Journal of International Scientific Publications:
Materials, Methods & Technologies, Volume 7, Part 2
ISSN 1313-2539, Published at: http://www.scientific-publications.net

A STUDY OVER THE MAIN METHODS USED IN CLUSTER ANALYSIS

Raluca-Mariana Stefan[1]

[1] Academy of Economic Sciences, 6 Piata Romana, Bucharest 010374, Romania

Abstract

These days, almost every domain involves work with data categorization and has to apply at least one of the procedures to group diverse and heterogeneous information provided by extremely large amounts of digital data.

Clustering is adequate if the low level of data homogeneity is considered along with the lack of labeled samples.

Grouping a lot of data using cluster analysis constitutes an exploratory technique of obtaining relevant information for the specialists that can use this knowledge to make appropriate and correct decisions for short term or long term in order to increase their prediction performances.

Key words: *clustering, data representation, data characterization, clustering algorithm, prediction performance*

1. INTRODUCTION

The complexity of economic and social phenomena involves working with a lot of data that need categorization. A system of forms or objects recognition is formed by a set of activities, rules, procedures, methods and techniques which have the main goal to identify the class a form or an object belongs to. Classifying data implies the choice between two procedures: supervised classification and unsupervised classification. The first procedure categorizes data with the help of a training sample extracted from initial data based on which the rest of data will be classified then. Unsupervised learning classifies raw data into clusters using hierarchical and partitioning methods. Each of these solutions to group data has advantages as well as limitations.

Prediction performance can be improved if it is based on specialists decisions made with the cluster analysis due to the low level of economic data homogeneity. Grouping a lot of data using cluster analysis constitutes an exploratory technique of obtaining relevant information for the decision makers that can use this knowledge to make appropriate and correct decisions for short term or long term in order to increase their prediction performances.

2. UNSUPERVISED LEARNING

Cluster analysis represents an exploratory way of learning from data. It has to identify the grouping trends of data, to decide on the sound clusters as well as to validate somehow the resulted structure. "The identification of the grouping tendency existing in a data collection assumes the selection of a framework stated in terms of a mathematical model allowing to express the similarity degree between couples of particular objects, quasi-metrics expressing the similarity between an object and a cluster and between clusters, respectively" (Cocianu, 2006).

A cluster represents a subset of mostly similar forms or objects. That brings the general criterion used to cluster forms or objects: the objects that form a cluster have the property that the similarity degree

of any two objects belonging to the same cluster is bigger than the similarity degree for any two objects that belong to different clusters.

Cluster analysis is fundamentally different from the statistical methods being an important and efficient instrument of exploratory analysis (Ruxanda, 2009).

3. IDENTIFYING CLASSES OF METHODS USED IN CLUSTER ANALYSIS

Cluster analysis is one of the main methods used for multidimensional data analysis. Input data are supposed to have a specific form and that depends on the methods, instruments and applied procedures. Data analysis which includes cluster analysis works with information consisting in computed statistics like mean, variance, covariance and correlation from primary data.

A cluster analysis generalization cannot be done if the numerous number of clustering methods is considered (Stefan, 2012).

3.1. Main Methods of Data Representation and Characterization

The basic fundamental concept used in classification is *data*. One of the factors that are influencing the clustering algorithm performance is data representation. If the data representation is adequate the clusters that are to be found will be well isolated and have compactness. Unfortunately, there is no universal choice for data representation and the data feature space is dependent on domain knowledge (Jain, 2010).

Data are obtained after the measurement and observation processes from investigated reality. In order to use data in a cluster analysis these must be processed. Methods and techniques are represented by rules, principles and data analysis, processing and interpretation procedures (Ruxanda, 2001). Most of the significance and relevance of the results is directly proportional to the accuracy of applied procedures on data. The choice of a method over another depends on *data representation* and *characterization.*

Data may differ based on their source, type and nature. Regardless of their diversity and variety data can be categorize as quantitative data, qualitative data and mixed data depending on the activities they express.

Having a group of data we are interested to know and to exploit the common features of that group so that a rule of data behavior results. Properties of data are called *characteristics* or *attributes* and a single data can have only one or more attributes. When data have many characteristics a multidimensional analysis is needed. Features nature divide data in data with *qualitative* and *quantitative characteristics*. Symbols are used to map reality processes data in order to have a better control over a large population of data and their characteristics.

In order to create a proper image of the changing nature of an economic process or phenomenon it is commonly used the concept of *variable.* Variable is an abstraction of the possible set of values that a data characteristic can register (Ruxanda, 2001).

