HIERARCHICAL CLUSTERING ALGORITHMS AND DATA SECURITY IN FINANCIAL MANAGEMENT

PREDA Bianca [1], ȘERBAN Mariuța [2], ȘTEFAN Raluca-Mariana [3]

[1] Faculty of Financial Accounting Management, Spiru Haret University
[2] Doctoral Studies, Pitești University, Pitești, Romania
[3] Institute of Doctoral Studies Academy of Economic Studies, Bucharest

***Abstract**:*

Abstract: Necessity of building typologies appears in the most various domains because they offer vast possibilities to analyse and interpret specific phenomena. An economic typology can be build based on existing similarities and dissimilarities between the objects of a given set of data.

Organizing data in structures having a high level of efficiency correlated with a low level of cost requires data categorization and also data security. Cluster analysis is a data classification technique that constitutes an efficient exploratory analysis instrument.

Financial management responsibilities include making decisions based on different analysis tools. This paper presents the results of hierarchical clustering algorithms applied over an economic dataset that provides useful description of secured data for decision makers by comparison.

***Keywords**: cluster analysis, hierarchical clustering algorithm, economic data security, financial management*

***JEL Classification**: A12, C15, C38, C52, C53, C63, C88, E22, M41*

[1] *Associate Professor Ph.D., Faculty of Financial Accounting Management, Spiru Haret University, Bucharest, Romania, bianca_preda1974@yahoo.com*
[2] *Assistant Professor Ph.D. Student, Doctoral Studies, Pitești University, Pitești, Romania, mariuta_serban@yahoo.com*
[3] *Assistant Professor Ph.D. Student, Institute of Doctoral Studies Academy of Economic Studies, Bucharest, Romania, rstefan2012@yahoo.com*

1. Introduction

As long as it involves data classification based on their features and their natural tendencies to group, unsupervised learning remains an important technique utilized in data analysis. Cluster analysis is a frequently used method in very numerous and various domains and the economic field makes no exception due to the complex and large sets of data. Searching for homogeneous groupings of data based on available data is a challenge for everyone, in every domain, but the economists have a tough job nowadays struggling to deal with economic crises and numerous huge amounts of data that have to be interpreted.

Cluster analysis specific techniques are very much necessary and useful in any process of data analysis, not just the ones that involves directly the necessity to classify data. For example, the importance of using these techniques is increased for those processes in which the quantity of information that needs processed is so varied and diverse that it can become impossible to extract what is rule, what is essential and what is significant, if the adequate instruments are not used to synthesize and summarize raw information.

Data that cluster analysis intends to group can be numerical, categorical or mixed: both numerical and categorical. Clustering data implies choosing between the two most common used methods to group data: hierarchical methods and partitioning methods. The solutions obtained as a result of applying hierarchical classification algorithms include many variants of objects classification, each of them containing cluster structures having a variable number of clusters. Such clusters structures are named multilevel cluster structures.

Hierarchical classification algorithms provide many solutions of multilevel type. These kinds of solutions are called cluster hierarchies and differ by the number of clusters that are included and by clusters aggregation level. Classical representation of a hierarchy is given by the shape of a tree and its name is dendrogram. The dendrogram root consists of a single cluster that contains all elements and the dendrogram leaves are corresponding to individual elements.

The most synthetic solution of a cluster structure obtained using hierarchical classification method formed by a single cluster that includes all

analysed objects. The most detailed solution of a hierarchical cluster structure includes a maximum number of clusters that is equal to the number of analysed objects, each cluster containing a single object. That means that the possible number of solutions from a hierarchical cluster structure is smaller by 1 than the number of classified objects. This number is determined by the number of hierarchical levels of the solution.

Choosing one of the cluster structure solutions that hierarchical algorithms provide remains the researcher privilege and is made based on the analysis objectives.

2. Hierarchical Clustering Algorithms

Hierarchical clustering represents an important technique for unsupervised learning. The hierarchical clustering methods are divided in two subclasses: agglomerative and divisive.

Agglomerative hierarchical clustering is based on bottom up approach and divisive hierarchical clustering works on top down approach (Agarwal et al., 2010). Hierarchical algorithms solve the basic problem of making data clusters by considering a set of m objects. Between each pair of these m objects a symmetric proximity measure is given or computed and the values form an m×m proximity matrix.

