USING THE AGGLOMERATIVE METHOD OF HIERARCHICAL CLUSTERING AS A DATA MINING TOOL IN CAPITAL MARKET[1]

Vera Marinova–Boncheva

Abstract: The purpose of this paper is to explain the notion of clustering and a concrete clustering method- agglomerative hierarchical clustering algorithm. It shows how a data mining method like clustering can be applied to the analysis of stocks, traded on the Bulgarian Stock Exchange in order to identify similar temporal behavior of the traded stocks. This problem is solved with the aid of a data mining tool that is called XLMiner™ for Microsoft Excel Office.

Keywords: Data Mining, Knowledge Discovery, Agglomerative Hierarchical Clustering.

ACM Classification Keywords: I.5.3 Clustering

Introduction

Data mining, the extraction of hidden predictive information from large databases, is a powerful new technology with great potential to help companies focus on the most important information in their data warehouses. Data mining tools predict future trends and behaviors, allowing businesses to make proactive, knowledge-driven decisions. The automated, prospective analyses offered by data mining move beyond the analyses of past events provided by retrospective tools typical of decision support systems. Data mining tools can answer business questions that traditionally are time consuming to resolve. They scour databases for hidden patterns, finding predictive information that experts may miss because it lies outside their expectations. Data mining consists of analysis of sets of supervised data with the aim of finding unexpected dependencies or to be generalized in a new way that is understandable and useful for owners of the data. There is a great deal of data mining techniques but we differentiate two of them like classification and clustering as supervised and unsupervised learning from data. [2]

The Analysis of Clustering

Clustering can be considered the most important unsupervised learning problem. So, as every other problem of this kind, it deals with finding a structure in a collection of unlabeled data. A cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

Cluster Analysis, also called data segmentation, has a variety of goals. They all relate to grouping or segmenting a collection of objects (also called observations, individuals, cases, or data rows) into subsets or "clusters", such that those within each cluster are more closely related to one another than objects assigned to different clusters. Central to all of the goals of cluster analysis is the notion of degree of similarity (or dissimilarity) between the individual objects being clustered that depends on the data and the application. Different types of similarity measures may be used to identify classes (clusters), where the similarity measure controls how the clusters are formed. Some examples of values that can be used as similarity measures include distance, connectivity, and intensity. [4]

The main requirements that a clustering algorithm should satisfy are:

[1] This work was supported by the Ministry of Education and Science under Project № MU-MI-1601/2006

- scalability;
- dealing with different types of attributes;
- discovering clusters with arbitrary shape;
- minimal requirements for domain knowledge to determine input parameters;
- ability to deal with noise and outliers;
- insensitivity to order of input records;
- high dimensionality;
- constrained - based clustering;
- interpretability and usability. [7]

Clustering algorithms may be classified as listed below:

- Exclusive Clustering
- Overlapping Clustering
- Hierarchical Clustering
- Probabilistic Clustering

In the first case data are grouped in an exclusive way, so that if a certain datum belongs to a definite cluster then it could not be included in another cluster. On the contrary the second type, the overlapping clustering, uses fuzzy sets to cluster data, so that each point may belong to two or more clusters with different degrees of membership. In this case, data will be associated to an appropriate membership value. A hierarchical clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations it reaches the final clusters wanted. Finally, the last kind of clustering uses a completely probabilistic approach. [5, 6]

There are a number of problems with clustering. Among them:

- current clustering techniques do not address all the requirements adequately (and concurrently);
- dealing with large number of dimensions and large number of data items can be problematic because of time complexity;
- the effectiveness of the method depends on the definition of “distance” (for distance-based clustering);
- if an obvious distance measure doesn’t exist we must “define” it, which is not always easy, especially in multi-dimensional spaces;
- the result of the clustering algorithm (that in many cases can be arbitrary itself) can be interpreted in different ways.