Generally, data can be represented as matrices. The main matrices used to represent primary data are *observations matrix, contingency matrix* or table and *proximity matrix* or table. An observations matrix is a rectangular table having a number of rows that represent the forms or objects and a number of columns that represent forms or objects characteristics. A contingency matrix or a contingency table is basically a way to record and display the relationships between more than two categorical variables. A square matrix having as elements data regarding certain measure of distance or similarity between the objects corresponding to rows and columns is named proximity matrix.

3.2. Algorithms Used for Data Clustering

Clustering can be defined as a multidimensional exploratory analysis that has the goal of organizing the data into relevant information and knowledge. Main techniques used for cluster analysis are divided in two large categories: hierarchical and non-hierarchical algorithms. Hierarchical algorithms provide multi-level solutions, cluster hierarchies, and non-hierarchical algorithms provide single-level solutions. Non-hierarchical algorithms are also named partitioning algorithms.

Hierarchical algorithms are also divided into two classes based on their nature and the offered solution: hierarchical classification methods by aggregation and divisive methods.

We can number some of the most used hierarchical aggregation algorithms: simple aggregation method, complete aggregation method, medium aggregation method, centroid method, Ward's method. The divisive methods contain a small number of algorithms due to the existence of a large number of hierarchical aggregation methods. There are two well-known divisive methods, one is DIANA algorithm and the other one is based on monothetic classification method but, unfortunately it can be used only if the variables describing the objects are binary (Ruxanda, 2009).

Some of the partitioning algorithms that are frequently used to cluster data are: k-means algorithm and k-medoids algorithm.

K-means algorithm is also a partitioning method to group observations. It uses the function *kmeans* to partition given unlabeled data in *k* mutually exclusive clusters. The algorithm assigns each observation to a cluster and the used function returns the clusters indices.

4. CLUSTERING ECONOMIC DATA USING K-MEANS

4.1. Why Use a Non-hierarchical Algorithm? – Advantages

A partitioning algorithm used to cluster a group of observations like *k-means* works on actual data creating a single cluster level. This gives k-means clustering the advantage of being more suitable than a hierarchical algorithm, for a large amount of data. A hierarchical algorithm operates on the larger set of dissimilarity measures (mathworks.com).

Another advantage is given by the fact that the number of clusters can be set and changed along with the data classification in order to obtain the best solution. This consists in the main characteristic of a partitioning algorithm.

4.2. Using K-means Algorithm to Cluster Economic Data

The data consists in 26x72 observations regarding six macroeconomic indicators, five prosperity indicators plus Gross Domestic Product (GDP), for a period of 12 years, 2000-2011. Data is represented by percentage.

K-means Matlab procedure was applied in order to classify countries. The real data regarding these economic indicators are grouped in two classes: prosper countries class and non-prosper countries class.

Knowing the fact that these 26 countries form in reality two clusters K-means algorithm can be utilize to see if the number of clusters is changed, how well-separated are the new clusters. Given the complexity of economic data a cluster analysis can explore reveal some new aspects. One of these aspects can be that the countries can be better separated into three clusters based on observations.

ISSN 1313-2539, Published at: http://www.scientific-publications.net

Table 1. Macroeconomic indicators for prosperity and GDP data used for clustering (first part)

Country/ Prosperity indicator	Exchange Rate	Inflation Rate	Long Term Interest Rate	Government Deficit or Surplus	Government Debt	Gross Domestic Product
EU (27 countries)	125.22	3.10	1.90	-6.30	85.50	2.14
Belgium	109.51	3.50	5.34	-2.40	96.50	2.07
Bulgaria	146.85	3.40	4.43	-1.90	18.00	2.66
Czech Republic	175.48	2.10	2.69	-3.80	40.80	1.8
Denmark	123.96	2.70	4.29	-2.50	45.60	1.42
Germany	88.56	2.50	4.39	-5.40	92.00	0.86
Ireland	141.15	1.20	5.24	-48.40	119.80	2.94
Greece	114.24	3.10	4.38	-5.40	160.50	4.44
Spain	122.41	3.10	4.45	-7.40	68.20	4.64
France	112.19	2.30	4.41	-6.70	85.60	2.29
Italy	121.33	2.90	5.71	-3.80	121.30	2.91
Cyprus	119.84	3.50	2.97	-4.50	64.50	2.24
Latvia	154.66	4.20	6.00	-6.90	52.70	3.1

