DSS Using Visualization of Multi-Algorithms Voting

Ran M. Bittmann
Graduate School of Business Administration – Bar-Ilan University, Israel

Roy M. Gelbard
Graduate School of Business Administration – Bar-Ilan University, Israel

INTRODUCTION

The problem of analyzing datasets and classifying them into clusters based on known properties is a well known problem with implementations in fields such as finance (e.g., pricing), computer science (e.g., image processing), marketing (e.g., market segmentation), and medicine (e.g., diagnostics), among others (Cadez, Heckerman, Meek, Smyth, & White, 2003; Clifford & Stevenson, 2005; Erlich, Gelbard, & Spiegler, 2002; Jain & Dubes, 1988; Jain, Murty, & Flynn, 1999; Sharan & Shamir, 2002).

Currently, researchers and business analysts alike must try out and test out each diverse algorithm and parameter separately in order to set up and establish their preference concerning the individual decision problem they face. Moreover, there is no supportive model or tool available to help them compare different results-clusters yielded by these algorithm and parameter combinations. Commercial products neither show the resulting clusters of multiple methods, nor provide the researcher with effective tools with which to analyze and compare the outcomes of the different tools.

To overcome these challenges, a decision support system (DSS) has been developed. The DSS uses a matrix presentation of multiple cluster divisions based on the application of multiple algorithms. The presentation is independent of the actual algorithms used and it is up to the researcher to choose the most appropriate algorithms based on his or her personal expertise.

Within this context, the current study will demonstrate the following:

- How to evaluate different algorithms with respect to an existing clustering problem.
- Identify areas where the clustering is more effective and areas where the clustering is less effective.
- Identify problematic samples that may indicate difficult pricing and positioning of a product.

Visualization of the dataset and its classification is virtually impossible using legacy methods when more than three properties are used, as is the case in many problems, since displaying the dataset in such a case will require giving up some of the properties or using some other method to display the dataset's distribution over four or more dimensions. This makes it very difficult to relate to the dataset samples and understand which of these samples are difficult to classify, (even when they are classified correctly), and which samples and clusters stand out clearly (Boudjeloud & Poulet, 2005; De-Oliveira & Levkowitz, 2003; Grabmier & Rudolph, 2002; Shultz, Mareschal, & Schmidt, 1994).

Even when the researcher uses multiple algorithms in order to classify the dataset, there are no available tools that allow him/her to use the outcome of the algorithms' application. In addition, the researcher has no tools with which to analyze the difference in the results.

The current study demonstrates the usage of a developed decision support methodology based upon formal quantitative measures and a visual approach, enabling presentation, comparison, and evaluation of the multi-classification suggestions resulting from diverse algorithms. The suggested methodology and DSS support a cross-algorithm presentation; all resultant classifications are presented together in a "Tetris-like format" in which each column represents a specific classification algorithm and each line represents a specific sample case. Formal quantitative measures are then used to analyze these "Tetris blocks," arranging them according to their best structures, that is, the most agreed-upon classification, which is probably the most agreed-upon decision.

Such a supportive model and DSS impact the ultimate business utility decision significantly. Not only can it save critical time, it also pinpoints all irregular sample cases, which may require specific examination. In this way, the decision process focuses on key issues instead of wasting time on technical aspects. The DSS is demonstrated using common clustering problems of wine categorizing, based on 13 measurable properties.

THEORETICAL BACKGROUND

Cluster Analysis

In order to classify a dataset of samples with a given set of properties, researchers use algorithms that associate each sample with a suggested group-cluster, based on its properties. The association is performed using likelihood measure that indicates the similarity between any two samples as well as between a sample, to be associated, and a certain group-cluster.

There are two main clustering-classification types:

- **Supervised** (also called categorization), in which a fixed number of clusters are predetermined, and the samples are divided-categorized into these groups.
- **Unsupervised** (called clustering), in which the preferred number of clusters, to classify the dataset into, is formed by the algorithm while processing the dataset.

There are unsupervised methods, such as hierarchical clustering methods, that provide visualization of entire "clustering space" (dendrogram), and in the same time enable predetermination of a fixed number of clusters.

A researcher therefore uses the following steps:

1. The researcher selects the best classification algorithm based on his/her experience and knowledge of the dataset.
2. The researcher tunes the chosen classification algorithm by determining parameters, such as the likelihood measure, and number of clusters.

Current study uses hierarchical clustering methods, which are briefly described in the following section.

Hierarchical Clustering Methods

Hierarchical clustering methods refer to a set of algorithms that work in a similar manner. These algorithms take the dataset properties that need to be clustered and start out by classifying the dataset in such a way that each sample represents a cluster. Next, it merges the clusters in steps. Each step merges two clusters into a single cluster until only one cluster (the dataset) remains. The algorithms differ in the way in which distance is measured between the clusters, mainly by using two parameters: the distance or likelihood measure, for example, Euclidean, Dice, and so forth, and the cluster method, for example, between group linkage, nearest neighbor, and so forth.

In the present study, we used the following well-known hierarchical methods to classify the datasets:

- **Average linkage (within groups):** This method calculates the distance between two clusters by applying the likelihood measure to all the samples in the two clusters. The clusters with the best average likelihood measure are then united.
- **Average linkage (between groups):** This method calculates the distance between two clusters by applying the likelihood measure to all the samples of one cluster and then comparing it with all the samples of the other cluster. Once again, the two clusters with the best likelihood measure are then united.
- **Single linkage (nearest neighbor):** This method, as in the average linkage (between groups) method, calculates the distance between two clusters by applying the likelihood measure to all the samples of one cluster and then comparing it with all the samples of the other cluster. The two clusters with the best likelihood measure, from a pair of samples, are united.
- **Median:** This method calculates the median of each cluster. The likelihood measure is applied to the medians of the clusters, after which the clusters with the best median likelihood are then united.
- **Ward:** This method calculates the centroid for each cluster and the square of the likelihood measure of each sample in both the cluster and

Related Content