Clustering is a method that is applicable in many fields like:

- Marketing: finding groups of customers with similar behavior when it is given a large database of customer data containing their properties and past buying records;
- Biology: classification of plants and animals given their features;
- Libraries: book ordering;
- Insurance: identifying groups of motor insurance policy holders with a high average claim cost; identifying frauds;
- City-planning: identifying groups of houses according to their house type, value and geographical location;
- Earthquake studies: clustering observed earthquake epicenters to identify dangerous zones;
- WWW: document classification; clustering weblog data to discover groups of similar access patterns.

Hierarchical Clustering

In hierarchical clustering the data are not partitioned into a particular cluster in a single step. Instead, a series of partitions takes place, which may run from a single cluster containing all objects to N clusters each containing a single object. Hierarchical Clustering is subdivided into agglomerative methods, which proceed by series of fusions of the N objects into groups, and divisive methods, which separate N objects successively into finer groupings. Agglomerative techniques are more commonly used, and this is the method implemented in the free version of XLMiner™ which is the Microsoft Office Excel add-in. [1]

If it is given a set of N items to be clustered and a N*N distance (or similarity) matrix then the basic process of agglomerative hierarchical clustering can be done iteratively following these four steps:

1. Start by assigning each item to a cluster. Let the distances (similarities) between the clusters are the same as the distances (similarities) between the items they contain;
2. Find the closest (most similar) pair of clusters and merge them into a single cluster;
3. Compute distances (similarities) between the new cluster and each of the old clusters;
4. Repeat step 2 and 3 until all items are clustered into a single cluster of size N.

Step 3 can be different because of the varieties in the definition of the distance (or similarity) between clusters:

- Single linkage clustering (nearest neighbor technique) – here the distance between groups is defined as the distance between the closest pair of objects, where only pairs consisting of one object from each group is considered i.e. the distance between two clusters is given by the value of the shortest link between clusters. At each stage the two clusters for which the distance is minimum are merged;
- Complete linkage clustering (farthest neighbor) – is the opposite of the single linkage i.e. distance between groups is defined as the distance between the most distant pair of objects, one from each group. At each stage the two clusters for which the distance is minimum are merged;
- Average linkage clustering – the distance between two clusters is defined as the average of distances between all pairs of objects, where each pair is made up of one object from each group. At each stage the two clusters for which the distance is minimum are merged;
- Average group linkage clustering – with this method, groups once formed are represented by their mean values for each variable, that is their mean vector and inter-group distance is defined in terms of distance between two such mean vectors. At each stage the two clusters for which the distance is minimum are merged. In this case, those two clusters are merged such that the newly formed cluster, on average, will have minimum pairwise distances between the points in it;
- Ward's hierarchical clustering – Ward (1963) proposed a clustering procedure seeking to form the partitions $P_n,...,P_1$ in a manner that minimizes the loss associated with each grouping and to quantify that loss in a form that is readily interpretable. At each step the union of every possible cluster pair is considered and the two clusters whose fusion results in minimum increase in "information loss" are combined. Information loss is defined by Ward in terms of an error sum-of-squares criterion. [3]

Hierarchical clustering may be represented by a two dimensional diagram known as dendrogram which illustrates the fusions or divisions made at each successive stage of analysis. By cutting the dendrogram at a desired level clustering of the data items into disjoint groups is obtained. [1]

Major weakness of agglomerative clustering methods is that:

- they do not scale well and time complexity is at least $O(n^2)$, where n n is the number of total objects;
- they can never undo what was done previously.