The concept used for hierarchical clustering is related to how close or how far two objects from a given set of data are, meaning the proximity measure defined for clusters. Based on that a list of seven most commonly used options is made as it follows:

- Complete linkage. In this case, the maximum proximity value is attained for pairs of objects within the union of two subsets, minimizing the maximum link.

- Single linkage. Here, the minimum proximity value attained for pairs of objects, where the two objects from a pair belong to separate classes is obtained minimizing the minimum link.

- Average linkage. The average proximity over the pairs of objects, defined across separate classes results by minimizing the average.

- Centroid linkage. The proximity between two clusters is computed as the distance between the clusters centroids.

- Median linkage. The proximity between two clusters is only the Euclidean distance computed for their weighted centroids.

- Ward linkage. The proximity between two clusters is defined as the increase in the squared error resulted after two clusters have merged and is adequate only for Euclidean distances.

- Weighted linkage. Weighted average distance is used to compute the proximity between two clusters.

Beside Euclidean distance that is recommended for uncorrelated variables with equal variances, there are many other distances that can be used:

-standardized Euclidean distance;
-city block metric;
-Minkowski metric;
-Chebychev distance;
-Mahalanobis distance;
-cosine distance etc. (www.mathworks.com).

Data standardization is recommended generally before applying hierarchical clustering algorithms. Mahalanobis distance is usually used in cluster analysis due to the fact that is a statistical distance and that it can compensate inter correlation between variables.

3. Financial Management and Security of Economic Data

In a world of globalized informational technologized economy, to progress and to overcome the challenges of an economic crisis means the management takes the right decisions using a certain infrastructure. This implies hardware and software for informatics systems based on new database management systems that involve storing huge amounts of digital data that need to be used. The essential elements are data and information.

Processing and using data become a lot easier lately, but it implies choosing adequate special technology devices and the hard work comes when management specialists need to sort information in order to make a decision. That is why a company resources management needs to be made with an increasingly regard to necessity of utilizing electronic computing technology by implementing modern and efficient economic informatics systems.

The role that financial management has, supposes knowledge of key documents used within a company, explaining the structure of annual financial statements, classification of accounting documents, computing performance indicators etc. Another financial management competence regards choosing

the proper method to apply so that with a minimum effort a maximum efficiency can be made.

Lately, financial personnel provide not only accounting and financial information (Buga-Stancu, 2011). They are able to offer to users or decision makers some other data and information. Security data services are involved in order to provide trusting and accurate data to business environment. Financial data needs to be protected from any threats, vulnerabilities and attacks they can be subject to while they are stored or transmitted to users or beneficiaries.

Financial management needs to give their assent to purchase a very good and adequate system for data security. The security system that will protect data stored in databases must comply with the specific operations that those data are used for. There are a lot of discrepancies between theoretical and practical advances made in data security and also in clustering.

In order to bridge the gap between the theoretical advances that have been made and the existing software implementations, which are widely used in different fields, Daniel Mülner (2011) found that suboptimal algorithms are used instead of improved methods suggested by the theoretical works. The numerous clustering schemes differ in the way that inter-cluster dissimilarity measure is updated after each step.

4. Financial Management and Hierarchical Algorithms Used to Classify Economic Data

The selection and classification of a set of indicators which reflect the level of aggregate economic activity remains a major interest of business cycle analysts. Decision makers are interested in grouping similar economic behavior based on data that represents their characteristics.

Understanding an economic phenomenon or its utility makes cluster analysis a very reliable instrument. As a result, this technique has been applied in various domains for different practical or theoretical reasons: social sciences, statistics, biology, information retrieval, pattern recognition, marketing and the list get longer every day due to benefits that clustering brings.

Since the beginning of the last century methods of object classification have been applied, especially but not only in biology for understanding purposes and in 1975 methods of numerical taxonomy were

used to classify economic indicators into various groups (Broder, Schoepfle, 1975).

Worldwide organizations provide divers and numerous measures and indicators data that can be utilize by researchers in scientific studies. An application of cluster analysis on 12 macroeconomic variables data was conducted regarding the countries competitiveness (Akkucuk, 2011) and a number of criteria were defined: cluster variability, size of the clusters and agreement of found solutions.

