ANNALS of the ORADEA UNIVERSITY.
Fascicle of Management and Technological Engineering, Volume VII (XVII), 2008

MODERN METHODS FOR DATA ANALYSIS. CLUSTER ANALYSIS

Simona Aurelia Bodog
University of Oradea
Faculty of Electrical Engineering and Information Technology
sbodog@uoradea.ro
Alexandru Constăngioară, Mirela Bucurean
University of Oradea, Faculty of Economic Science
sandu_oradea@yahoo.com

Key words: quantitative management, market research, business application, cluster analysis.

Abstract. Data analysis is paramount to modern business word. Numerous supervised and unsupervised techniques have become available for supporting efficient managerial decisions. This paper focuses on cluster analysis and its uses in business world.

1. Introduction.

Unsupervised classification is classification with an unknown target. That is, the class of each case is unknown. The total number of classes is also unknown. The aim is to segment the cases into disjoint classes that are homogenous with respect to the inputs. However, there is no guarantee that the resulting clusters will be meaningful or useful. Unsupervised classification is also useful as a step in a supervised prediction problem. For example, customers could be clustered into homogenous groups based on sales of different items. Then a model could be built to predict the cluster membership based on some more easily obtained input variables. K-means clustering is an unsupervised classification method. One setback of problem formulation is failing to appreciate the limitations of the analytical methods i.e. forcing a number of clusters with no practical relevance.

2. Business Problem.

A catalog company periodically purchases lists of prospects from outside sources. They want to design a test mailing to evaluate the potential response rates for several different products. Based on their experience, they know that customer preference for their products depends on geographic and demographic factors. Consequently, they want to segment the prospects into groups that are similar to each other with respect to these attributes.

After the prospects have been segmented, a random sample of prospects within each segment will be mailed one of several offers. The results of the test campaign will allow the analyst to evaluate the potential profit of prospects from the list source overall as well as for specific segments.

3. Cluster Analysis.

A dataset of 5078 observations was employed. I use cluster analysis in order to form homogenous classes of subjects. Analysis indicates 6 clusters have been identified. Results are presented in what follows.

Fig. 1 summarizes three statistics for each of the six clusters. The height of the slice indicates the number of cases in each cluster. Clusters 3 and 5 contain the most cases, and cluster 4 contains the fewest. The width of each slice is set to Std. deviation, which is the root-mean-square standard deviation (root-mean-square distance) between cases in the cluster. The color is set to radius, which is the distance of the farthest cluster member from the cluster seed.

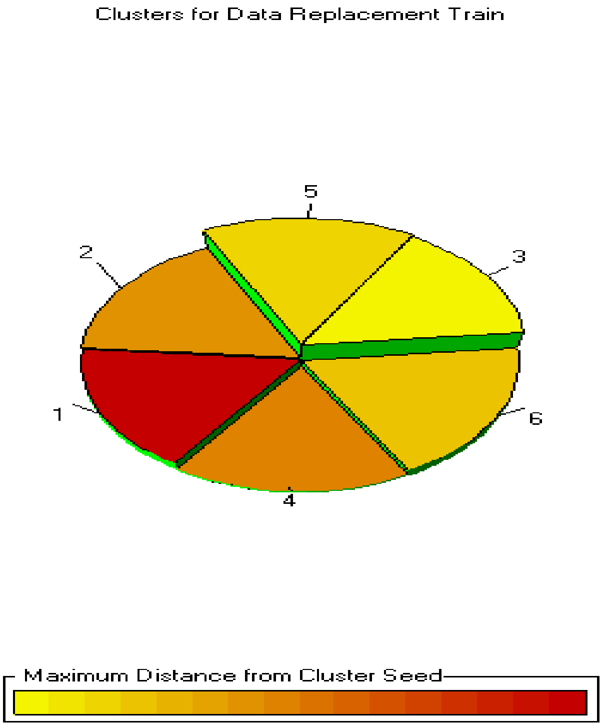

Fig.1. Clusters for dataset in use

Further analysis is necessary for our purposes. Figure 2 presents a graphical representation of the clusters. Multi-dimensional scaling is used to construct two dimensions of the clusters.

Fig. 2 provides a graphical representation of the size of each cluster and the relationship among clusters. The axes are determined from multidimensional scaling analysis using a matrix of distances between cluster means as input. The asterisks are the cluster centers, and the circles represent the cluster radii. A cluster that contains only one case is displayed as an asterisk. The radius of each cluster depends on the most distant case in that cluster, and cases may not be distributed uniformly within clusters. Hence, it may appear that clusters overlap, but in fact each case is assigned to only one cluster.