Clustering of Stocks, traded on the Official Market of BSE

As inputs we have taken data for 16 stocks from the Bulgarian Stock Exchange in a single day. (Table 1) These data are listed on the Internet address: http://www.econ.bg/capital.html. It contains information for each stock as the code and the name of the company, the nominal, prices (low, high, last, medium), the change in price in comparison to the previous day and the traded amount of this kind of stock.

company code	nominal	prices				change	amount
		low	high	last	medium		
CENHL	1	29	30.1	29.78	29.78	0.91	1231
SFARM	1	7.72	7.96	7.9	7.9	0.09	130848
CCB	1	8.17	8.29	8.17	8.17	-0.06	379598
PETHL	1	11.36	11.99	11.79	11.79	0.7	30508
DOVUHL	1	5.25	5.4	5.3	5.3	-0.19	17201
IHLBL	1	7.7	8	7.97	7.97	0.04	7608
ALBHL	1	16.01	16.5	16.38	16.38	-0.02	6493
GAZ	1	10.01	10.2	10.13	10.13	-0.07	24693
PET	1	4.86	4.95	4.95	4.95	0.05	303240
ORGH	1	144.5	146	145.04	145.04	-0.76	292
HVAR	1	38.12	44.49	41.92	41.92	3.23	1929
SEVTO	1	6.47	6.72	6.64	6.64	0.11	4637
ODES	1	185	190	185.2	185.2	-1.01	75
CHIM	1	10.8	11.3	11.02	11.02	0.1	229116
MONBAT	1	9.53	9.7	9.6	9.6	-0.07	67937
KTEX	1	24.5	25	24.77	24.77	0.19	700

Table 1. Information about stocks, traded on the Official Market of Bulgarian Stock Exchange

We use the data mining tool named XLMiner™ for MS Excel. We select the agglomerative method of hierarchical clustering to find clusters of stocks. We experiment on all five variants of agglomerative method of hierarchical clustering and we have founded that the average linkage method will give the best results. We use as a stop rule for the process of clustering the number of clusters which is 4. [1]

Row Id.	Cluster Id	Sub Cluster Id	Var1	Var2	Var3	Var4	Var5	Var6
1	1	1	29	30.1	29.78	29.78	0.91	1231
2	2	2	7.72	7.96	7.9	7.9	0.09	130848
3	3	3	8.17	8.29	8.17	8.17	-0.06	379598
4	1	4	11.36	11.99	11.79	11.79	0.7	30508
5	1	5	5.25	5.4	5.3	5.3	-0.19	17201
6	1	6	7.7	8	7.97	7.97	0.04	7608
7	1	7	16.01	16.5	16.38	16.38	-0.02	6493
8	1	8	10.01	10.2	10.13	10.13	-0.07	24693
9	4	9	4.86	4.95	4.95	4.95	0.05	303240
10	1	10	144.5	146	145.04	145.04	-0.76	292
11	1	11	38.12	44.49	41.92	41.92	3.23	1929
12	1	12	6.47	6.72	6.64	6.64	0.11	4637
13	1	13	185	190	185.2	185.2	-1.01	75
14	4	14	10.8	11.3	11.02	11.02	0.1	229116
15	1	15	9.53	9.7	9.6	9.6	-0.07	67937
16	1	16	24.5	25	24.77	24.77	0.19	700

