HOW APPROPRIATE IS THE CHOICE OF APPROPRIATE METHOD IN CLUSTERING ANALYSIS?

Necati Alp Erilli

Abstract

In order to make a general diagnosis or definition, it is necessary to examine the structures of the data we observe. By looking at these structures, we can place data into similar groups, contents, and other characteristics. In general, these studies are called classifications. The most commonly used method in classification studies is cluster analysis. Cluster Analysis is a method that enables to classify the units examined in a study by gathering them into certain groups according to their similarities, to reveal the common features of the units and to make general definitions about these classes. There are different clustering methods suitable for different data structures in the literature. These methods can be classified as Crisp Clustering, Fuzzy Clustering, Anti-Clustering, Biclustering and Gray Clustering. To find a complete answer for the best clustering method question is very hard. In this study, the performances of the 5 main clustering methods briefly introduced above in data sets that are frequently used in the literature and where cluster memberships are certain are examined. The estimation percentage of these methods in determining cluster memberships and the success of assigning new memberships correctly were compared.

Key words: clustering analysis, classification, fuzzy clustering, biclustering, anti-clustering

JEL Code: C38, C30

Introduction

Although its structure in nature is known, the desire to investigate the underlying causes has never decreased. From molecular medicine studies to nanotechnology, human beings keep their research curiosity up to date and learn new information almost every day. New knowledge can be obtained from observations, research, or experiments. To make a general diagnosis or definition, it is necessary to examine the structure of the observed data. By examining these structures, we can place data into similar groups, contents, and other characteristics. Generally, these studies were called as classification. The subject of classification is to investigate relationships within a set of “objects” to determine whether data can be validly summarized by fewer classes (clusters) based on their similarity. Classification can be summarized as the task of dividing individuals or objects into similar groups according to their similarities. The classification of individuals (observations) into classes according to their similarities is called Q Analysis Techniques (Discriminant Analysis, Cluster Analysis, Logistic Regression, Multidimensional Scaling), and the grouping of related variables is called R analysis techniques (Principal Component Analysis, Factor Analysis, Correlation Analysis). Perhaps the most preferred of these techniques is Cluster Analysis.

Cluster analysis is a comprehensive concept that refers to the large number of operations that can be used to create classifications. These operations are empirically derived from the groups or clusters created by this method. Clustering is a multivariate statistical method that starts with datasets containing information about a sample consisting of units and rearranges these units into similar (homogeneous) groups (Hair et al., 2009). Cluster analysis is a group of multivariate techniques, the main purpose of which is to group objects according to their properties. Cluster analysis investigates the cluster structure and number of clusters of different types of data. While finding the grouping structure, it aims to have the same structure for observations within the cluster and different structures for observations between clusters. For observations, these decompositions were made using similarity and difference criteria. These criteria are based on distance, correlation, or similarity/dissimilarity measures.

One of the most important problems encountered in classification studies is the application of this method will be applied to the data at hand. For datasets with a small number of variables, examining the binary graphs of the variables may be preferable to determine the number of clusters or cluster structures. However, when the number of variables is large or the datasets in which the observation values are nested, it is possible that the desired or expected results cannot be obtained. Several clustering methods based on different data types have been proposed in the literature. In contrast to classical clustering methods, these methods can show computational differences according to the data type or the algorithm used. It is difficult to find a complete answer to the question of which of these methods gives more successful results. In this study, the performances of five main clustering methods, which are frequently used in the literature and briefly introduced in the second section, are examined in datasets where cluster membership is certain. The estimation percentages of these methods for determining cluster memberships and the success of correctly assigning new memberships were compared.

1 Clustering Analysis

Cluster Analysis is a method that allows to classify the observations examined in a study into groups close to each other according to some characteristics, to determine the common features of the observations and to make general statistics about these classes. The purpose here is to classify ungrouped data according to their similarities and to help researchers obtain appropriate and useful summative information (Hair et al., 2009). In other words, similar data are collected in the same group or cluster, considering the similarities between the data. Cluster Analysis is a collection of methods that helps to separate individuals or variables in the data matrix whose natural groupings are not known precisely into subsets that are like each other. In addition, cluster analysis is used in the determination of real types, facilitating model fitting, preliminary estimation for groups, testing hypotheses, clarifying the data structure, reducing data size, and finding outliers.

