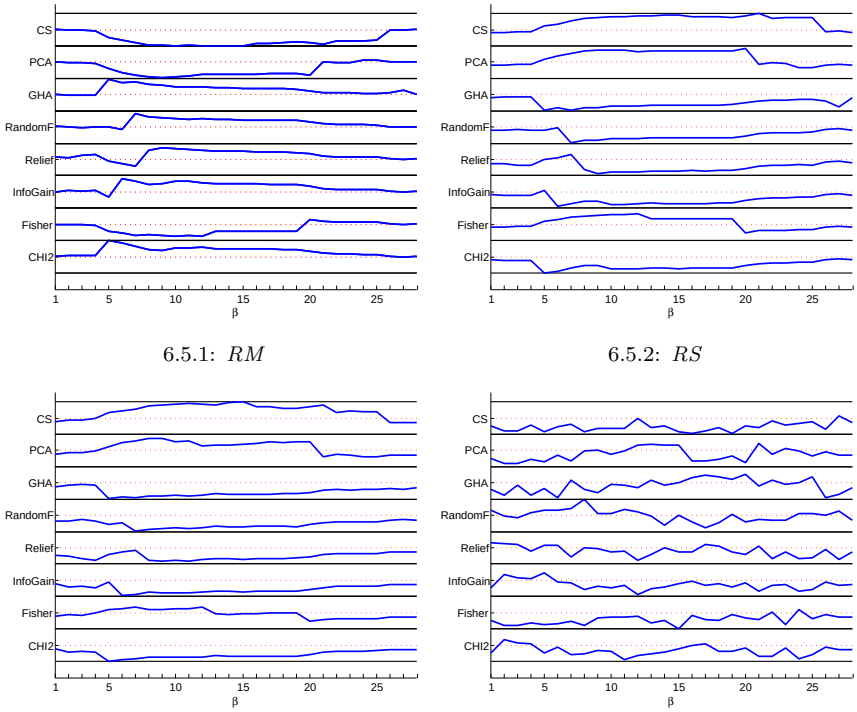
Table 6.2: Summary of the different candidate variables, together with their ranges.

according to this respective strategy.

As can be seen on Table 6.3, the different ranking strategies are leading to different results. In order to select the one to adopt, some experiments are conducted based on the results of Table 6.3. The purpose of these experiments is to evaluate the different strategies by measuring the quality of the segments obtained when following these respective strategies. Each strategy is thus considered to build a series of partitions by iteratively adding new variables from the set of candidate variables to the set of original variables while following the order of the respective rankings. 224 (8×28) subsets of the candidate variables are thus formed, 186 of them being unique, leading to 186 executions of the clustering approach discussed in Section 6.2. The four evaluation metrics of Section 6.2 are then applied on the different clustering results in order to assess their quality. Since the absolute values of the different metrics are difficult to interpret, a relative comparison of the different strategies is to be preferred. These comparisons are reported on Figures 6.5.1, 6.5.2, 6.5.3 and 6.5.4 where the relative values of the respective metrics are reported for each of the 8 strategies. As an example of the reading of such a plot, Figure 6.5.1 allows to compare the 8 ranking strategies using the RMSSTD index introduced in Section 6.2. For each of the strategies, a blue line represents the RMSSTD values obtained after adding a number of candidate variables $\beta \in [1..28]$ (represented on the abscissa) following the given strategy. The solid black lines represent respectively the maximum and the minimum values of the RMSSTD of the 186 clustering output, such that the respective RMSSTD curves evolve between them. The dotted red line is a visual marker equidistant from the minimum and the maximum helping the analysis. This allows a visualization and a comparison of the different strategies using relative values, hence avoiding a difficult interpretation of the absolute values. As an example, comparing the PCA and the RandomForest rankings using the CH evaluation metric leads, referring to Figure 6.5.3, to the straightforward conclusion that the PCA strategy generates better clustering results on this database since a high value of CH is preferable. However, making the same comparison using the DB index, hence referring to Figure 6.5.3,

	Chi2	**Fisher**	**InfoGain**	**Relief**	**Random Forest**	**GHA**	**PCA**	**CS**
FamilyConcertCount	24	15	24	23	23	4	11	8
JazzCount	26	23	26	26	15	20	14	11
JazzWorldMusicCount	27	27	27	28	21	11	26	12
ClassicalCount	1	1	1	18	18	2	1	1
PopularCount	19	14	16	22	20	3	16	13
RedSofaCount	23	16	25	25	11	24	9	6
WorldMusicCount	25	19	22	24	9	6	13	14
FamilyConcertRatio	20	21	19	10	8	16	17	23
JazzRatio	22	17	23	13	27	27	22	9
JazzWorldMusicRatio	28	28	28	27	22	8	28	15
ClassicalRatio	2	6	2	1	17	12	10	19
PopularRatio	12	8	9	3	12	28	12	16
RedSofaRatio	17	26	20	19	6	25	27	17
WorldMusicRatio	18	11	18	7	2	17	15	10
MondayCount	21	10	21	6	24	19	8	20
TuesdayCount	15	9	15	20	14	10	6	4
WednesdayCount	16	7	17	21	25	13	7	18
ThursdayCount	9	4	11	17	26	23	5	5
FridayCount	7	2	8	14	16	1	2	2
SaturdayCount	8	3	7	16	10	21	3	3
SundayCount	13	5	12	15	13	26	4	7
MondayRatio	14	18	14	12	4	9	20	21
TuesdayRatio	10	13	10	9	5	7	24	22
WednesdayRatio	11	25	13	11	28	15	23	24
ThursdayRatio	5	20	6	8	7	5	21	26
FridayRatio	4	12	4	2	3	14	19	25
SaturdayRatio	3	22	3	5	1	22	25	27
SundayRatio	6	24	5	4	19	18	18	28

Table 6.3: Position of each of the 28 variables for each of the 8 ranking strategies.



6.5.1: *RM* 6.5.2: *RS*

6.5.3: *CH* 6.5.4: *DB*

Figure 6.5: Relative values for the 4 evaluation metrics RMSSTD, RS, CH and DB for different values of β considering the 8 strategies.

is less straightforward since the curves are more similar.