Table 2. Clusters of stocks taken from table 1

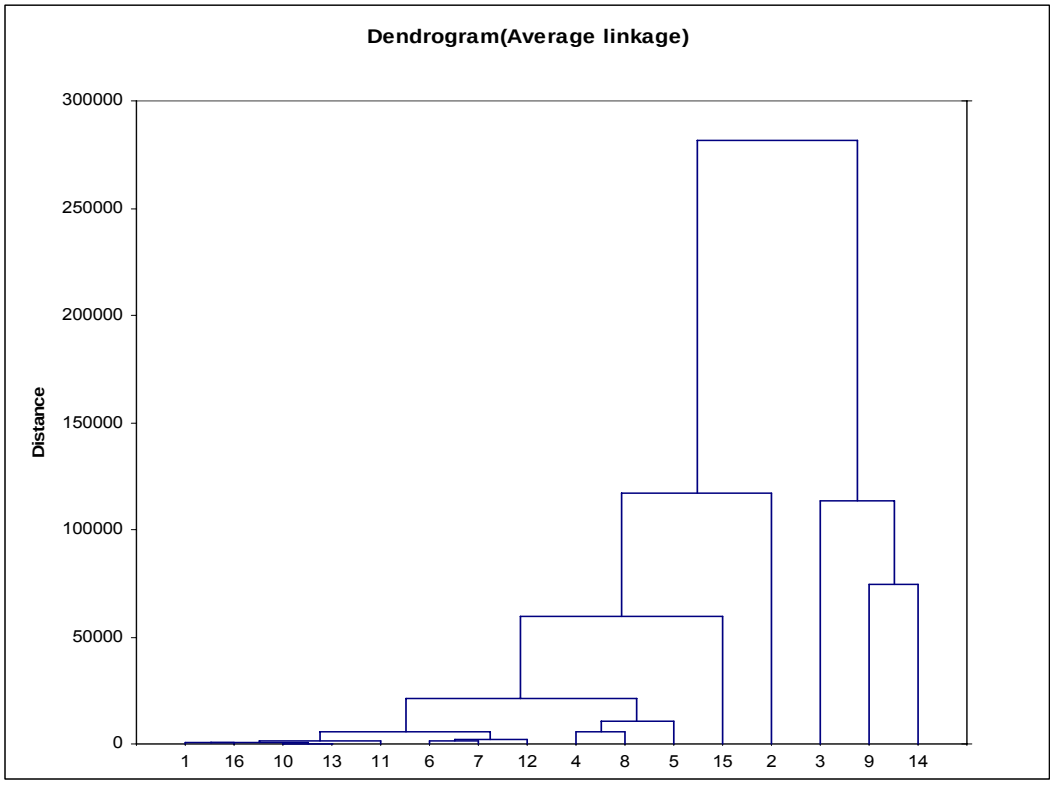

Figure 1. Dendrogram of the clusters from table 1

The dendrogram in Figure 1 shows how the numbered stocks are divided into the following four clusters: {1,4,5,6,7,8,10,11,12,13,15,16}, {2}, {3}, {9,14}. (Table 2) The last cluster is composed by two stocks that have the least prices, the greatest amounts traded and positive change. They are the most interesting for the investor. The

second and the third cluster consist of only one stock. They have approximately equal prices and high amounts of them are traded but they differ from each other because stock 2 has positive change but stock 3 has negative change. The rest of stocks are grouped in another cluster. So this method is a good way to combine stocks that are preferred by the investors.

Conclusion

Data mining software allows users to analyze large databases to solve business decision problems. Data mining is, in some ways, an extension of statistics, with a few artificial intelligence and machine learning. Like statistics, data mining is not a business solution, it is just a technology. In this article it has been shown how a hierarchical clustering method can support an investor decision to choose stocks which can pretend to be participants in an investment portfolio by using a data mining tool. So the identification of clusters of companies of a given stock market can be exploited in the portfolio optimization strategies.

Bibliography

1. G, Nitin, R. Patel, P. C. Bruce. Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner. Hardcover, 2007.
2. Chris Westphal, Teresa Blaxton, Data Mining Solutions, John Wiley, 1998.
3. A.K.Jain, R.C. Dubes. Algorithms for Clustering Data. Englewood Cliffs, NJ: Prentice Hall, 1988.
4. L. Kaufman, P.J. Rousseeuw. Finding Groups in Data: An Introduction to Cluster Analysis. New York: John Wiley&Sons, 1990.
5. J.A. Harigan. Clustering Algorithms. New York: John Wiley&Sons,1975.
6. A.K. Jain, M.N. Murty, and P.J. Flynn. Data clustering: A survey. ACM Comput. Surv., 31:264-323, 1999.
7. J. Han, M. Kamber. Data mining: Concepts and Techniques, Morgan Kaufmann, 2000.

Authors' Information

Vera Marinova-Boncheva - Institute of Information Technologies, Bulgarian Academy of Science, Sofia-1113, Bulgaria; e-mail: vboncheva@iit.bas.bg