In classical cluster analysis, an observation is either a member of a cluster or not. Classical clustering analysis is grouped into two sub-headings: hierarchical and non-hierarchical. Hierarchical methods are divided into two sub-methods, aggregated and divisive, whereas non-hierarchical methods are divided into four sub-methods: centroid, density, distribution, and connectivity (Everitt et al., 2011). A good clustering method will produce high quality clusters with high intra-class similarity and low inter-class similarity. The quality of a clustering result depends on both the similarity measure used by the method and its implementation. Besides the quality of a clustering method is also measured by its ability to discover some or all the hidden patterns.

1.1 Miscellaneous clustering methods

In recent years, several successful studies have been conducted using different clustering analysis methods. The most widely used of these methods is the fuzzy clustering analysis. The use of fuzzy sets for clustering was proposed by Bellman et al. (1966). Fuzzy Clustering analysis use fuzzy logic to cluster data and an object can be classified into more than one cluster in these methods. Because this type of algorithm handles the uncertainty of real numbers, it helps develop clustering patterns suitable for daily life experiences. This approach emerges as a suitable method if the clusters are not clearly separated from each other or if some units are undecided in cluster membership. The concept of fuzzy set can be summarized as functions that determine each unit between zero and one, which is determined as the membership value of the related units in the clusters. Units close to each other have high membership degrees and are in the same cluster. Fuzzy Cluster analysis provides membership values that are useful for interpretation, are flexible in the use of distance, and can be combined with numerical optimization when some of the membership values are known. Fuzzy Clustering Algorithms are divided into two categories. One uses fuzzy relations within fuzzy clustering, and the other uses the objective function. While fuzzy relations-based clustering deals with the relational structure between original individuals, algorithms based on the objective function aim to solve the clustering problem by transforming it into an optimization problem. In this method, the objective function is used to measure uniqueness within the cluster, and this objective function is minimized to obtain the best division.

The term binary clustering was first used by Mirkin (1996) in his study. In this study, the concept of binary clustering is explained with the definition of "simultaneous clustering of both row and column clusters in a data matrix". Binary clustering allows rows and columns in a data matrix to be included in more than one binary cluster. Thus, it allows a gene or condition to be identified by more than one set of methods. This added flexibility accurately reflects the reality of genes functionality and overlapping factors in tissue samples and experimental conditions. Microarray studies have become an increasingly important field in medical and biological research and represent an important field in the identification of gene groups, analysis and evaluation of gene data. Conventional clustering methods divide an expression matrix into submatrices that cover the entire set of features, since they give equal weights to all conditions. However, for certain biological or health problems, not all genes can be expected to produce results in the same way under all available conditions. To account for this, binary clustering approaches can perform grouping in both dimensions for genes and conditions simultaneously. It would thus allow subsets of genes to produce similar results under certain subset of conditions. Also, if a gene participates in more than one differentially regulated pathway, this gene is expected to be included in multiple clusters, which cannot be achieved by conventional clustering (Cheng and Church, 2000).

The gray cluster analysis is based on the Gray System Theory (GST) developed by Deng (1982). GST is a theory that allows the examination of problems involving small samples and limited information (Liu and Forrest, 2006). Clustering for the separation of objects, it can be divided into gray cluster analysis and gray–white weight function cluster analysis. The grey-white weight function cluster analysis was used to check whether an observation belonged to the different categories identified. Gray cluster analysis is used for similar factors to simplify complex systems (Ke et al., 2012). Gray cluster analysis is based on the whitening weight function of gray numbers to classify observations or to include objects in a defined group. Gray cluster analysis is used to simplify complex systems by classifying similar factors (Liu and Forrest, 2006). The grey clustering method using whitening weight functions is mainly applied to test whether the observation objects belong to predetermined classes. The Gray Clustering Method was developed to classify observation objects into identifiable groups and can be performed using gray incidence matrices or whitening weight functions. The advantage of grey cluster analysis is that it handles the sample distribution of the clustering object without any special requirements, and more useful clustering results can be obtained through a simple computational process (Ke et al., 2012).

