International Journal of Business Analytics
Volume 3 • Issue 4 • October-December 2016

Supervised Regression Clustering:
A Case Study for Fashion Products

Ali Fallah Tehrani, Technology Campus Grafenau, Deggendorf Institute of Technology, Grafenau, Germany
Diane Ahrens, Technology Campus Grafenau, Deggendorf Institute of Technology, Grafenau, Germany

ABSTRACT

Clustering techniques typically group similar instances underlying individual attributes by supposing that similar instances have similar attributes characteristic. On contrary, clustering similar instances given a specific behavior is framed through supervised learning. For instance, which fashion products have similar behavior in term of sales. Unfortunately, conventional clustering methods cannot tackle this case, since they handle attributes by a same manner. In fact, conventional clustering approaches do not consider any response, and moreover they assume attributes act by the same importance. However, clustering instances with respect to responses leads to a better data analytics. In this research, the authors introduce an approach for the goal supervised clustering and show its advantage in terms of data analytics as well as prediction. To verify the feasibility and the performance of this approach the authors conducted several experiments on a real dataset derived from an apparel industry.

KEYWORDS

Data Mining, Fashion Products, Forecast, Hadamard Product, k-means Clustering, Minkowski Distance, Supervised Clustering

1. INTRODUCTION

Clustering techniques, traditionally as primary tools in exploratory data analytics have received notably attention due to the fact that identifying similar objects allows for categorization, which simplifies the input space. By conventional clustering methods the implicit emphasis is that all attributes have the same importance. However, this assumption is not reliable for several reasons (Garcia-Escudero & Gordaliza, 2007). Firstly, adding non-relevant attributes to a feature space leads to different similarities. Secondly, the magnitude of attributes plays a crucial role, i.e. while two instances maybe similar on the whole feature space, considering a new attribute can lead to a significant dissimilarity [1]. Nevertheless, an incorrect understanding of the underlying similarity can cause a flawed understanding of the data. We shall illustrate this point by mentioning an example: Assume several cars from several brands, each of which is characterized by horsepower, consumption, weight and number of doors with the car price as output. Since prices vary, the goal may be to cluster cars by price. One might think of a price clustering, however, a price clustering could indeed lead to the clustering of very dissimilar cars, e.g. a large car may be in the same group as a small and speedy car since both are in the same price range. The simple reason is that projecting an object specified by features to a one-dimensional space (output) leads to information lost. Note that in a straightforward manner response can be integrated in the feature space, and immediately the clustering techniques can be conducted, however, when the price as a factor is taken into account the simple clustering approach cannot sparkle its effect

DOI: 10.4018/IJBAN.2016100102

correctly. The simple reason is that the price effect thanks to the other input factors can be ignored, especially in the light of the large number of input attributes the effect can be neglected.

More concretely, we are aiming at clustering the articles from fashion products regarding the number of sales. One of the challenging tasks of fashion retailers is to estimate approximately the number of orders w.r.t. several products for the next season or even the next year. Due to the costs associated with purchasing and transferring, typically apparel retailers refrain from ordering more than two times per year. Seen from this perspective, an accurate sales-forecasting is required and prevents reordering. Accounting for the fact that each product is characterized by several qualitative and quantitative attributes, the goal is to find established patterns in the sales records on the use of a reliable ordering. A conventional solution is addressed under fitting a regression curve based on available sales data, however, in the presence of outliers a simple regression model delivers poor results. To overcome this problem our idea refers to identify the products which have similar characteristic in terms of the number of sales. We should again emphasize that thresholding solely based on the regression output may lead to the wrong conclusion, namely thresholding on the number of sales sacrifices a part of information, i.e., while a product is highly sold due to a low price a trendy product my sell well due to its trendiness. To cope with this inconsistency, the model should incorporate the response with other input factors, which leads to a more robust solution. In this regard, we modify the conventional clustering approach by integrating the response into the model, which at the core of the idea lies a proper weighting approach. Another drawback of the conventional clustering is to assume the same weights for all input factors, which indeed leads to a poor clustering, due to the fact that irrelevant input factors easily contribute to the model, even more in the case that all input factors are relevant, they are incorporating in different manners.

This paper is organized as follows: in the next section we give an overview on existing semi-supervised clustering approaches. In Section 3 we discuss comprehensive about conventional clustering. Section 4 is dedicated to our proposal. In Section 5 the algorithm is presented and in Section 7 the first preliminary results are shown. Finally, in Section 8 the concluding remarks and future horizon are discussed.

2. RELATED WORK

While the methods underlying conventional clustering approaches (Ackerman, & Ben-David, 2009; Aparna, & Mydhili, 2015; Bach, & Jordan, 2003; Bouveyron, Girard, & Schmid, 2007; Bock, 1996; Cardot, Cenac, & Monnez, 2012; Cuevas, Febrero, & Fraiman, 2001; Dhillon, Guan, & Kogan, 2002; Hruschka, & Natter, 1999; Jain, & Dubes, 1988; Klein, Kamvar, & Manning, 2002; Mulvey, & Beck, 1984; Sarkar, Yegnanarayana, & Khemani, 1997) have been studied almost comprehensively, methods in supervised clustering (Awasthi, & Bosagh Zadeh, 2010; Grira, Crucianu, Boujemaa, 2004) have been studied far less deeply and are typically addressed under *semi-supervised clustering* (Gao,, Tan, & Cheng, 2006; Hochbaum, & Shmoys, 1985; Xing, Jordan, Russell, & Ng, 2003; Wagstaff, Cardie, Rogers, & Schrödl, 2001). Seen from this view more investigation in this regard is desirable. In the following, we give a survey on the existing approach w.r.t. the semi-supervised clustering.

In (Eick, Zeidat, & Zhao, 2004) it is proposed the idea to use a fitness function on the basis of class impurity. The class impurity indicates the percentage of minority examples in the different clusters. By assuming the fitness function based on impurity it is possible to cluster instances associating the same class. In (Finley, & Joachims, 2005) it is proposed to define parametrized k-means clustering underlying support vector machine framework equipped by cutting plane algorithm. In addition, it is supposed that as training data some instances are already clustered, and henceforth can be applied for training phase. To fit the optimal parameters then the cutting plane algorithm is used, and once the parameters are trained the parameters are used for the goal clustering. In (Basu, Bilenko, & Mooney, 2004) a probabilistic framework applied to cluster data points. In this regard, the objective function for clustering is weighted by probability density function. In addition (Demiriz, Bennett,