In order to conclude on which strategy to adopt, one is thus looking for a strategy showing a superior behavior for the 4 different evaluation metrics (i.e. curves with higher values of RMSSTD, lower values of RS, higher values of CH and lower values of DB). Starting with the RMSSTD index, two strategies are showing relatively better performances (curves with higher values), namely the CS and the FCA. Looking closer on Figure 6.5.1, one could even say that the CS strategy is showing a relatively similar behavior to PCA until the first 20 new variables are added and performs slightly better for the 8 last variables until converging to the same point (which makes sense since

the final clustering is similar for all strategies because all the candidate variables are added). Referring to Figure 6.5.2, 3 strategies can be identified as performing better (higher RS values than the other strategies), namely the CS, PCA and Fisher Score strategies. As for the previous evaluation metric, a closer look tends to confirm that even if the curves are relatively similar when adding the first candidate variables, the CS strategy is performing equally or better when adding the last variables. Referring to Figure 6.5.3, the CS strategy clearly outperforms the other strategies, showing higher values of CH when adding new variables. Finally, when referring to the DB index of Figure 6.5.4, the CS and Fisher Score strategies seem to perform better than the other strategies, showing relatively lower values for the DB index. As a conclusion, the CS, PCA and the Fisher Score strategies seem to perform better on the data set at hand than the other 5 strategies. Looking closer and comparing these 3 strategies, the CS approach is showing similar or better results than the other strategies when analyzing the different evaluation metrics. Because of these results and the ease of calculation of the CS strategy, this approach is considered, for this case, as the strategy to adopt. Note that the experiments leading to the plots of Figure 6.5 are computationally expensive and are reported here in order to assess the proposed feature selection approaches by simulating the different possible outcomes. Other experiments with additional data sets may enrich this analysis but, given the business objective of this application, focusing on this data set will already allow us to make different decisions. For example, based on the selected strategy, the CS strategy, and referring to Table 6.3, it could be said that the ratios about the preferences for the different days are not relevant candidates to add to the original variables since they are ranked last according to the CS ranking. One could then decide to choose a set of β new variables to add to the existing segmentation by adding to the original variables the β variables ranked first when adopting this same strategy (using Table 6.3 for the rankings). Since the focus of this chapter is on providing a ranking of the candidate variables and not on providing an optimal set of variables to add, the parameter β is in this case an arbitrary parameter

fixed by the user. The reason behind this choice is that proposing a framework for finding an optimal value for β would be difficult to defend. It could be approached by considering the curve of a strategy for a given evaluation metric, as reported on Figure 6.5, and by trying to identify a value of β maximizing this quality criterion. However, the arbitrary parameter would then become the evaluation metric to choose. For this reason, the different strategies are objectively assessed using four well-known evaluation techniques, leading to the identification of the best strategy for the data at hand, while keeping β as the arbitrary parameter. The business can thus for example be interested in adding the 10 next best variables, such that β is set to 10. Following the CS strategy, the 10 variables *ClassicalCount*, *FridayCount*, *SaturdayCount*, *TuesdayCount*, *ThursdayCount*, *RedSofaCount*, *SundayCount*, *JazzRatio*, *FamilyConcertCount* and *WorldMusicRatio* are thus added to the original variables in order to form a new segmentation taking into account the previous one while aiming at obtaining a better segmentation of the customers. After running the clustering approach of Section 6.2 on the new input, 20 clusters are obtained including the original variables and the 10 new candidate variables. The centers of these clusters and the number and percentage of customers belonging to each of them are reported in Figure 6.6, the interpretation of which being the same as for Figure 6.4.

6.6 Conclusion

In conclusion, this paper contributes to both research and business societies by proposing and applying an approach allowing to identify next relevant variables for segmentation by using feature selection techniques. From a research point of view, a new problem is identified and formalized while discussing and unifying 7 candidate solutions issued from different domains. In addition, a new ranking approach is proposed based on fundamental criteria from the clustering discipline, adapted to provide a solution to the identified problem of ranking a new set of candidate variables conditionally to the structure of an ex-

