Available online at www.sciencedirect.com
SCIENCE DIRECT®
European Journal of Operational Research 173 (2006) 705–716
EUROPEAN JOURNAL OF OPERATIONAL RESEARCH
www.elsevier.com/locate/ejor

Data mining in a bicriteria clustering problem

E. Abascal, I. Garcia Lautre, F. Mallor *

Department of Statistics and Operations Research, Public University of Navarra, 31006 Pamplona, Spain

Received 26 October 2004; accepted 28 March 2005
Available online 15 November 2005

Abstract

In this paper, we address the issue of clustering elements, described by a large set of non-negative variables, first using quantitative criteria to differentiate variable values, and then qualitative criteria to focus on whether or not the variables take a zero value. A zero value is relevant in a managerial context, for example, where it may indicate non-consumption of a certain product. In this case, a zero versus a positive value constitutes, in itself, an primary point of interest. This is the type of situation, moreover, in which there is usually a high frequency of zero values. We suggest two different approaches to the analysis of these data. One uses multiple factor analysis (MFA), which allows a compromise between qualitative and quantitative criteria. The other proposes a family of functions for transforming the original data in such a way that the parameter used to index the functions is interpreted as the weight assigned to each criterion. We have tested both procedures on a real-world data set to obtain a customer typology for a telecommunications company. The results were encouraging and useful to the managers.

Keywords: Data mining; Clustering; Multiple factor analysis; Telecommunications

1. Introduction

Data analysis aimed at grouping elements into homogeneous classes plays an important part in business decision-making. The two most widely used techniques are classification and clustering. The main difference between them is whether the classes are pre-defined or unknown. In the first case, classification techniques are used; in the second clustering techniques are required.

A review of multicriteria classification methods is presented in Zopounidis and Doumpos (2002), which also includes a long list of practical applications (including numerous references). The real-world applications of the classification problem include the fields of medicine, pattern recognition, human resources management, production systems

* Corresponding author.
E-mail address: mallor@unavarra.es (F. Mallor).

0377-2217/$ - see front matter
doi:10.1016/j.ejor.2005.10.006

management and technical diagnosis, marketing, environmental and energy management, financial management, economics, and so on. Obviously, when the groups are not known in advance, then the problem will not be one of classification but of clustering. For a comparative assessment of classification methods we would recommend Kiang (2003).

In this paper, we consider the problem of grouping a large set of elements, namely customers, described by a large set of non-negative integer variables, into homogeneous classes that are not pre-defined.

In this case, we need to identify potential groups, or clusters, such that the elements in the same group are more similar to each other than to elements in other clusters. To deal with this clustering problem, many tools, such as neural networks, mathematical programming, or multivariate statistics have been proposed and studied (Gordon, 1999). Our work is based on a multicriteria approach combined with classical statistical procedures. As far as we know, there are few unsupervised multicriteria techniques (Zopounidis and Doumpos, 2002).

Thus, we approach the problem of clustering elements, using both a quantitative criterion, in the sense that two individuals with different values for the variables are different, and also a qualitative criterion, which focuses on whether or not the variables take a zero value. This relevance of the zero value arises in situations in which it indicates the non-consumption of a certain product (in a managerial context) or the absence of a symptom (in a medical context). Therefore, differentiating between a zero value and a positive value constitutes an interesting classification in itself. These situations, moreover, tend to show a high frequency of zero values.

The problem that concerns us arose in a telecommunications company when, after a decade of considerable growth, the management wished to know more details concerning the client portfolio. Very large amounts of data on clients' consumption behaviour had been recorded, but no personal data or social information was available about them. The managers agreed to do a cluster analysis, using information from their data warehouse, in order to identify homogeneous customer clusters. The information dealt with was the services were used, when they were used and how often they were used. A similar issue is treated in Daskalaki et al. (2003) where data mining is used to analyse insolvency among the customers of a telecommunications company. Their case is approached as a classification problem.

After a preliminary analysis with disappointing results, we reformulated the problem in order to tackle it from a multicriteria perspective. Not finding in the literature any analysis of a similar problem, we proposed two different methodologies of solution: one is based on prior transformation of the data, the other on the use of multiple factor analysis. Both methods performed well in our practical experience.

The paper is organised as follows. In Section 2, we formally introduce the problem and the available data. We also explain how to apply the two methodologies for solving the bicriteria clustering problem and propose a graphical procedure to compare the results of several classifications. In Section 3, we discuss the difficulties of this problem when dealing with real data and illustrate the application of the two methodologies to a set of simulated data with a statistical pattern similar to that of the original ones, although in a more reduced space. Confidentiality requirements prevent us from revealing either the characteristics of the real data or the results of their analysis. We end the paper with some concluding remarks and a list of references.

2. Data and methodology

In the first part of this section, we describe our bicriteria clustering problem, including such aspects as the individuals and variables involved, and the criteria for grouping the elements. For this purpose we will use the same terminology as in the application: the elements we have to group are the customers of a company and the clusters we identify should be interpreted in terms of different customer typologies. In the second part, we describe our two approaches to the problem: multiple factor analysis (MFA) and prior transformation of