Classifying countries based on macroeconomic indicators and economic policies has a long tradition and it was never an easy thing to do considering that real data can prove that specialists were wrong not considering all possible set of variants. Based on their economic institutions and policies countries have been classified into one of a small number of categories (Ahlquist, Breunig, 2009) and this classification generated controversy.

5. Hierarchical Algorithms Applied to Economic Data

A set of data consisting in values of macroeconomic indicators that characterize a country level of prosperity were hierarchical clustered. The primary goal of this paper is to form homogeneous groups for 26 countries by applying ward hierarchical clustering algorithm over a set of complex economic data.

The data from Table 1 is a sample of the data that corresponds to a period of 12 years, 2000-2011, regarding six macroeconomic indicators: exchange rate, inflation rate, long term interest rate, government deficit or surplus, government debt and gross domestic product. Data are expressed as percentage and were provided by European Commission site.

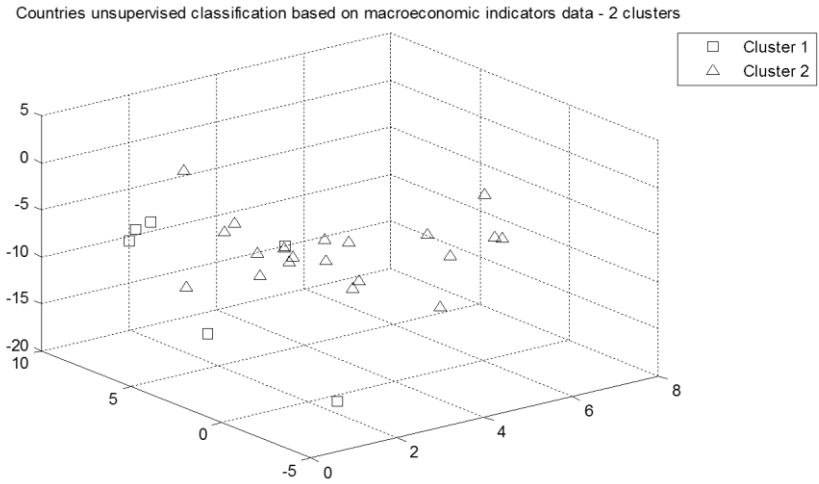
First, Ward's method was used and the default distance, Euclidean distance. The cophenetic correlation coefficient was computed and the obtained value is 0.7203. The cophenetic correlation coefficient is a measure of how accurate a dendrogram preserves the pairwise distances between the original data points. The cophenetic distance between two observations is represented in the resulted dendrogram by the height of the link at which those two observations are first joined (www.mathworks.com). That height is the distance between the two subsets of clusters that are merged by that link. This value should be very close to 1 for indicating a good solution and this measure

can be used to compare alternative cluster solutions obtained using different algorithms.

Table 1: Prosperity macroeconomic indicators data characterizing countries for year 2011

	COUNTRY	Exchange Rate	Inflation Rate	Long Term Interest Rate	Government Deficit or Surplus	Government Debt	Gross Domestic Product
1	EU (27 countries)	125.22	3.10	1.90	-6.30	85.50	2.14
2	Belgium	109.51	3.50	5.34	-2.40	96.50	2.07
3	Bulgaria	146.85	3.40	4.43	-1.90	18.00	2.66
4	Czech Republic	175.48	2.10	2.69	-3.80	40.80	1.8
5	Denmark	123.96	2.70	4.29	-2.50	45.60	1.42
6	Germany	88.56	2.50	4.39	-5.40	92.00	0.86
7	Ireland	141.15	1.20	5.24	-48.40	119.80	2.94
8	Greece	114.24	3.10	4.38	-5.40	160.50	4.44
9	Spain	122.41	3.10	4.45	-7.40	68.20	4.64
10	France	112.19	2.30	4.41	-6.70	85.60	2.29
11	Italy	121.33	2.90	5.71	-3.80	121.30	2.91
12	Cyprus	119.84	3.50	2.97	-4.50	64.50	2.24
13	Latvia	154.66	4.20	6.00	-6.90	52.70	3.1
14	Lithuania	137.00	4.10	5.23	-4.50	46.60	3.29
15	Hungary	152.04	3.90	7.64	-3.80	82.90	2.61
16	Malta	126.38	2.40	4.94	-3.50	70.20	1.38
17	Netherlands	115.68	2.50	5.49	-4.60	65.00	0.76
18	Austria	96.68	3.60	3.91	-4.70	74.10	0.83
19	Poland	101.26	3.90	5.99	-8.30	58.90	2.58
20	Portugal	115.56	3.60	5.69	-9.50	103.60	3.01
21	Romania	195.67	5.80	6.98	-4.80	38.40	2.37
22	Slovenia	110.87	2.10	4.88	-5.50	42.30	1.57
23	Slovakia	188.47	4.10	4.62	-7.40	46.50	3.32
24	Finland	109.65	3.30	2.47	-2.50	53.30	2.01
25	Sweden	94.53	1.40	2.42	1.10	36.70	2.29
26	UK	91.99	4.50	4.20	-9.10	90.20	2.11