	recency	OnlyFuture	frequency	monetary	FutureEvent	TrendVariable	DayDelta	Distance	FirstLog	NbTicketsConcert	RatioAbo	RPhORatio	FamilyConcertCount	ClassicalCount	RedSofaCount	JazzRatio	WorldMusicRatio	TuesdayCount	ThursdayCount	FridayCount	SaturdayCount	SundayCount	# customers	%
c1	407.31	0.00	84.57	211.19	0.18	0.72	228.14	16.98	6.30	2.18	0.67	0.41	0.96	22.73	0.83	0.01	0.02	2.71	2.11	3.99	8.44	6.58	983	1.20
c2	374.27	0.00	5.66	38.18	0.00	0.49	38.25	34.20	1.18	2.50	0.00	0.11	0.05	0.47	0.02	0.01	0.07	0.06	0.11	0.16	0.36	0.25	8781	10.70
c3	309.59	0.00	7.65	54.90	0.00	0.57	50.12	24.95	1.34	2.56	0.00	0.76	0.10	1.11	0.01	0.00	0.01	0.05	0.24	0.43	0.32	0.39	3831	4.67
c4	285.39	0.00	19.85	100.59	0.03	0.71	98.76	23.96	3.43	2.56	0.09	0.67	0.20	3.84	0.09	0.01	0.02	0.23	0.70	1.19	1.15	1.06	1898	2.31
c5	92.30	0.01	8.10	91.40	0.02	0.97	40.01	35.20	0.49	2.45	0.00	0.45	0.08	0.82	0.01	0.02	0.03	0.06	0.18	0.46	0.27	0.37	3209	3.91
c6	669.59	0.00	14.38	53.52	0.00	0.50	70.37	15.01	5.98	2.39	0.06	0.47	0.39	3.60	0.07	0.01	0.04	0.33	0.55	1.00	1.11	1.38	5573	6.79
c7	362.44	0.00	149.15	323.61	0.31	0.77	308.41	16.12	6.58	2.24	0.81	0.50	3.50	38.65	1.33	0.00	0.01	2.97	3.74	8.69	11.93	14.26	795	0.97
c8	1450.74	0.00	2.28	15.48	0.00	0.21	26.66	20.40	4.55	2.82	0.00	0.03	0.13	0.37	0.01	0.02	0.29	0.07	0.09	0.18	0.31	0.37	12074	14.71
c9	372.29	0.00	43.25	136.14	0.04	0.72	109.31	16.23	6.16	2.21	0.25	0.52	0.34	11.32	0.25	0.01	0.03	0.96	1.83	3.24	3.07	2.87	1547	1.88
c10	113.07	0.00	130.08	387.16	0.77	0.94	216.12	16.66	6.55	2.30	0.61	0.73	0.26	33.61	0.36	0.01	0.01	1.65	8.61	11.99	5.34	5.49	1853	2.26
c11	1358.15	0.00	3.67	19.09	0.01	0.25	60.76	17.47	5.50	2.62	0.02	0.54	0.27	1.22	0.01	0.01	0.01	0.10	0.16	0.33	0.39	0.60	6039	7.36
c12	136.94	0.00	71.49	230.67	0.83	0.89	176.72	19.68	5.31	2.40	0.42	0.56	0.35	15.29	0.37	0.02	0.03	0.96	2.47	5.15	3.70	3.46	1264	1.54
c13	706.67	0.00	5.84	31.94	0.00	0.37	55.65	24.90	2.62	2.61	0.00	0.67	0.14	1.11	0.02	0.00	0.01	0.07	0.23	0.38	0.32	0.39	5718	6.97
c14	914.76	0.00	4.13	22.22	0.00	0.29	42.35	32.61	2.67	2.69	0.00	0.09	0.13	0.52	0.02	0.01	0.07	0.08	0.12	0.19	0.33	0.36	13598	16.57
c15	274.19	0.00	6.41	51.56	0.00	0.63	55.53	72.11	1.12	2.14	0.00	0.10	0.02	0.40	0.02	0.44	0.02	0.05	0.24	0.32	0.26	0.30	2827	3.44
c16	175.30	0.00	143.14	323.57	0.55	0.88	226.42	15.48	6.62	2.18	0.70	0.33	0.34	36.98	1.40	0.00	0.01	5.77	2.92	5.21	15.95	7.23	1298	1.58
c17	325.13	0.05	30.59	121.75	0.93	0.82	120.96	22.55	3.46	2.61	0.17	0.37	0.61	4.77	0.17	0.03	0.03	0.48	0.83	1.22	1.48	1.80	2015	2.45
c18	1062.40	0.00	4.25	21.04	0.00	0.29	35.69	19.99	3.91	2.65	0.00	0.55	0.24	0.96	0.02	0.01	0.01	0.09	0.20	0.27	0.30	0.47	2819	3.43
c19	936.37	0.00	21.94	75.95	0.01	0.42	218.28	20.39	5.77	2.32	0.69	0.42	1.02	6.64	0.13	0.01	0.04	0.76	0.87	1.40	2.00	2.58	4079	4.97
c20	287.99	0.82	7.62	80.77	0.98	0.99	95.50	36.73	0.43	3.03	0.00	0.11	0.07	0.53	0.00	0.03	0.01	0.31	0.16	0.19	0.20	0.30	1881	2.29

Figure 6.6: The centroids of the 20 clusters obtained after adding 10 of the candidate variables to the original variables.

isting segmentation with as objective to improve the quality of the resulting partition. In order to evaluate this quality, four evaluation metrics widely used and discussed in the literature are discussed and summarized. Experiments are then conducted in order to assess the different ranking strategies and are reported in detail. From a business perspective, a detailed real-life segmentation exercise is reported using a recent case from the concert industry while sharing insights collected during this application. After describing the different steps and output leading to the original segmentation of a customer base, the operational problem of how to update this segmentation by adding new variables is studied. The different techniques introduced in this paper are therefore combined in order to identify the best strategy to adopt given the problem and data set at hand. A ranking of a set of new variables is then obtained using the identified best strategy, allowing to discriminate irrelevant variables and to propose a new segmentation of the customer base enriched with a set of relevant new variables.

The main implication of applying this approach is that a new segmentation can be obtained conditionally to a previously obtained segmentation by adding new relevant variables. Practitioners should consider this approach if they face a situation in which a recent segmentation is already integrated and used in daily practices and new variables are available. This approach should then be preferred to a traditional re-clustering since the knowledge present in the first segmentation will be used to guide the new segmentation. In other words, since the current practices are based on the current segmentation, guiding the new segmentation using the existing knowledge will also guide the new practices by valuing existing ones, easing their acceptance by the business.

Finally, these experiments and associated results open new tracks for research since:

- new ranking techniques could be developed or compared using the same setup,
- new evaluation metrics for clustering could be added,

- new questions could be answered as how to identify the optimal value for the parameter β and
- similar experiments could be conducted using other data sets.

Note that the data, output and problems mentioned in this paper are based on a real case involving a leading concert organizer based in the Netherlands. We extend our gratitude to Mathijs Bouwman for his input as business expert, which significantly increased the value of this work.

Chapter 7

Impact on business

Two different sets of requirements have been considered while conducting the different projects reported in this document. The first set of requirements came from the scientific world, since each project was conducted with the objective to contribute to the scientific community by generating high quality publications targeting international peer reviewed journals. Considering this as the main objective of this work, and because scientific requirements were aligned with the academic requirements of a PhD, the operational business considerations and implications of the respective projects may not be properly highlighted. In order to cope with this shortcoming, this chapter will highlight the different business implications of the projects conducted during this PhD journey, allowing the reader to better understand the second set of requirements: the business requirements.

In what follows, the main business implications and lessons learned from the five projects respectively reported in Chapters 2, 3, 4, 5 and 6 are summarized.

- ***A new SOM-based method for profile generation: theory and an application in direct marketing***

The project leading to the work reported in Chapter 2 was a kick-off project initiated in collaboration with Ticketmatic in order to assess the opportunities of data analytics in a ticketing context. Fueled

by the data of a Belgian concert organizer collected Ticketmatic-side, extensive experiments were conducted. Unconstrained regarding the scope of our analysis, we investigated both unsupervised and supervised techniques and proposed a stepwise methodology allowing us to illustrate interesting functionalities. The data summarization and visualization offered by the self-organizing maps combined with a second step clustering on top of the SOM output proved to be of great value for the business. It quickly triggered the interest of other customers of Ticketmatic and we decided to use this approach as a necessary first step in the following analyses.