Table 1. Macroeconomic indicators for prosperity and GDP data used for clustering (second part)

Country/ Prosperity indicator	Exchange Rate	Inflation Rate	Long Term Interest Rate	Government Deficit or Surplus	Government Debt	Gross Domestic Product
Lithuania	137.00	4.10	5.23	-4.50	46.60	3.29
Hungary	152.04	3.90	7.64	-3.80	82.90	2.61
Malta	126.38	2.40	4.94	-3.50	70.20	1.38
Netherlands	115.68	2.50	5.49	-4.60	65.00	0.76
Austria	96.68	3.60	3.91	-4.70	74.10	0.83
Poland	101.26	3.90	5.99	-8.30	58.90	2.58
Portugal	115.56	3.60	5.69	-9.50	103.60	3.01
Romania	195.67	5.80	6.98	-4.80	38.40	2.37
Slovenia	110.87	2.10	4.88	-5.50	42.30	1.57
Slovakia	188.47	4.10	4.62	-7.40	46.50	3.32
Finland	109.65	3.30	2.47	-2.50	53.30	2.01
Sweden	94.53	1.40	2.42	1.10	36.70	2.29
UK	91.99	4.50	4.20	-9.10	90.20	2.11

Being a partitioning method for clustering, k-means allows us to set the number of clusters from the start and to run the procedure. The number k can be changed in order to establish the best separation of data. Initially, *k-means* algorithm was applied for k=2.

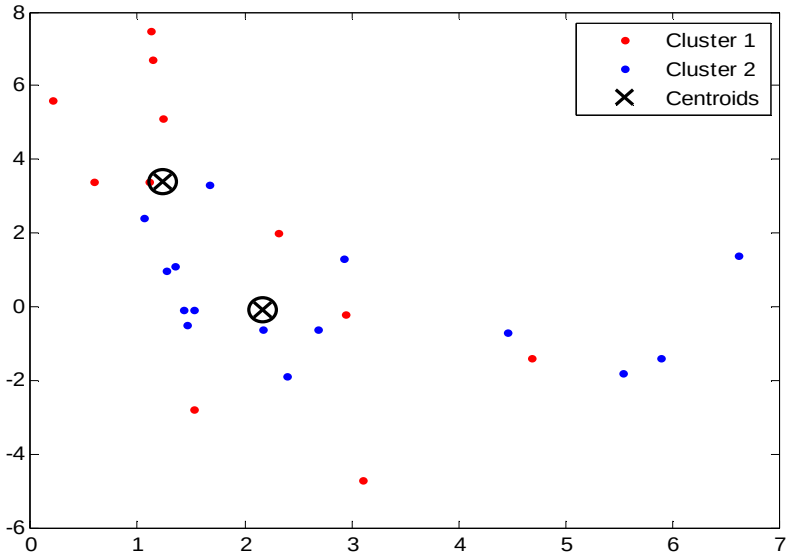


Fig. 1. K-means algorithm for k=2 (author's Matlab output)

The 26 countries used for analysis formed two clusters: first cluster has 11 countries – 6 countries are closest to the first centroid, and the second has 15 countries – 9 countries are closest to the second centroid.

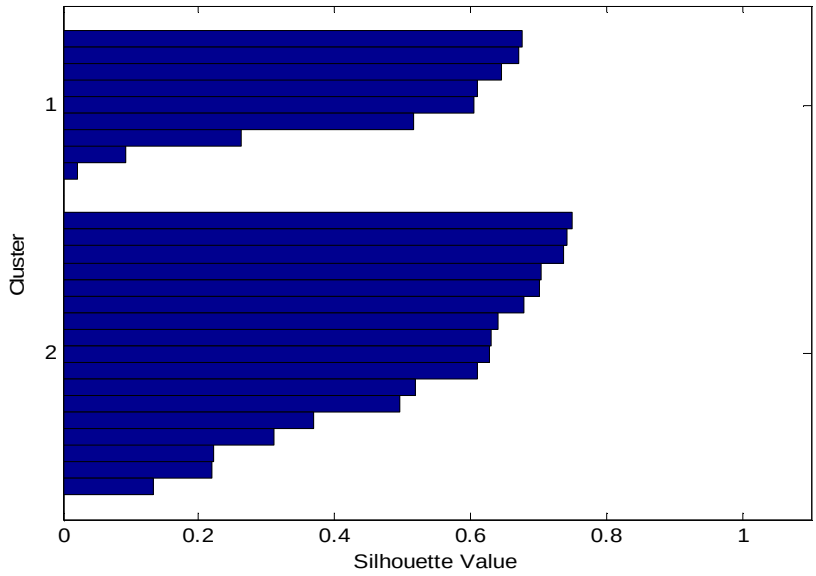


Fig. 2. Silhouette plot (author's Matlab output)

ISSN 1313-2539, Published at: http://www.scientific-publications.net

The silhouette plot was used to see how well are the data grouped into two or three clusters. The silhouette plot displays a closeness measure between each cluster point to the other points in the neighboring clusters.