Figure 1: Hierarchical clustering of countries based on macroeconomic data (Ward's method, Euclidean distance)

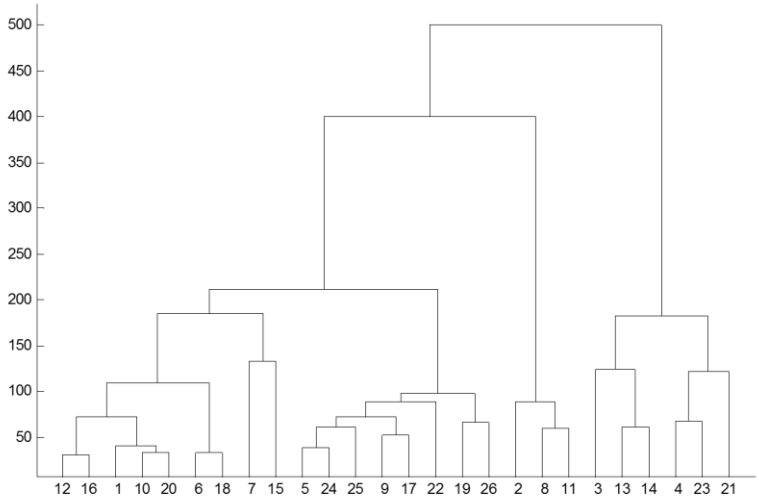


Source: Authors' Matlab output

Hierarchical clustering is a method of unsupervised classification that can provide correct solutions to an economic problem. The two clusters resulted are graphically described in figure 1. Cluster 1 has 6 objects that represent corresponding 6 countries and cluster 2 has 20 objects that correspond to the rest of the countries considered for the study.

A dendrogram is a tree drawing structure that is often used to show the way clusters were formed through hierarchical clustering. Dendrograms are often used in various fields to illustrate the clustering of observations at every step of the algorithm and figure 2 shows the solutions provided by hierarchical algorithm applied on data. The dendrogram also contains the countries that form the two clusters we were looking for.

Figure 2: Resulted dendrogram – clusters of countries based on macroeconomic indicators data



Source: Authors' Matlab output

Phenograms and cladograms represent the level of phenetic similarity due to the fact that they are based on phenetic data. In biology, there can be some differences but in any other domain there are not. Thus, some consider that phenograms and cladograms are synonymous with dendrogram, while others consider them subtypes of a dendrogram. Regardless of their point of view, for general purposes, all are diagrams.

The lines drawn at right angles to the phenogram represent values of similarity as percentages and the values are directly proportional with the degree of similarity. The statistical reliability of a phenogram was not known and because of that, a measure of its effectiveness, the cophenetic correlation coefficient, has been proposed by Sokal and Rohlf in 1962 (Broder, Schoepfle, 1975). A number of coefficients have also been proposed (Sneath, Sokal, 1973; Johnson, Wichern, 1988).

In figure 2 it can be seen the two clusters contain: first 20 countries and the second 6 countries. Input data regarding country prosperity based on six macroeconomic indicators correspond to every country. It is very easy to

identify the countries for each of the two clusters as the dendrogram displays the observation number for every object. For the cluster that contains six countries that correspond to the following numbers: 3, 4, 13, 14, 21, 23. The countries are: Bulgaria, Czech Republic, Latvia, Lithuania, Romania and Slovakia. Ward hierarchical algorithm proved its efficiency over a set of highly heterogeneous data like the economic data that was used in this case.