From a usability perspective, it allowed business users with no or limited background in data analytics and statistics to explore large amounts of data while drawing interesting conclusions about the structure of it. Allowing unexperienced users to get quickly in touch with new insights about the data previously unused is a key achievement of this first project. It opened new tracks for research and triggered the interest of other business partners. Capturing the visual patterns using a second clustering step, a transparent segmentation of the customers was a first added value of the methodology used in this project. As a next step in the methodology, we investigated the use of the resulting segments in order to generate profiles targeting specific characteristics of the customers. The added value of it was twofold. On the one hand, we illustrated a way to build on the knowledge previously obtained during the segmentation. This concept of incorporating available knowledge in the analysis has been of great help for the acceptance of the different techniques. Moving stepwise while creating value at each step allowed customers with different maturities to consider the proposed methodology, less mature customers limiting themselves to the first steps of the approach. On the other hand, we were able to answer the question of the impact of the quantity of data available on the analysis by studying the performance of our profile generator. Even if the conclusions drawn about the quantity of data are limited to the dataset at hand, it opened the mind of Ticketmatic regarding topics as data quality, availability and volume which are key for them given their position of data provider.

After this first project, the focus of the research was changed from data analytics in general to a more specific study of the operational challenges related to the implementation of successful segmentation practices and associated topics.

- ***A new knowledge-based constrained clustering approach: theory and application in direct marketing***

The second project started in order to cope with some business limitations of traditional segmentation algorithms. The idea of a traditional "blind" clustering was thus opposed to a "business-guided" constrained clustering approach, developed and illustrated using the same dataset than the one explored during the first project. The main purpose was to modify traditional clustering algorithms to allow the users to incorporate prior knowledge in the segmentation task. By doing this, the hypothesis was that the subjective quality of the resulting partition would be increased. Although using prior knowledge to guide clustering algorithms was not a new field in the research world, we managed to propose a way to incorporate a source of knowledge which was new and made sense to the business partners we were working with. The idea of capturing the a priori knowledge about the relative importance of the variables selected for a segmentation task was proposed and formalized during this second project. The strength of this approach, from a user perspective, was the natural and simple way to input the prior knowledge. The user would just have to assign priorities to the variables prior to the clustering task and weights would be generated to bias the perception of the algorithm. Asking different business partners about the relevance of such a guided clustering approach in their respective business contexts, we came to the conclusion that this prior knowledge was almost always present when preparing a task but was never used as input. In order to build on the previous knowledge and training of the partners involved, we decided to apply this strategy using the SOM algorithm as baseline, again helping the acceptance and understanding of the new approach.

As a logical consequence of the artificial and subjective constrain on a task, we analyzed the impact of such a guidance on the quality

of the resulting output. In order to properly assess it, we decided to differentiate the subjective quality from the objective one. This idea of subjective quality reflected the importance in the structure of the obtained clustering of the variables considered as more important by the user, hence introducing a quality metric depending on the objectives of the user. Surprisingly or not, the business users understood this idea of subjective quality even better than traditional objective metrics. Preforming experiments on the data at hand, we managed to show that constraining the clustering task by using the extra prior knowledge was significantly improving the subjective quality while only slightly reducing the objective quality.

The intangible added-value of this project was the operationalization of existing techniques by allowing the user to interact and guide the algorithms. This idea of using and valuing the prior knowledge of the user was thenceforth considered as a must and contributed to the acceptance of the proposed clustering strategy as a way to improve the understanding of the customers of the partner involved.

- ***Business knowledge based segmentation of online banking customers***

For this project, the interest triggered by the constrained clustering approach crossed the borders of the ticketing industry and led us to apply it in a banking environment in which analysts were trying to understand their online customers. The setup of their problem was very similar to the situations faced in the marketing departments of the customers of Ticketmatic since a subset of the variables considered for the segmentation was perceived by the business as more important than the rest while considering the rest as important enough to be kept in the analysis. Using data collected by the customer intelligence department of the bank combined with data available at the corporate level, we designed two segmentation exercises. The first segmentation was considered as the quantitative segmentation while the second, called the qualitative segmentation by the business, took the prior knowledge about the importance of the variables into account. Although the original idea was to compare both approaches in order

to identify the most suitable for the business involved, we came to the conclusion that both segmentations were relevant for the business. While the quantitative segmentation gave a general segmentation at a corporate level, the qualitative one allowed to enrich the general patterns and insights by focusing more on the variables captured by the department of interest. Considering multiple segmentations as the solution, instead of trying to obtain the best unique segmentation, helped us to achieve valuable results for the bank and inspired us in our research for value creation.

Following this project, we propagated this concept of a segmentation being dependent on the user and his personal objectives as a complement to traditional segmentation practices and received positive feedback from our business partners. Changing this belief that a customer base should be uniquely segmented was a great achievement from a business perspective since marketers were now considering the clustering algorithms as tools used on a daily basis to take pictures of their datasets from different angles and with different magnitudes of zoom.

- ***A dynamic understanding of customer behavior processes based on clustering and sequence mining***

A natural next step following these static pictures of a data set was the making of a movie by introducing the time dimension in the analysis. This need for understanding the dynamics associated with a customer base emerged when considering the fact that customers characterized by some dynamic attributes may soon or later move from one segment to another. This consideration triggered the interest of the users and two main directions were identified. First, a strategy was necessary to deal with the consequences of a segmentation involving dynamic attributes. Business partners started indeed to raise questions concerning the update of the knowledge captured through the segments. The most simple approach consisted of starting a new segmentation with a fresh data set after a certain period of time, keeping or not the original attributes. From a purely practical point of view, this approach could have been considered as the

best way to cope with this situation since the necessary infrastructure, techniques and skills were available. However, such an approach would not take into consideration the previous segmentation and would neglect the work already done and the business understanding of the segments. As a consequence, we jointly decided to strive for operational solutions keeping former segmentations into account such that movements between segments could be traceable at a customer level. While this approach would require to wait for pictures of the customer base in the future, the other direction consisted of an understanding of the historical data leading to the current segmentation. This second approach was at the origin of this fourth project, in which historical data is used to construct and understand trajectories leading to segments showing characteristics of interest for the user. By proposing mechanisms allowing to identify trajectories from the past, we aimed at triggering the interest of the users concerning the problematic of these high dimensional movements. Reaffirming our ambition to build on previous knowledge, we integrated the time dimension in techniques our users already mastered, namely the SOM and P-SOM algorithms combined with a second clustering step.