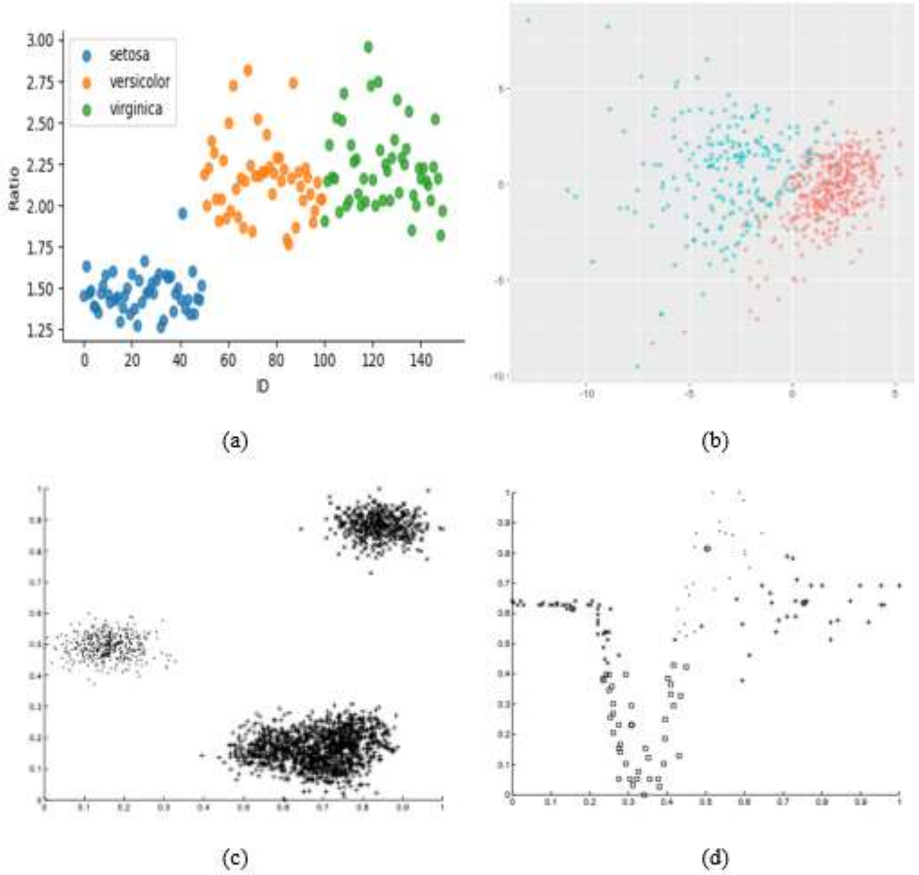
In many cluster analysis methods, the aim is to divide the studied data into homogeneous and well-separated groups among clusters. Thus, observations within the same cluster should be similar to each other (low in-group variance) and different among other clusters (high intergroup variance). While much less common, there are also practical situations where the goal is to divide a set of objects into groups with high in-group and low in-group variance. That is, objects within a group should be as diverse as possible, while pairs of groups should be very similar in their composition. This type of partitioning is known as anti-clustering, a term originally coined by Spath (1986) as anti-k-means partitioning, where the goal is to maximize the in-group sum of squares criterion rather than minimize it (Brusco et al., 2019). When dividing a pool of items into groups (i.e., anti-clusters) to create high within-group heterogeneity and high inter-group similarity, anti-clustering is accomplished by maximizing rather than minimizing the clustering objective function.

2 Application

In application, the clustering analysis methods briefly introduced above were applied to some data sets whose cluster numbers and distribution of observations are known in the literature, and the results were evaluated. In fact, the main theme of the study is the answer to the question of what will be the results if any clustering analysis method introduced in the literature is applied for any data set. Applications have been made in R (version 4.2.2) package program with author own codes or through suitable packages for methods (Kaiser et al. (2023); Papenberg et al. (2023)).

The first data set used in the study is Iris data. The data set consists of 50 samples from each of three species of Iris. Four features were measured from each sample: the length and the width of the sepals and petals, in centimeters. The scatter plot for the data divided into three clusters is as given in Figure 1(a). As can be seen, the spread of the data is quite good. The second data set is Wisconsin Breast Cancer data set which consists of a total of 683 samples given in Figure 1(b). There are a total of 9 variables and 2 clusters in this data. Synthetic data has 400 samples, 2 variables and 3 clusters (Figure 1c). Motorcycle data has 133 samples, 2 variables and 4 clusters (Figure 1d).

Fig. 1: Scatter plots for first four different data sets



(a)

(b)

(c)