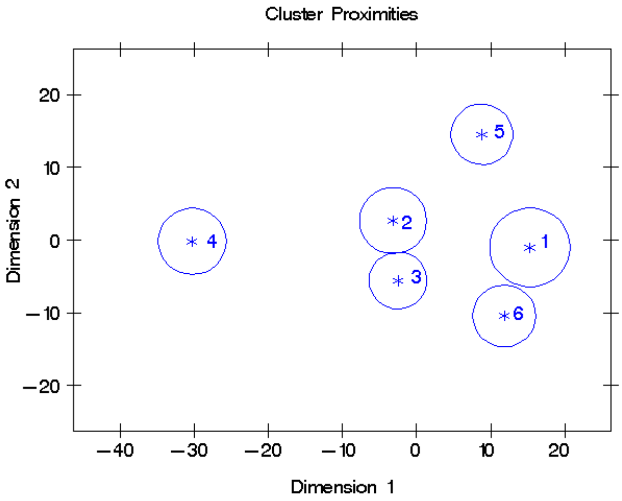

Fig.2. Graphical representation of clusters

Figure 3 and Figure 4 depicts several distributions of the variables by clusters. We see that for nominal variables column bar graph is used whereas for interval variable a box plot is employed. We see that males predominates in cluster 5 and females in cluster 6. Climate chard is more difficult to interpret since more climate zones are available.

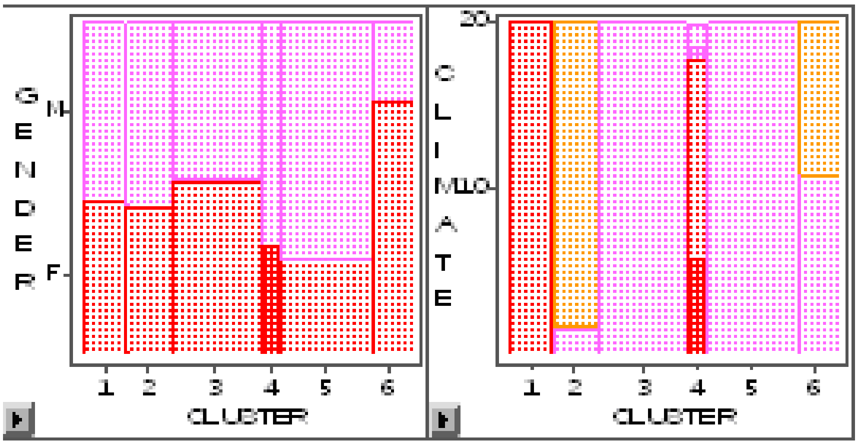

Fig.3. Distribution of nominal variables by clusters

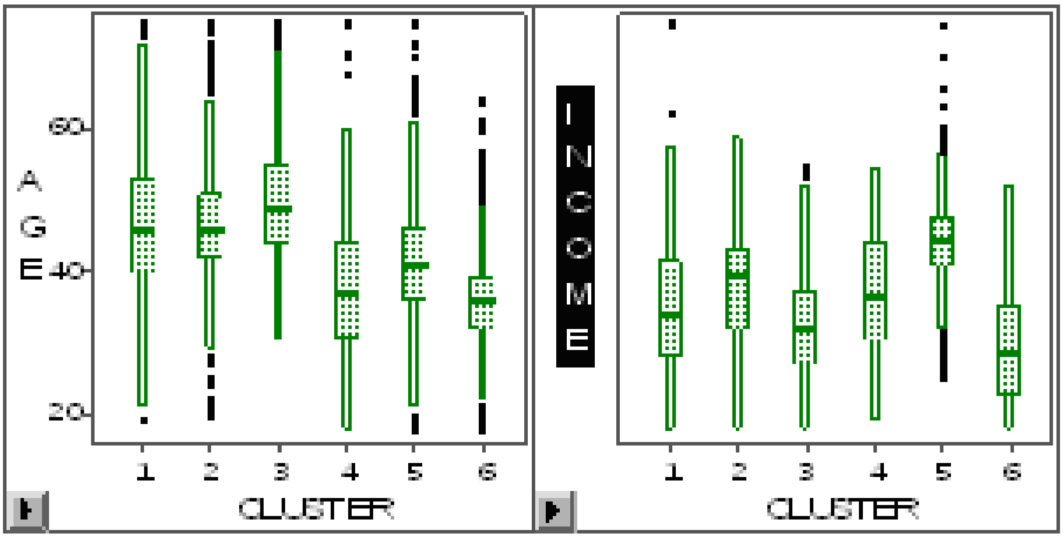

Fig.4. Distribution of scale variables by clusters

Figure 4 shows that young people are more numerous in cluster 4 and 6 whereas in cluster 3 the average age is highest. Similarly, average income is greatest in cluster 5 and lowest in cluster 6.

4. Conclusion.

In summary, the six clusters can be described as follows:

Cluster 1	married persons living in climate zone 10
Cluster 2	married persons living in climate zone 30
Cluster 3	married persons living in climate zone 20
Cluster 4	younger, unmarried persons with lower FICO scores, living in climate zone 10
Cluster 5	younger, unmarried men with higher incomes, living in climate zone 20
Cluster 6	younger, unmarried women living in climate zone 20 or 30.

These clusters may or may not be useful for marketing strategies, depending on the line of business and planned campaigns.

References.

1. Cole, A. J. & Wishart, D. (1970). An improved algorithm for the Jardine-Sibson method of generating overlapping clusters. The Computer Journal 13(2):156-163.
2. Ester, M., Kriegel, H.P., Sander, J., and Xu, X. 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, USA
3. Heyer, L.J., Kruglyak, S. and Yooseph, S., Exploring Expression Data: Identification and Analysis of Coexpressed Genes, Genome Research