European Journal of Operational Research 220 (2012) 199–209

Contents lists available at SciVerse ScienceDirect

European Journal of Operational Research

journal homepage: www.elsevier.com/locate/ejor

Decision Support

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

Alex Seret [a,*], Thomas Verbraken [a], Sébastien Versailles [a], Bart Baesens [a,b,c]

[a] Department of Decision Sciences and Information Management, Katholieke Universiteit Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[b] School of Management, University of Southampton, Highfield Southampton SO17 1BJ, United Kingdom
[c] Vlerick, Leuven-Gent Management School, Reep 1, B-9000 Gent, Belgium

ARTICLE INFO

Article history:
Received 6 July 2011
Accepted 20 January 2012
Available online 1 February 2012

Keywords:
Data mining
Customer profiling
SOM
Direct marketing

ABSTRACT

The field of direct marketing is constantly searching for new data mining techniques in order to analyze the increasing available amount of data. Self-organizing maps (SOM) have been widely applied and discussed in the literature, since they give the possibility to reduce the complexity of a high dimensional attribute space while providing a powerful visual exploration facility. Combined with clustering techniques and the extraction of the so-called salient dimensions, it is possible for a direct marketer to gain a high level insight about a dataset of prospects. In this paper, a SOM-based profile generator is presented, consisting of a generic method leading to value-adding and business-oriented profiles for targeting individuals with predefined characteristics. Moreover, the proposed method is applied in detail to a concrete case study from the concert industry. The performance of the method is then illustrated and discussed and possible future research tracks are outlined.

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 paper, 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 paper is structured as follows. Section 2 provides the necessary background about customer segmentation and direct marketing, introducing the concepts of segmentation bases, customer profitability, the RFM framework and two techniques of segmentation, namely self-organizing maps and salient dimensions extraction. In Section 3 the SOM-based profile generator is presented and completed with ad hoc definitions of performance measures. Section 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 5.

2. Customer segmentation and 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,

* Corresponding author. Tel.: +32 16 326 880; fax: 32 16 326 624.
E-mail addresses: alex.seret@econ.kuleuven.be (A. Seret), thomas.verbraken@econ.kuleuven.be (T. Verbraken), sebastien.versailles@gmail.com (S. Versailles), bart.baesens@econ.kuleuven (B. Baesens).

0377-2217/$ - see front matter © 2012 Elsevier B.V. All rights reserved.
doi:10.1016/j.ejor.2012.01.044

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 paper 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 Kotler et al. (2006) for additional information on the segmentation bases.

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 (Nobibon et al., 2011), Bayesian neural networks (Baesens et al., 2002, 2004) and decision trees (Kim et al., 2006) 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 Lee and Li (2001), Li (2011), Yong et al. (2011), Xing and Xin-feng (2010), Shukla and Tiwari (2009), Nanda et al. (2010)) and find applications in all domains where data grouping and summarization using prototypes or profiles make sense. The evolution of these techniques, outlined in Jain (2010), offers new ways of dealing with existing problems such as the segmentation of a customer base. Moreover, new approaches combining existing techniques (e.g. Farajian and Mohammadi (2010)) reveal synergy possibilities that have to be exploited. This paper 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. Techniques used in the proposed method

This section will focus on the two major techniques used in this study, namely self-organizing maps (SOM) and salient dimensions extraction in order to provide the reader with the necessary background. It is assumed that most of the readers are already familiar with the *k*-means clustering algorithm that is used in this paper. For more information, the reader is referred to Tan et al. (2006).

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 (Smith and Ng, 2003), industrial and medical diagnostics (Schwartz et al., 2003), and corruption analysis (Huysmans et al., 2006) are contemporary examples of SOM analysis applications and successes. This section is based on Kohonen (1995) and aims at giving a theoretical background to the reader. An application of the technique can be found in Section 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):

$$\|n_i - m_c\| = \min_r\{\|n_i - m_r\|\}. \tag{1}$$

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:

$$m_r(t+1) = m_r(t) + \alpha(t)h_{cr}(t)[n(t) - m_r(t)]. \tag{2}$$

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. 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 Kohonen (1995).

2.2.2. Salient dimensions extraction

Extracting salient dimensions (SD) for automatic SOM labeling is a methodology developed by Azcarraga et al. (2005) 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:

$$|df(k,d) - \mu_{df}| \geqslant z_2\sigma_{df}(k) \tag{3}$$

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