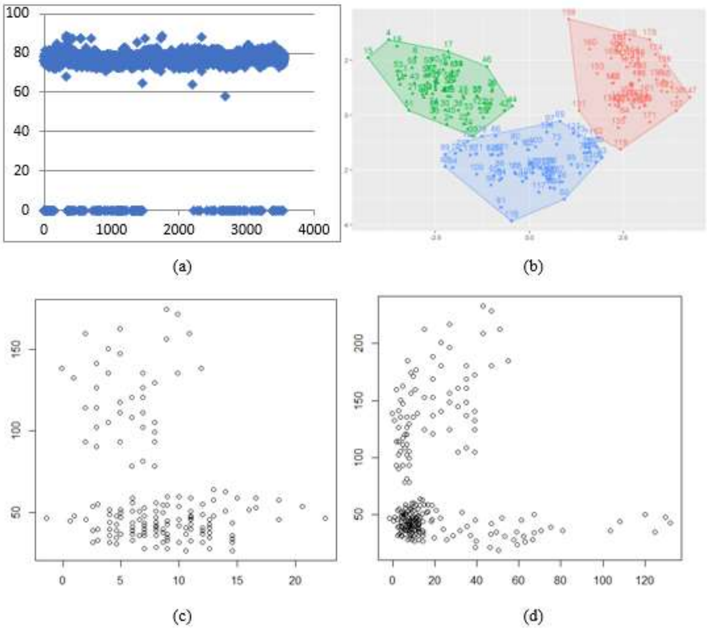
(d)

Source: a-URL1, b-Niknam et al. (2008), c-d-Balasko et al. (2005)

Fifth data is yellow wagtail bird data consist of 4506 samples and 4 variables with 7 clusters given in Figure 2a. Sixth data is Wine data with 178 samples, 13 variables and 3 clusters (Figure 2b). Last 2 dataset are simulation data sets. First one has 155 samples, 2 variables and 3 clusters and second one has 270 samples, 3 variables and 4 clusters.

Some of the previously introduced clustering methods were applied to these eight data sets, and the assignment percentages are given in Table 1. Looking at the results in Table 1, we cannot say that any method always gives more successful results, as expected. The accuracy of the clustering algorithm on the dataset may vary depending on various factors, such as the choice of features, preprocessing steps, and the specific implementation and parameters of the clustering algorithm.

Fig. 2: Scatter plots for second four different data sets



Source: a-Gürsoy (2007), b-URL1, c-d by author

Tab. 1: Accurate clustering percentages according to different methods

		Iris	WBC	Synth.	M.Cycle	Wagtail	Wine	Sim.1	Sim.2
Crisp	Between Groups	75	83	98	65	42	88	68	77
	Within Groups	51	86	98	76	41	89	71	74
	Nearest Neighbor	68	77	99	73	46	89	69	73
	Furthest Neighbor	52	79	99	80	36	85	73	70
	Ward	89	90	99	82	44	96	80	84
	k-means	92	90	100	84	46	93	86	85
Fuzzy	Fuzzy C-Means	81	95	100	88	50	93	87	86
	Gustafsson-Kessel	79	87	98	83	51	92	88	82
	Gath-Geva	71	88	99	83	46	89	81	80
Gray	Gray	77	76	99	89	53	88	88	85
BiClustering	Bimax	79	71	99	81	54	89	69	74
	CC	76	80	99	80	50	87	64	72
	Questmotif	81	83	99	79	48	85	69	72
AntiClustering	AntiClustering	80	84	100	79	40	84	68	70

Note that success rates are low in datasets where the data distributions are relatively nested. This is perhaps the most critical and challenging part of cluster analysis.

Conclusion

Cluster analysis techniques are concerned with exploring datasets to assess whether they can be summarized meaningfully for a relatively small number of objects, sets of objects, or individuals that are like each other and differ in some respects from individuals in other clusters. Different measures of similarity or dissimilarity calculated from the same set of individuals can often lead to different solutions when used as the basis for cluster analysis. As a result, it would be extremely helpful to know which specific measures are in some sense “optimal.” Although this subject has been frequently studied in the literature, there is no definite answer to this question. The effect of factors such as the researcher's prior knowledge about the subject, the type of variables, or their degree of importance on the results is important. Sokal and Sneath (1973) suggested selecting the simplest coefficient applicable to a dataset. Perhaps, it is best to interpret things in the simplest way possible.

In practice, hierarchical methods form the backbone of cluster analysis. They are widely available in almost every statistical software package and provide fast results even in large data structures. The researcher can change the proximity measure, the clustering method suitable for the data, and the number of clusters instantly and evaluate different results. The main problem in practice is that a particular clustering method cannot be recommended, because methods with positive mathematical properties often do not seem to produce empirically interpretable results. Standard clustering methods have been developed in many ways to cover realistic situations, such as data constraints and overlapping clusters. Many of these developments allow for a more comprehensive interpretation of sets in terms of both objects and variables or include features such as fuzzy memberships.