The impact of this project on the practices of our business partner is mainly linked to the opening of a new track for future analyses. Since such an analysis requires historical data or, at least, multiple snapshots of a data set, a clear need for structured data with timestamps was thus identified. Analyses that were previously conducted without taking the time and movements into considerations have now evolved and a need for exploratory tools in high dimensional spaces has been confirmed. Together with these partners, both directions of the analysis of the dynamics of their customers are currently being considered, allowing us to prepare new best practices while identifying next challenges.

- ***Identifying next relevant variables for segmentation by using feature selection approaches***

In the last project reported in this manuscript, a new operational problem is addressed. This project involves data collected by the main customer of Ticketmatic and deals with the challenge of identifying

next relevant variables for segmentation. After a successful segmentation exercise performed in collaboration with the marketing manager of this customer, a new strategy based on the insights of the obtained segments was applied in their daily practices. This segmentation was based on a set of variables identified during long meetings and discussions with the business. Because this set of original variables evolved during the segmentation process, different iterations were necessary before converging to a solution satisfying the different business experts. Once obtained, this set of variables was considered as final and a segmentation was obtained. However, this final aspect was linked to a context, a task and the availability of the data at this specific moment. What if new variables would later become available? In order to answer this question, we designed a straightforward approach to rank new variables according to their relevance to an existing segmentation. In this project, the goal of the selection was to enrich the existing segmentation without disturbing too much its structure. The relevance of a variable was then the extent to which this variable was correlated with the existing structure. On the other hand, in another business context, one may be interested in the identification of variables questioning the actual structure of the clustering output, hence not being correlated with it.

The lessons learned during this project are related to the previous project, in which the time dimension was considered. In this case too, a dynamic aspect is studied by considering the evolution of the data sources through time. Helping the business to think about their decisions and insights as time dependent concepts was a significant contribution of this project. By understanding this, this business partner is now planning long term experiments and analyses while considering the aspects related to the maintenance and update of the knowledge as key elements.

Together with this new approach to rank candidate variables, the original segmentation process is described into detail. It consists of a scientific summary of an article written by the business, using a business terminology. This article (in the last phase of publication at the moment of the elaboration of this manuscript) is probably the major

contribution of this PhD from a business perspective. Not because of the scientific value of it, but because it is written by a business expert after a successful data mining project. The idea behind this article is to share the experience collected during a series of projects while highlighting the main business implications of the decisions made. As a response to this article, other customers of Ticketmatic are starting the journey that may lead them soon or later to monetize their data.

Chapter 8

Concluding notes and future work

In this document, a subset of the output generated during the PhD of Alex Seret is reported. Five main chapters are reported since they form a coherent body. In Chapter 2, an exploratory methodology combining both descriptive and predictive techniques is presented. This work is an initial exploration of the possibilities linked to the usage of data mining on the data collected by Ticketmatic. In the third chapter, the focus is put on how to guide the clustering algorithm in order to obtain better perceived results. The link with Chapter 2 is clear since the techniques proposed in Chapter 3 are designed to reduce the limitations of the purely unsupervised algorithms used in the first steps of the methodology of Chapter 2. In the fourth chapter, the techniques proposed in Chapter 3 are applied in a marketing context of another industry. It shows how business knowledge can be used to offer new perspectives on data sets. Chapter 5 reports a logical next step following a first segmentation. By putting the attention on the dynamic aspects while reusing the approaches presented in Chapter 3, the fourth chapter opens new tracks on how to build on the knowledge previously generated. Finally, Chapter 6 proposes and evaluates feature selection approaches allowing to identify next relevant variables conditionally to an existing segmentation. This chapter is a first step

approaching one of the operational challenges linked to the need to update and enrich existing models.

Other operational challenges have been identified during the redaction of this manuscript and may require additional efforts in a future work. These different research tracks are briefly introduced in what follows, following the outline of this manuscript.

- In Chapter 2, an attempt to combine predictive and descriptive techniques is proposed and used in order to assess the impact of the size of the dataset and the obtained segments on the predictive task. Although the focus of this thesis is on descriptive approaches, another ongoing project aims at evaluating the opportunities that could be offered by recommender systems combined with the knowledge embedded in a segmentation. To do so, different collaborative filtering methods have been applied on the data collected by Ticketmatic. Although offline results have already been obtained and discussed in different reports, the results of the first online tests are only being collected at this stage. Guiding recommender systems with knowledge issued from the different segmentation exercises would be an interesting way to calculate the ROI of such segmentation exercises, which triggered a deep interest from the different business partners.

- In Chapter 3, an approach allowing the incorporation of existing knowledge as input of a data mining task is proposed, discussed and illustrated in a marketing context. On the one hand, future research may investigate how additional knowledge sources could be used in order to guide the analysis. On the other hand, other data mining techniques may be tuned in order to allow such a practice. Allowing the user to guide and interact with the algorithms is definitely an interesting topic that may benefit analysts of different industries, as illustrated in Chapter 4 were a priori knowledge is used to guide a segmentation task in a banking context.

- Another research track results from the conclusions of Chapter 5, in which the challenging task of understanding trajectories in

high dimensional spaces is discussed. The study of the possibilities that fuzzy logic could offer in the quest of understanding trajectories in high dimensional spaces has been identified as a topic for future research. As discussed in Chapter 5, understanding movements in high dimensional spaces remains a challenge. New techniques allowing the interpretation of such trajectories are necessary and we believe that fuzzy approaches could help for this purpose.

- Finally, the Chapter 6 opens a myriad of new research tracks by approaching one of the operational challenges related to the updating and enriching of existing models and insights. Since more and more companies are investing in analytics, the need for next steps and monitoring best practices will be of crucial importance in the coming years. Managing and understanding the evolution of the patterns and models will be key in order to obtain a long term monetization of the available data sources. Future research should start by identifying these operational challenges and understand their interrelationships.

List of Figures

List of Tables

Bibliography

[1] P. Kotler, K. L. Keller, B. Dubois, and D. Manceau, *Marketing Management*, 12th ed. Prentice Hall, 2006.

