International Journal of Advanced Research in
IT and Engineering

ISSN: 2278-6244

CLUSTERING AND ITS APPROACHES

Shabnam Azari*

Tooraj Samimi Behbahan*

Abstract: *Clustering and Classification are among methods that widely used in Data Analysis. Extracting patterns in the data by individuals grouping and variables, the main goal of these methods is subject. Clustering and classification of a wide variety of methods that can be used in many sciences.*

*Department of Computer Engineering, Behbahan Branch, Islamic Azad University, Behbahan, Iran

1- INTRODUCTION

Today, as the volume of database increases, care about clustering methods is increasing day to day. Clustering is useful at exploration of pattern analysis in the fields of identifying pattern, grouping, decision making, machine learning, data mining, documents retrieval, and segmentation, classification of patterns, biology, Medicine, and information retrieval.

A clustering algorithm classifies the existing samples within a multi-dimensional space into identified groups from features having certain values. These identified groups are called clusters. There is no information that in which class or cluster the existing samples lie. Therefore, clustering is classified as an unsupervised learning technique.

2- DEFINITION OF CLUSTERING

A set of $X = \{x_1, x_2, ..., x_n\}$ consists of n objects, each equal to one vector of length s of features $(x_i \in R^s)$. These objects should be clustered within k group called $C = \{C_1, C_2, ..., C_k\}$ that are non-overlapping, so that:

$$C_1 \bigcup C_2 \bigcup \cdots \bigcup C_K = X \ , \quad C_i \neq \phi \ , \quad and \quad C_i \bigcap C_j = \phi \quad for \ \ i \neq j$$

2-1- Initialization

In some clustering algorithms such as k-means, initialization must be done properly in order to get a good performance. Some initialization methods have been provided for troubleshooting, such as memetic algorithms using a combination of accidental search and k-means for clustering. By several initial values and by using evaluation functions, an appropriate clustering could be achieved.

2-2- Classification stage

There are different methods for clustering. Clustering methods have been divided to seven categories including hierarchical, theory graph, partitioning, estimation of cones probability density, fuzzy, neural network, complementary algorithms, punishment and reward-based methods, and compound algorithms. There is no clustering algorithm could function well for all applications, and clustering algorithm is used as for any given matter. Clustering algorithm often includes an implicit assumption about cluster form or about organizing several clusters based on similarity, and also on used classification criterion.

2-3- Evaluation of clusters

Analysis of clustering validation is studied by clustering output estimation. In this analysis, a specified criterion is used for optimization. This criterion examined subjectively, however, validation determination is objective. This criterion used for determining meaningfulness of the output.

2-4- Feedback

In some applications, clustering algorithm cannot get an appropriate answer with only one performance, and there is a need for applying changes and reimplementation of clustering algorithm in new conditions. Feedback can be automatic and can use human comments. In order to improve the efficiency of clustering algorithm, methods for validation determination usually used to achieve an appropriate feature, appropriate parameters, an appropriate similarity criterion, and/or an appropriate initialization.

2-5- Interpretation of results and knowledge discovery

One of the most important applications for clustering is for realizing data structure. Clustering makes an appropriate knowledge to be produced regarding the subject, in case there is no initial knowledge about it. One of the most important applications for the obtained knowledge about clustering is to use it for solving classification problems.

The act of clustering is basically subjective. The same data set should be partitioned when clustering. There are different forms of partitioning for various purposes. For example, if whale, elephant, and fish considered, whale and elephant will be categorized within mammals cluster. However, if user classifies the animals based on their location, fish and whale will be included within the same class. So, clustering criterion and knowledge scope are combined.

3- REASONS FOR USING CLUSTERING ALGORITHMS

Unsupervised methods are attractive because of the following reasons [1]:

- ✓ Unsupervised methods could provide insight into the structure or the nature of data. Discovery of the structure could be very valuable.
- ✓ Data without labeling needs an expert person. This process is very costly, time-consuming, and error-prone.
- ✓ In some databases, there is detailed information, and employing unsupervised algorithms has little cost.

- ✓ In order to determine features and an appropriate similarity criterion, unsupervised algorithms may be used automatically.
- ✓ As the volume of data set increases, supervised algorithms will be used less because of long time training, and need for many training patterns.

4- DISPLAY OF PATTERNS

Display of pattern is a process of selecting the number of classes, the number of the existing patterns, number and criterion of the existing patterns for clustering algorithm. One technique for feature selection is detection process for a subset of features which is more effective to be used in clustering. Another technique for feature extraction is the use of one or more conversion on input features to produce a new prominent feature. In figure 1.A it is observed the data are not separated and it should be converted. However, in figure 1.B data are separated. In figure 2 it is considered that data are on a curve. If Cartesian coordinate selected, it could be divided into two or more clusters that are not compact. However, if polar coordinate selected to be displayed, data could be easily clustered. Data conversion within feature space could cause data separation and better clustering [2].