Cluster analysis methods have emerged as important tools for the investigation of multivariate data. Clustering can help researchers discover features of any existing structure or model by organizing such data into subgroups or clusters. However, the practical application of these methods requires great care if overinterpretation of the obtained solutions is avoided. Applying a particular cluster analysis method to a dataset and accepting the solution as it appears are often insufficient. If a precise and accurate result about the research is desired, different methods (with different distance measures and cluster validity indices) should be tested, and the results obtained should be evaluated as a whole.

The best clustering method depends on the specific data set and the desired outcome. Here is a brief overview of the different clustering methods and their strengths and weaknesses:

K-means clustering is a simple and efficient clustering method that is often used as a baseline for comparison with other clustering methods. It is not as flexible as some other clustering methods, but it is relatively easy to understand and implement.

Fuzzy clustering allows data points to belong to multiple clusters, which can be useful for data sets that are not well-clustered. However, fuzzy clustering can be more computationally expensive than other clustering methods.

Biclustering identifies clusters of data points that are similar to each other in two dimensions. This can be useful for finding patterns in data sets that are not well-clustered in one dimension.

Anticlustering identifies data points that are dissimilar to the rest of the data set. This can be useful for finding outliers or anomalies in data sets.

Gray clustering is a type of clustering that allows data points to have different degrees of membership in different clusters. This can be useful for data sets that have a continuous range of values.

In general, k-means clustering is a good choice for clustering data sets that are well-clustered in one dimension. Fuzzy clustering is a good choice for clustering data sets that are not well-clustered in one dimension. Biclustering is a good choice for clustering data sets that have two dimensions. Anticlustering is a good choice for finding outliers or anomalies in data sets. Gray clustering is a good choice for data sets that have a continuous range of values. The best way to choose a clustering method is to experiment with different methods and see which one produces the best results for your specific data set.

References

Balasko, B., Abonyi, J., Feil, B. (2005). *Fuzzy clustering and data analysis toolbox.* Department of Process Engineering, University of Veszprem, Veszprem.

Bellman, R. E., Kalaba, R., Zadeh, L.A. (1966). Abstraction and pattern classification. *J. Math. Anal. Appl.*, 1-7.

Brusco, M. J., Cradit, J. D., Steinley, D. (2019). Combining diversity and dispersion criteria for anticlustering: A bicriterion approach. *British J. of Math. and Statistical Psychology*.

Cheng, Y., Church, G. M. (2000). Biclustering of expression data. *Proc Int Conf Intell Syst Mol Biol.,* 8: 93-103.

Deng, J.L. (1982). Control Problems of Grey Systems. *Systems & Control Letters*, 1 (5), 288-294.

Everitt, B. S., Landau, S., Leese, M., Stahl, D. (2011). Cluster analysis. John Wiley & Sons.

Gürsoy, A. (2007). *Motacilla flava'nın Kızılırmak Deltası'ndan göç eden populasyonlarının morfometrik özelliklerinin ve göç oriyantasyonlarının incelenmesi* (In Turkish), Phd Thesis, 19 Mayis University, Samsun, Turkiye.

Hair Jr., J.F., Black, W.C., Babin, B.J., Anderson, R.E. (2009) *Multivariate Data Analysis.* 7th Edition, Prentice Hall, Upper Saddle River, 761.

Kaiser, S., Santamaria, R., Khamiakova, T., Sill, M., Theron, R., Quintales, L., Leisch, F., De Troyer, E., Leon, S. (2023). *Package 'biclust'*. The Comprehensive R Archive Network.

Ke, L., Xiaoliub, S., Zhongfua, T., Wenyan, G. (2012). Grey Clustering Analysis Method for Overseas Energy Project Investment Risk Decision. *Systems Engineering Procedia,* 3, 55-62

Liu, S., Forrest, J.Y.L. (2006). *Grey Information: Theory and Practical Applications.* Springer: London.

Niknam, T., Bahmani Firouzi, B., Nayeripour, M. (2008). An efficient hybrid evolutionary algorithm for cluster analysis. *World Applied Sciences Journal*, 4, 2, 300–307.

Papenberg, M., Michalke, M., Klau, G.W., Nagel, J.V., Breuer, M., Schaper, M.L., Diekhoff, M. (2023). *Package 'anticlust'*. The Comprehensive R Archive Network.

Spath, H. (1986). Anticlustering: Maximizing the variance criterion. *Control and Cybernetics,* 15, 213–218.

URL1: https://www.kaggle.com/datasets

Contact

Necati Alp Erilli
Sivas Cumhuriyet University, Department of Econometrics
Sivas Cumhuriyet University, IIBF, Ekonometri Bolumu, 58200, Sivas/Turkey
e-Mail: aerilli@cumhuriyet.edu.tr