Cophenetic correlation coefficient was utilized to compare results of different hierarchical algorithms. Table 2 displays the cophenetic correlation coefficient values obtained as a result of hierarchical clustering algorithms.

The five used methods are showed in this table and indicate a good solution for the method that used average linkage with Chebyshev metric having a value of 0.7932. The smallest value was resulted by applying single linkage with cityblock distance, 0.6943.

Table 2: Cophenetic correlation coefficient for applied hierarchical clustering methods

Number of method of hierarchical clustering	Linkage	Distance	c
1	Ward	Euclidean	0.7203
2	Single	Cityblock	0.6943
3	Complete	Minkowski	0.7284
4	Average	Chebyshev	0.7932
5	Weighted	Minkowski	0.7766

Source: Authors' results

6. Conclusions

In order to increase future competitiveness and efficiency based on present information, financial management must use cluster analysis and economic data security. This interdisciplinary challenge can be consider one of the very good opportunities to use informational technology for making the best decision that can be made considering the complex, multidimensional economic context. Methods of clustering are useful to point out some of the particular connections that exist between data observed and their structure. The results obtained after a hierarchical algorithm was applied over economic data regarding country prosperity are very good considering how close to reality are the two clusters formed and described in this paper. Consequently,

hierarchical clustering algorithms are efficient and further research is required in this direction applying these methods in order to choose the best solution.

There are other hierarchical clustering algorithms that have been omitted here, there are a lot of economic indicators and they all need to be analysed in order to benefit from a good opportunity offered by them for us to understand the way they affect a country economy and a country categorization. Being far from excluding human intervention, further studies on other hierarchical clustering algorithms are indicated in any economic analysis in order to beneficiate the maximum efficiency from these classification methods.

7. References

- Agarwal, P.; Alam, M.A.; Biswas, R. (2010) Analysing the agglomerative hierarchical Clustering Algorithm for Categorical Attributes, International Journal of Innovation, Management and Technology, Vol. 1, No. 2, June 2010.
- Agathon, D.M.; Sava, C. (2012) A Golden Mind & Spirit analysis – The figures that tear us apart. How far are we from the developed countries from the E.U. Available at www.buzznews.ro.
- Ahlquist, J.; Breunig, C. (2009) Country Clustering in Comparative Political Economy, Max Planck Institute for the Study of Societies, Cologne, Discussion Paper.
- Akkucuk, U. (2011) A Study on the Competitive Positions of Countries Using Cluster Analysis and Multidimensional Scaling, European Journal of Economics, Finance and Administrative Sciences, Issue 37. Available at www.eurojournals.com.
- Broder, I.; Schoepfle, G. (1975) Classification of Economic Indicators: An Alternative Approach, Annals of Economic and Social Measurement 4/5. Available at http://www.nber.org/books/aesm75-3.
- Buga-Stancu, M. (2010) Servicii de asigurare a informațiilor – misiuni de audit, revizuiri, servicii conexe, Tribuna Economică, Nr. 7, 16 feb 2011.
- Cocianu, C.-L. (2006) Supervised and Unsupervised Classification for Pattern Recognition Purposes, Revista Informatica Economică, nr. 4 (40)/2006, p. 5-13.
- Heijden, F.; Duin, R.P.W.; Ridder, D.; Tax, D.M.J. (2004) Classification, Parameter Estimation and State Estimation An Engineering Approach using MATLAB, Wiley.

• Hepşen, A. (2012) Using Hierarchical Algorithms for Turkish Residential Market, International Journal of Economics and Finance, Vol. 4, No. 1, January 2012. Avaiable at www.ccsenet.org/ijef.

• Nanni, L. (2006) Cluster-based Pattern Discrimination: A Novel Technique for Feature Selection, Pattern Recognition Letters 27, Elsevier.

• Seong, S.C.; William, D.W. (2006) Effect of Using Principal Coordinates and Principal Components on Retrieval of Clusters, Computational Statistics & Data Analysis 50, Elsevier.