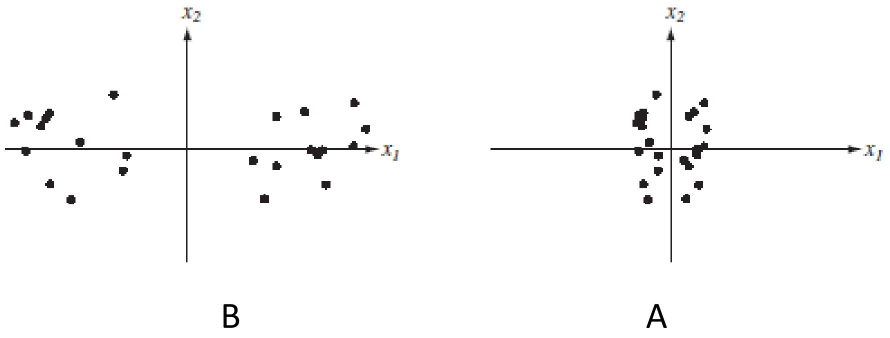

Figure 1: Display the way of data distribution

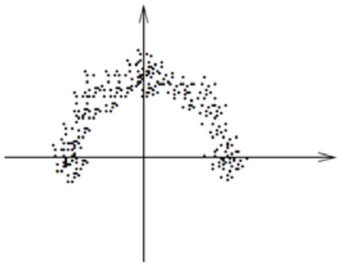

Figure 2: Data in the form of a curve, and the same distribution

5- CLUSTERING ALGORITHMS

Although while clustering, all solutions are considered, c^n/c different partitions for n patterns within the cluster could be considered. For example, in order to partition 100 patterns in 5 clusters, more than 10^{67} partitions could be considered [3]. So considering all possible clusters, and achieving the best state is almost impossible. In [4], clustering analysis has been classified to the following groups:

- ✓ Partitioning algorithms such as K-means
- ✓ Hierarchical methods such as Linkage, and minimum variance method
- ✓ Distribution-based algorithms such as DBSCAN
- ✓ Grid-based algorithms such as STING
- ✓ Other techniques like Fuzzy clustering and neural network

The existing clustering algorithms could be divided to 9 groups based on the used method and technique. Figure 3 displays these algorithms.

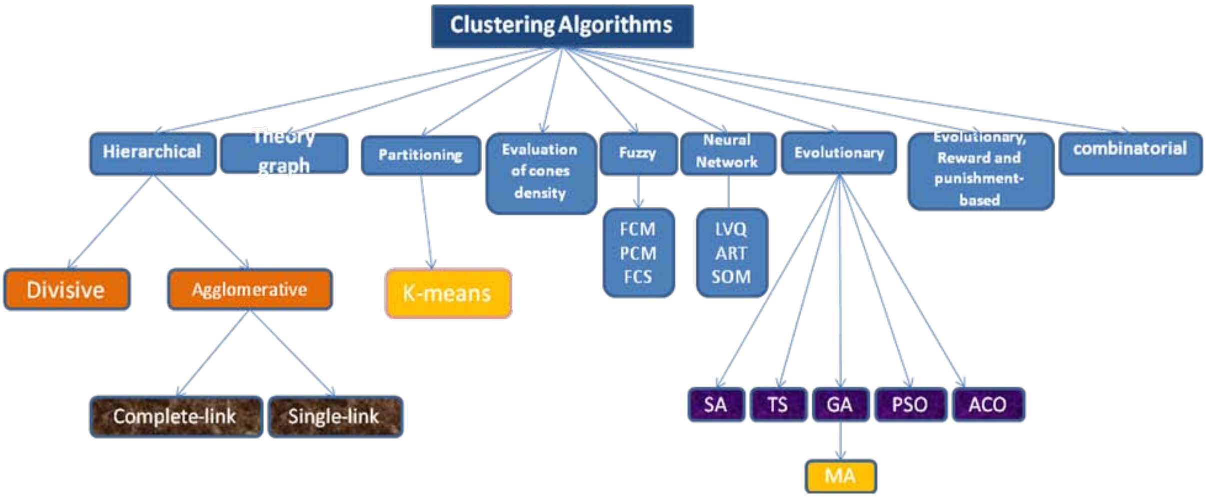

Figure 3: Clustering algorithms

6- CONCLUSION

In this chapter, different clustering approaches, and their advantages and disadvantages were considered. In order to implement clustering operations, it is necessary to do feature selection on the training set, whose method types discussed. In many clustering algorithms, the critical problem is the way of determining the number of clusters often used by user familiar with domain. Some methods provided to determine the number of clusters. Various similarity criteria considered. Also, a comparison made between different clustering algorithms. In this thesis, clustering has been used as an introduction to word documents margin operations for determining thematic scope of the document.

REFERENCES

[1] Webb A.R., "Statistical Pattern Recognition, Second Edition", 2002 John Wiley & Sons, Ltd.

[2] Jain A.K., Murty M.N., Flynn P.J., "Data clustering, a review, ACM Comput. Survay.", 31 (3) (2000) 264–323

[3] Duda R., Hart P., Stork D., "Pattern Classification", 2nd Edition, 2001.

[4] Cheng C.H., Wei L.Y., "An Evolutionary Computation Based on GA Optimal Clustering", Proceedings of the Sixth International Conference on Machine Learning and Cybernetics, Hong Kong, 19-22 August 2007.