[2] F. T. Nobibon, R. Leus, and F. C. R. Spieksma, "Optimization models for targeted offers in direct marketing: Exact and heuristic algorithms." *European Journal of Operational Research*, vol. 210, no. 3, pp. 670–683, 2011.

[3] B. Baesens, S. Viaene, D. Van den Poel, J. Vanthienen, and G. Dedene, "Bayesian neural network learning for repeat purchase modelling in direct marketing," *European Journal of Operational Research*, vol. 138, no. 1, pp. 191–211, 2002.

[4] B. Baesens, G. Verstraeten, D. Van den Poel, M. Egmont-Petersen, P. Van Kenhove, and J. Vanthienen, "Bayesian network classifiers for identifying the slope of the customer lifecycle of long-life customers," *European Journal of Operational Research*, vol. 156, no. 2, pp. 508–523, 2004.

[5] S.-Y. Kim, T.-S. Jung, E.-H. Suh, and H.-S. Hwang, "Customer segmentation and strategy development based on customer lifetime value: A case study," *Expert Systems with Applications*, vol. 31, no. 1, pp. 101–107, 2006.

[6] J.-H. Lee and T.-C. Li, "Supporting user participation design using a fuzzy analytic hierarchy process approach." *Eng. Appl. of AI*, vol. 24, no. 5, pp. 850–865, 2011.

[7] S. Li, "Research about e-commerce customer value classification based on affiliations cloud clustering algorithm," in *Management and Service Science (MASS), 2011 International Conference on*, aug. 2011, pp. 1–4.

[8] Z. Yong, Z. Hao, C. Zheng, Z. Pingyuan, and X. Yating, "Cluster analysis of the consumption value based on the clothing market," in *Management Science and Industrial Engineering (MSIE), 2011 International Conference on*, jan. 2011, pp. 537–541.

[9] B. Xing and W. Xin-feng, "The evaluation of customer potential value based on prediction and cluster analysis," in *Management Science and Engineering (ICMSE), 2010 International Conference on*, nov. 2010, pp. 613–618.

[10] S. K. Shukla and M. Tiwari, "Soft decision trees: A genetically optimized cluster oriented approach," *Expert Systems with Applications*, vol. 36, no. 1, pp. 551–563, 2009.

[11] S. R. Nanda, B. Mahanty, and M. K. Tiwari, "Clustering indian stock market data for portfolio management," *Expert Syst. Appl.*, vol. 37, no. 12, pp. 8793–8798, December 2010.

[12] A. K. Jain, "Data clustering: 50 years beyond k-means," *Pattern Recognition Letters*, vol. 31, no. 8, pp. 651–666, 2010.

[13] M. A. Farajian and S. Mohammadi, "Mining the banking customer behavior using clustering and association rules methods," *International Journal of Industrial Enginering and Production Research*, vol. 21, no. 4, pp. 239–245, 2010.

[14] K. A. Smith and A. Ng, "Web page clustering using a self-organizing map of user navigation patterns," *Decision Support Systems*, vol. 35, no. 2, pp. 245–256, may 2003.

[15] D. Schwartz, K. A. Smith, L. Churilov, M. Dally, and R. Weber, "Design and application of hybrid intelligent systems," Amsterdam, The Netherlands, 2003, ch. Improving risk grouping rules for prostate cancer patients using self-organising maps, pp. 126–135.

[16] J. Huysmans, D. Martens, B. Baesens, J. Vanthienen, and T. Van Gestel, "Country corruption analysis with self organizing maps and support vector machines," *Intelligence and Security Informatics*, pp. 103–114, 2006.

[17] T. Kohonen, *Self-Organizing Maps*. Springer, 1995.

[18] ——, *Self-Organizing Maps*, ser. Springer Series in Information Sciences Series. Springer-Verlag GmbH, 2001.

[19] G. Pölzlbauer, "Survey and comparison of quality measures for self-organizing maps," in *Proceedings of the Fifth Workshop on Data Analysis (WDA'04)*, J. Paralič, G. Pölzlbauer, and A. Rauber, Eds. Sliezsky dom, Vysoké Tatry, Slovakia: Elfa Academic Press, jun. 2004, pp. 67–82.

[20] P. Tan, M. Steinbach, V. Kumar *et al.*, *Introduction to Data Mining*. Pearson Addison Wesley Boston, 2006.

[21] A. P. Azcarraga, M. H. Hsieh, S. L. Pan, and R. Setiono, "Extracting salient dimensions for automatic SOM labeling," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, vol. 35, no. 4, pp. 595–600, 2005.

[22] J. Vesanto and E. Alhoniemi, "Clustering of the self-organizing map," *IEEE transactions on neural networks*, vol. 11, no. 3, pp. 586–600, 2000.

[23] D. Davies and D. Bouldin, "A cluster seperation measure," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. PAMI-1, no. 2, pp. 224–227, 1979.

[24] M. Cottrell, P. Gaubert, C. Eloy, D. François, G. Hallaux, J. Lacaille, and M. Verleysen, "Fault prediction in aircraft engines using self-organizing maps," in *Proceedings of the 7th International Workshop on Advances in Self-Organizing Maps*, ser. WSOM '09. Berlin, Heidelberg: Springer-Verlag, 2009, pp. 37–44.

[25] C.-F. Hsu, C.-M. Lin, and R.-G. Yeh, "Supervisory adaptive dynamic rbf-based neural-fuzzy control system design for unknown nonlinear systems," *Appl. Soft Comput.*, vol. 13, no. 4, pp. 1620–1626, 2013.

[26] F. J. Martinez-Lopez and J. Casillas, "Artificial intelligence-based systems applied in industrial marketing: An historical overview, current and future insights," *Industrial Marketing Management*, vol. 42, no. 4, pp. 489 – 495, 2013.

[27] K. Singh, S. Gupta, P. Ojha, and P. Rai, "Predicting adsorptive removal of chlorophenol from aqueous solution using artificial intelligence based modeling approaches," *Environmental Science and Pollution Research*, vol. 20, no. 4, pp. 2271–2287, 2013.

[28] Y. K. Lam and P. W. Tsang, "exploratory k-means: A new simple and efficient algorithm for gene clustering," *Applied Soft Computing*, vol. 12, no. 3, pp. 1149 – 1157, 2012.

[29] D. Karaboga and C. Ozturk, "A novel clustering approach: Artificial bee colony (abc) algorithm," *Applied Soft Computing*, vol. 11, no. 1, pp. 652 – 657, 2011.