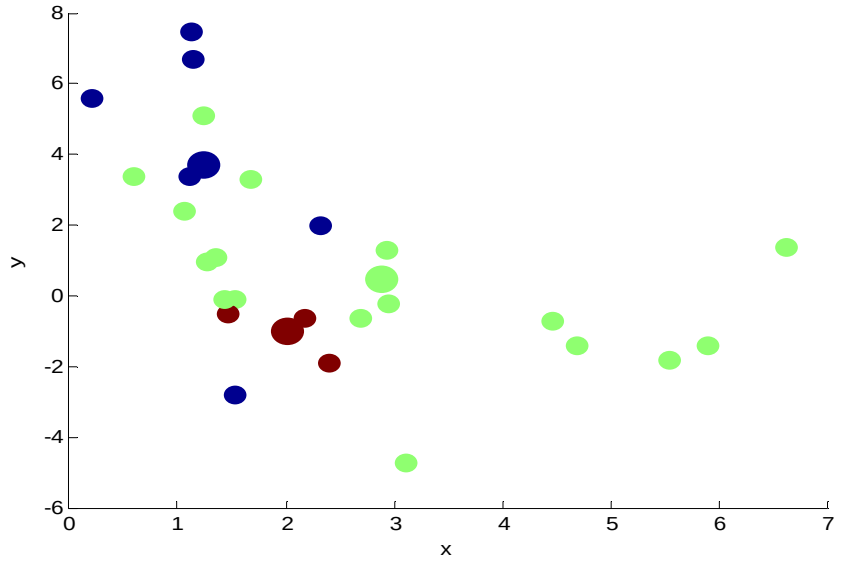


Fig. 3. K-means algorithm for k=3 (author's Matlab output)

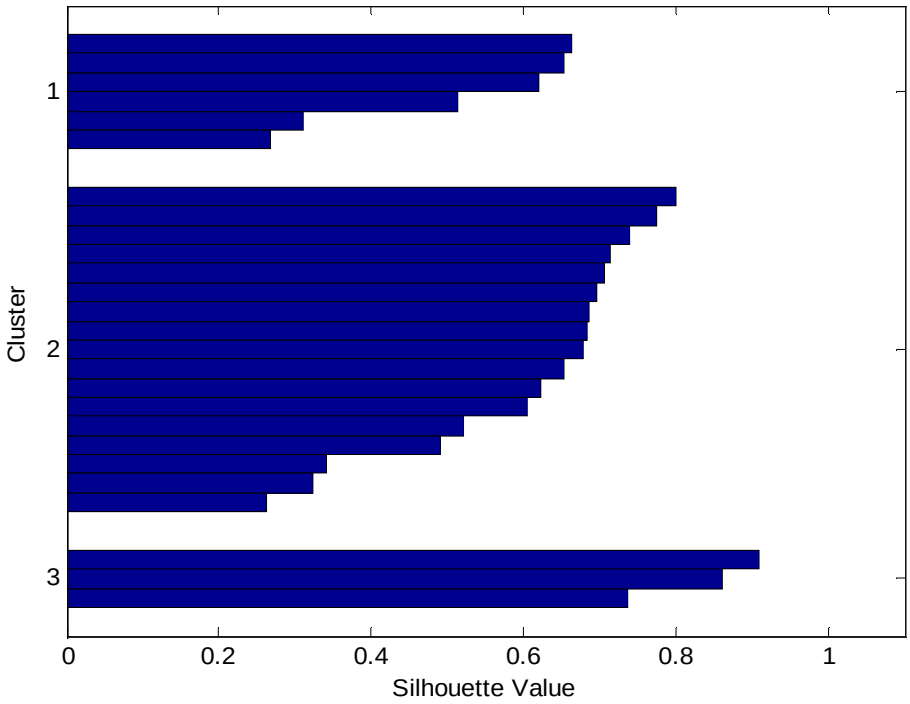


Fig. 4. Silhouette plot (author's Matlab output)

Data grouped in three well separated clusters having 6 countries in the first cluster, 17 countries in the second cluster and 3 countries in the third cluster. The second cluster has a larger part with silhouette values that are greater than 0.6, indicating a good cluster separation. The third cluster has all values greater than 0.6 and that indicates a very good separation from its neighbors.

