Expert Systems with Applications 42 (2015) 6255–6266

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Identifying next relevant variables for segmentation by using feature selection approaches

Alex Seret [b,a,*], Sebastian Maldonado [b], Bart Baesens [a,c,d]

[a] Department of Decision Sciences and Information Management, KU Leuven, Naamsestraat 69, B-3000 Leuven, Belgium
[b] Faculty of Engineering and Applied Sciences, Universidad de los Andes, Las Condes, Santiago, Chile
[c] School of Management, University of Southampton, Highfield, Southampton SO17 1BJ, United Kingdom
[d] Vlerick, Leuven-Gent Management School, Reep 1, B-9000 Gent, Belgium

ARTICLE INFO

Article history:
Available online 6 March 2015

Keywords:
Clustering
Feature selection
Filter methods
Customer segmentation
Business analytics

ABSTRACT

Data mining techniques are widely used by researchers and companies in order to solve problems in a myriad of domains. While these techniques are being adopted and used in daily activities, new operational challenges are encountered concerning the steps following this adoption. In this paper, the problem of updating and improving an existing clustering model by adding relevant new variables is studied. 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.

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 (Baesens, 2014). 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 customers can thus be explored (see e.g. Seret, vanden Broucke, Baesens, & Vanthienen, 2014a). 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. Peters, 2012). 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 paper 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

* Corresponding author at: Faculty of Engineering and Applied Sciences, Universidad de los Andes, Las Condes, Santiago, Chile.

E-mail addresses: alex.seret@kuleuven.be (A. Seret), smaldonado@uandes.cl (S. Maldonado).

http://dx.doi.org/10.1016/j.eswa.2015.01.070
0957-4174/© 2015 Published by Elsevier Ltd.

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 Seret, Verbraken, and Baesens (2014b) and Wagstaff, Cardie, Rogers, and Schroedl (2001), 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 paper is structured as follows. In Section 2, the theoretical background on unsupervised learning for customer segmentation is summarized, including a description of relevant evaluation metrics for clustering performance. Section 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 4. In Section 5, the proposed methodology is applied in a real case involving a marketing department from the event industry. Section 6, concludes this paper and identifies new tracks for research.

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, Hiziroglu (2013) 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 Hiziroglu (2013) 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 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.

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 Louis, Seret, and Baesens (2013), Seret, Verbraken, Sebastien, and Baesens (2012), Azcarraga, Hsieh, Pan, and Setiono (2005) and Vesanto and Alhoniemi (2000). 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 Kohonen, 1995, 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 previously trained are clustered using a classical partitioning algorithm as the *k*-means algorithm (Tan, Steinbach, & Kumar, 2006). 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 Vesanto and Alhoniemi (2000) and illustrated in Section 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 Fig. 1 and applied in Section 5. The reader is referred to Kohonen (1995) for an exhaustive discussion of the SOM algorithm and to the previously cited works for examples of applications.

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 Aggarwal and Reddy (2014). 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 5 in 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