[30] P. J. Herrera, G. Pajares, and M. Guijarro, "A segmentation method using otsu and fuzzy k-means for stereovision matching in hemispherical images from forest environments," *Applied Soft Computing*, vol. 11, no. 8, pp. 4738 – 4747, 2011.

[31] X. Li, Y. Huang, S. Li, and Y. Zhang, "Hybrid retention strategy formulation in telecom based on k-means clustering analysis," in *E -Business and E -Government (ICEE), 2011 International Conference on*, may. 2011, pp. 1–4.

[32] T. Zhou, H. Lu, D. Yang, J. Ma, and S. Tuo, "Rough kernel clustering algorithm with adaptive parameters," in *Artificial Intelligence and Computational Intelligence*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2011, vol. 7004, pp. 604–610.

[33] F. Shang, L. Jiao, J. Shi, M. Gong, and R. Shang, "Fast density-weighted low-rank approximation spectral clustering," *Data Mining and Knowledge Discovery*, vol. 23, pp. 345–378, 2011.

[34] K. Wagstaff, C. Cardie, S. Rogers, and S. Schrödl, "Constrained k-means clustering with background knowledge," in *Proceedings of the Eighteenth International Conference on Machine Learning*. Morgan Kaufmann Publishers Inc., 2001, pp. 577–584.

[35] M. Halkidi, D. Gunopulos, M. Vazirgiannis, N. Kumar, and C. Domeniconi, "A clustering framework based on subjective and objective validity criteria," *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 4, pp. 4:1–4:25, 2008.

[36] J. Schmidt, E. Brandle, and S. Kramer, "Clustering with attribute-level constraints," in *Data Mining (ICDM), 2011 IEEE 11th International Conference on*, dec. 2011, pp. 1206 –1211.

[37] A. Mahmood, T. Li, Y. Yang, H. Wang, and M. Afzal, "Semi-supervised clustering ensemble for web video categorization," in *Multiple Classifier Systems*. Springer Berlin Heidelberg, 2013, vol. 7872, pp. 190–200.

[38] K. Sim, V. Gopalkrishnan, A. Zimek, and G. Cong, "A survey on enhanced subspace clustering," *Data Mining and Knowledge Discovery*, vol. 26, no. 2, pp. 332–397, 2013.

[39] S. Anand, S. Mittal, O. Tuzel, and P. Meer, "Semi-supervised kernel mean shift clustering," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. PP, no. 99, pp. 1–1, 2013.

[40] H. Huang, Y. Cheng, and R. Zhao, "A semi-supervised clustering algorithm based on must-link set," in *Advanced Data Mining and Applications*. Springer Berlin / Heidelberg, 2008, vol. 5139, pp. 492–499.

[41] P. S. Bradley, K. P. Bennett, and A. Demiriz, "Constrained k-means clustering," MSR-TR-2000-65, Microsoft Research, Tech. Rep., 2000.

[42] J. Sun, W. Zhao, J. Xue, Z. Shen, and Y. Shen, "Clustering with feature order preferences," in *PRICAI 2008: Trends in Artificial Intelligence*, ser. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, 2008, vol. 5351, pp. 382–393.

[43] J. Wang, S. Wu, and G. Li, "Clustering with instance and attribute level side information," 2010, vol. 3, no. 6, pp. 770–785.

[44] K. L. Wagstaff, "Intelligent clustering with instance-level constraints," Ph.D. dissertation, Faculty of the Graduate School of Cornell University, August 2002.

[45] M. H. C. Law, A. Topchy, and A. K. Jain, "Clustering with soft and group constraints," in *In Proc. joint IAPR international workshops on structural, syntactic, and statistical pattern recognition*, 2004, pp. 662–670.

[46] M. Charikar, V. Guruswami, and A. Wirth, "Clustering with qualitative information," in *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, 2003, pp. 524–.

[47] J. Wang, S. Wu, C. Wen, and G. Li, *Computing and Informatics*, no. 3, pp. 597–.

[48] A. Seret, T. Verbraken, S. Versailles, and B. Baesens, "A new som-based method for profile generation: Theory and an application in direct marketing," *European Journal of Operational Research*, vol. 220, no. 1, pp. 199–209, 2012.

[49] J. R. Miglautsch, "Thoughts on rfm scoring," *The Journal of Database Marketing*, vol. 8, no. 1, pp. 67–72, 2000.

[50] G. H. McDougall and T. J. Levesqu, "Benefit segmentation using service quality dimensions: An investigation in retail banking," *International Journal of Bank Marketing*, vol. 12, no. 2, pp. 15–23, 1994.

[51] S. M. Achim Machauer, "Segmentation of bank customers by expected benefits and attitudes," *International Journal of Bank Marketing*, vol. 19, pp. 6 – 18, 2001.

[52] R. Kuo, L. Ho, and C. Hu, "Integration of self-organizing feature map and k-means algorithm for market segmentation," *Computers and Operations Research*, vol. 29, no. 11, pp. 1475 – 1493, 2002.

[53] H. W. Shin and S. Y. Sohn, "Segmentation of stock trading customers according to potential value," *Expert Syst. Appl.*, vol. 27, no. 1, pp. 27–33, Jul. 2004.

[54] S. E. H. H. Kim S., Jung T., "Customer segmentation and strategy development based on customer lifetime value: A case study," *Expert Systems with Applications*, vol. 31, pp. 101–107, 2006.

[55] C. Jayawardhena, L. T. Wright, and C. Dennis, "Consumers online: Intentions, orientations and segmentation," *International Journal of Retail Distribution Management*, vol. 35, no. 6, pp. 515–526, 2007.

[56] K. N. Dimitriadis S., Kouremenos A., "Trust-based segmentation. preliminary evidence from technology-enabled bank channels," *International Journal of Bank Marketing*, vol. 29, no. 1, pp. 5–31, 2010.

[57] S. HV and S. Varadarajan, "Article: Customer segmentation of bank based on data mining - security value based heuristic approach as a replacement to k-means segmentation," *International Journal of Computer Applications*, vol. 19, no. 8, pp. 13–18, 2011.