The results obtained prove that the data regarding prosperity indicators can categorize better the 26 countries into three groups rather than the real two groups. This can be interpreted in terms of economic prosperity of a country that macroeconomic indicators data categorize not only two groups of prosper or non-prosper country, and a third category is formed by emergent countries. K-means algorithm revealed a new group of countries that was determined based on the prosperity economic data. In the context of world global economy specialists must take into consideration that when it comes to the level of prosperity of a country it is not always a case of "black" or "white", it can be also "grey" and their decisions can be affected decreasing prediction performance.

5. CONCLUSIONS

Decisions that are made in the present based on the past economic evolutions and data will affect all of us in the future. The increasingly importance given to cluster analysis in the economic field helps the people involved in the global economic process to make proper and correct decisions. The advantage of automated studying a large amount of data with no intervention makes the specialists think twice weather to use statistical methods or cluster analysis.

Obtaining relevant and efficient information by clustering economic data means saving a lot of time because no prior information is necessary.

We propose further research and studies using macroeconomic indicators data that characterize and affect the economy of a country in order to get important information so that the level of future prediction performance can be increased.

REFERENCES

1. Jain, A.K. (2010) Data clustering: 50 years beyond K-means, Pattern Recognition Letters, Volume 35, 8 Issue, Elsevier B.V., pp. 651-666.

2. Ruxanda, G. (2009) Multidimensional data analysis – Doctoral School Course Notes, 133 pp.

3. Murtagh F., Legendre P., Ward's (2011) Hierarchical Clustering Method: Clustering Criterion and Agglomerative Algorithm, http://arxiv.org/abs/1111.6285 .

4. Jain, A.K., Dubes, R.C. (1988) Algorithms for Clustering Data, Prentice Hall Inc., 320 pp.

5. Van der Heijden F., Duin R.P.W., de Ridder D., Tax D.M.J. (2004) Classification, Parameter Estimation and State Estimation: An Engineering Approach Using MATLAB, Wiley, 423 pp.

6. Pedrycz W. (2005) Knowledge-Based Clustering: From Data to Information Granules, Wiley, 316 pp.

7. Tibshirani R., Walther G. (2005) Cluster Validation by Prediction Strength, American Statistical Association, Institute of Mathematical Statistics and Interface Foundation of North America, Journal of Computational and Graphical Statistics, Volume 14, Number 3, Pages 511-528.

8. Growe G. (1999) Comparing Algorithms and Clustering Data: Components of the Data Mining Process, Department of Computer Science and Information Systems, Grand Valley State University.

9. Chae S., Warde W. (2006) Effect of using principal coordinates and principal components on retrieval of clusters, Computational Statistics and Data Analysis, Volume 50, Elsevier B.V., pp. 1407-1417.

10. Ruxanda G. (2001) Analiza datelor, Ed. ASE, București.

10. Nanni L. (2006) Cluster-based pattern discrimination: A novel technique for feature selection, Pattern Recognition Letters, Volume 27, Elsevier B.V., pp.682-687.

11. Fränti P., Virmajoki O. (2006) Itterative shrinking method for clustering problems, Pattern Recognition, Volume 39, Elsevier B.V., pp. 761-775.

12. Jain A.K., Murty M.N., Flynn P.J. (2000) Data Clustering: A Review, ACM Computing Surveys, Vol. 31, No. 3.

13. Halkidi M., Batistakis Y., Vazirgiannis M. (2001) On Clustering Validation Techniques, Journal of Intelligent Information Systems, 17:2/3, pp. 107-145, Kluwer Academic Publishers, Netherlands.

14. Cocianu C. (2006) Supervised and Unsupervised Classification for Pattern Recognition Purposes, Revista Informatica Economică, nr.4 (40)/2006.

15. Stefan R.M. (2012) An Overview of the Frequently Used Algorithms to Build Clusters, International Conference Materials, Methods & Technologies, Sunny Beach, Bulgaria, published in Journal of International Scientific Publications, vol. 6, part 2, pp. 163-174, Ed. Info Invest Ltd.

16. www.mathworks.com .