[58] A. Seret, T. Verbraken, and B. Baesens, "A new knowledge-based constrained clustering approach: Theory and application in direct marketing," *Applied Soft Computing*, vol. 24, no. 0, pp. 316 – 327, 2014.

[59] N. Chen, B. Ribeiro, A. Vieira, and A. Chen, "Clustering and visualization of bankruptcy trajectory using self-organizing map," *Expert Systems with Applications*, vol. 40, no. 1, pp. 385 – 393, 2013.

[60] P. Louis, A. Seret, and B. Baesens, "Financial efficiency and social impact of microfinance institutions using self-organizing maps," *World Development*, vol. 46, no. 0, pp. 197 – 210, 2013.

[61] M.-Y. Chen, "Visualization and dynamic evaluation model of corporate financial structure with self-organizing map and support vector regression," *Applied Soft Computing*, vol. 12, no. 8, pp. 2274 – 2288, 2012.

[62] M. Zorrilla and D. Garcia-Saiz, "A service oriented architecture to provide data mining services for non-expert data miners," *Decision Support Systems*, vol. 55, no. 1, pp. 399–411, 2013.

[63] V. Carlei, A. Marra, and C. Pozzi, "Public governance, human capital and environmental outcomes: an analysis based on self-organizing maps," *Environmental Policy and Governance*, vol. 22, no. 2, pp. 116–126, 2012.

[64] A. Seret, S. vanden Broucke, B. Baesens, and J. Vanthienen, "An exploratory approach for understanding customer behavior processes based on clustering and sequence mining," in *Proceedings of the 1st International Workshop on Decision Mining and Modeling for Business Processes (DeMiMoP'13)*, in press.

[65] A. Skupin and R. Hagelman, "Visualizing demographic trajectories with self-organizing maps," *Geoinformatica*, vol. 9, no. 2, pp. 159–179, 2005.

[66] M. Sperandio and J. Coelho, "Constructing markov models for reliability assessment with self-organizing maps," in *Probabilistic Methods Applied to Power Systems, 2006. PMAPS 2006. International Conference on*, 2006, pp. 1–5.

[67] R. Agrawal, H. Mannila, R. Srikant, H. Toivonen, and A. Verkamo, *Fast discovery of association rules*. Menlo Park, CA, USA: American Association for Artificial Intelligence, 1996, pp. 307–328.

[68] R. Agrawal, D. Gunopulos, and F. Leymann, "Mining process models from workflow logs," in *Proceedings of the 6th International Conference on Extending Database Technology (EDBT'98)*, 1998, pp. 469–483.

[69] J. Cook and A. Wolf, "Discovering models of software processes from event-based data," *ACM Transactions on Software Engineering and Methodology*, vol. 7, no. 3, pp. 215–249, 1998.

[70] W. M. P. van der Aalst, A. J. M. M. Weijters, and L. Maruster, "Workflow mining: Discovering process models from event logs," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 9, pp. 1128–1142, 2004.

[71] A. Rozinat and W. M. P. van der Aalst, "Decision mining in ProM," in *Business Process Management*, 2006, pp. 420–425.

[72] R. Srikant and R. Agrawal, "Mining sequential patterns: Generalizations and performance improvements," in *Advances in Database Technology - EDBT '96*. Springer Berlin Heidelberg, 1996, vol. 1057, pp. 1–17.

[73] G. Peters, R. Weber, and R. Nowatzke, "Dynamic rough clustering and its applications," *Appl. Soft Comput.*, vol. 12, no. 10, pp. 3193–3207, 2012.

[74] B. Baesens, *Analytics in a Big Data World: The Essential Guide to Data Science and its Applications*. Wiley, 2014.

[75] A. Seret, S. K. vanden Broucke, B. Baesens, and J. Vanthienen, "A dynamic understanding of customer behavior processes based on clustering and sequence mining," *Expert Systems with Applications*, vol. 41, no. 10, pp. 4648 – 4657, 2014.

[76] W. R. N. R. Peters, G., "Dynamic rough clustering and its applications," *Applied Soft Computing*, vol. 12, no. 10, pp. 3193–3207, 2012.

[77] K. Wagstaff, C. Cardie, S. Rogers, and S. Schroedl, "Constrained k-means clustering with background knowledge," in *In ICML*. Morgan Kaufmann, 2001, pp. 577–584.

[78] A. Hiziroglu, "Soft computing applications in customer segmentation: State-of-art review and critique," *Expert Systems with Applications*, vol. 40, no. 16, pp. 6491 – 6507, 2013.

[79] P. Tan, M. Steinbach, V. Kumar *et al.*, *Introduction to Data Mining*. Pearson Addison Wesley Boston, 2006.

[80] C. C. Aggarwal and C. K. Reddy, Eds., *Data Clustering: Algorithms and Applications*. CRC Press, 2014.

[81] J. A. Lee and M. Verleysen, *Nonlinear Dimensionality Reduction*, ser. Information Science and Statistics. Berlin, Heidelberg: Springer, 2007.

[82] R. Duda, P. Hart, and D. Stork, *Pattern Classification*, 2nd ed. John Wiley & Sons, New York, 2001.

[83] K. Pearson, "On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling," *Philosophical Magazine*, vol. 50, no. 302, pp. 157–175, 1900.

[84] Y. Yang and J. O. Pedersen, "A comparative study on feature selection in text categorization." Morgan Kaufmann Publishers, 1997, pp. 412–420.

[85] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[86] I. Guyon, S. Gunn, M. Nikravesh, and L. A. Zadeh, *Feature extraction, foundations and applications*. Springer, Berlin, 2006.

[87] I. Kononenko, "Estimating attributes: Analysis and extensions of relief." Springer Verlag, 1994, pp. 171–182.

[88] T. D. Sanger, "Optimal unsupervised learning in a single-layer linear feedforward neural network," *Neural Networks*, vol. 2, no. 6, pp. 459–473, 1989.

[89] D. E. Hebb, *The Organization of Behavior*. New York: Wiley, 1949.

[90] R. Colombo and W. Jiang, "A stochastic rfm model," *Journal of Interactive Marketing*, vol. 13, no. 3, pp. 2–12, 1999.

Doctoral dissertations from the faculty of business and economics

A full list of the doctoral dissertations from the Faculty of Business and Economics can be found at:

www.kuleuven.ac.be/doctoraatsverdediging/archief.htm.