CLUSTERING DATA FOR KNOWLEDGE

Assistant Professor Ph.D. Student, Raluca-Mariana Ştefan

Academy of Economic Studies, Bucharest, Romania
rstefan2012@yahoo.com

Abstract*: Based on their natural senses and their knowledge people could always easily group data or objects into different categories. But the past few decades have brought overwhelming amounts of digital data as well as the necessity of increasing the performances of classifying multidimensional objects into relevant information and knowledge. The technique of determining the natural trend of data to form groups of similar objects is called cluster analysis. This type of analysis defines unsupervised data classification. It is used when data or objects have no labelled samples so that in order to establish to what class they belong, division of data set based only on their similarities and dissimilarities is needed. Cluster analysis is fundamentally different from statistical procedures because is not based on any a priori specific hypothesis, but some researchers do not agree. This paper describes principles, methods, algorithms and optimal criteria used by clustering to categorize complex data into real and accurate knowledge.*

Keywords: cluster analysis, data clusters, knowledge, optimal criterion

1. Introduction

Cluster analysis tries to identify groups of similar objects, helps to discover distribution of patterns and interesting correlations in large data sets and it has been subject for research since it arises in many application domains like engineering, business and social sciences [4]. Huge transactional and experimental amounts of digital data sets made clustering algorithms necessary and proved to be efficient in various fields. Fundamental clustering concepts, clustering principles, frequently used clustering algorithms and clustering validity measures and approaches constitute the main preoccupation in this paper.

Clustering and classification are two different terms used to express two different methods to organize a set of data and they must not be confused. Clustering process regards grouping data according to their similarities or / and dissimilarities not knowing the classes they belong to. That is called unsupervised learning. Classification procedure assigns objects from a set to predefined classes using a training set of examples. Training phase makes this procedure to be known under the name of supervised learning. The choice of one or the other method for a data set is the researcher privilege according to expected results.

2. Cluster Analysis

The paradigm of knowledge-based clustering already emerged and is concerned with reconciling two important driving forces of clustering activities: gaining data and domain knowledge and building a coherent platform of navigation in highly dimensional and lately, very heterogeneous data spaces [6]. The main goal of clustering is descriptive being set to discover new data categorizations based on their inner structure. This idea is applicable in many domains and clustering is found under many different names: partition, numerical taxonomy, typology etc. [4]. Unsupervised learning uses an optimal partition of feature space in terms of a certain optimal criterion, without using a priori information.

The advantage of these types of methods is given that is automatic and not requiring any user intervention. Thus, it can be used for making data clusters from data sets for which there are no information regarding their content. On the other hand, because it is an automatic process, the relevance of classes tends to be lower than for supervised learning, because is dependent on used method and discriminatory power of features of the space that was chosen.

2.1. Clustering Principles

Some of the most important principles that guide cluster analysis are: grouping data and reducing dimensionality.

Clustering principles form a prominent class of methods designed to discover and extract structures that lie in data sets [5].

Clustering algorithms impose structure to the given data and the resulting clusters may not always be the correct ones. Clustering algorithms also have a good performance when applied to experimental data sets and they have solution stability.

2.2. Clustering Methods and Algorithms

Clustering methods objective for a data set is discovering existing significant groups. These methods usually look for clusters that contain objects as closer to each other as possible and this indicates a high level of objects similarity.

Algorithms that can be used for carrying out the construction of clusters, know a great variety including heuristic algorithms, optimization algorithms and fuzzy algorithms. The differences between how to build clusters by an algorithm or another are determined mainly by the nature of the method used to evaluate distances between clusters. The type of cluster analysis results from the nature of the algorithm used to build clusters.

There are many types of clustering algorithms and they are clustered based on the following actions:

- The type of data input;

- The clustering criterion defining the similarity between data points;

- The theory and fundamental concepts on which clustering analysis techniques are based on [4].

Based on their nature, type of operation and type of solutions they provide, cluster analysis methods can be divided into two broad categories: hierarchical methods and iterative methods or partition type.

Hierarchical methods or algorithms have as a purpose producing many cluster solutions and these kinds of solutions are named cluster hierarchies. Their main characteristic is that the clusters number is not known at the beginning of the procedure. There are two categories of hierarchical clustering algorithms, namely aggregation and divisive algorithms.

By applying a hierarchical clustering algorithm the results contain more objects clustering variants and each variant contains cluster structures that have a variable clusters number. Cluster structures that result by running such algorithms are called multilevel cluster structures.

Iterative algorithms or methods aim to produce a cluster structure that is formed by a single cluster solution. This type of solution is called unilevel cluster structure because it contains a single cluster that contains a fixed number of objects clusters. Thus, partitioning (or objective function-based) clustering algorithms provide unique solutions after they are applied.

Main characteristic of partitioning clustering algorithms is that the number of clusters is fixed by user.

Other types of clustering algorithms according to the method applied to define clusters are: density-based algorithms (DBSCAN), grid-based algorithms, model-based methods (COBWEB) and fuzzy clustering algorithms.

2.3. Clustering Optimal Criteria

Clustering faces the problem of deciding which number is an optimal number of clusters so that it fits considered data set. A clustering optimal criterion is needed to evaluate how good the results obtained after a clustering algorithm is applied. Evaluating the results of a clustering algorithm is known under the term cluster validity [4]. Three approaches are used for cluster validity: external

quality criteria, internal quality criteria and relative quality criteria.

Table 1 Evaluation criteria measures

Validity indices	*Description*
I.INTERNAL QUALITY CRITERIA	Based on statistical methods
Sum of squared error (SSE)	SSE is appropriate for clusters well separated
Other minimum variance criteria	In addition to SSE, they are also appropriate for well separated clusters
Scatter criteria	Derived from scatter matrices, used to compute within, between and total cluster scatter.
Condorcet Criterion	Maximal number of clusters is not predetermined, it can be determined
C-Criterion	Represent an extension of Condorcet criterion
Category utility metric	For small number of nominal features that have a small number of elements
Edge cut metrics	Represent problem as a an edge cut minimization problem
Cophenetic correlation criteria	Validation of the clusters hierarchy
Validating a single clustering scheme	Validation of agreement degree between a clustering scheme and proximity matrix
II.EXTERNAL QUALITY CRITERIA	Based on statistical methods
Mutual information based measure	An external measure for mutual clustering validity
Precision-recall measure	Verifies correct clustered objects
Rand index	Compares an induced cluster and a given cluster having a value between 0 and 1
Monte Carlo techniques	Compute probability density function for validity indices
Comparison of cluster structure with partition P	Not useful for hierarchical clustering
Comparison of proximity matrix with partition P	Compares two matrices similarity
III.RELATIVE QUALITY CRITERIA	Involves no statistical tests; based on choosing the best algorithm from different results according to an a priori criterion. Appropriate for fuzzy clustering and crisp clustering

Data clustering procedure consists of the following operations that are performed over a data set:

Step 1. Data features are selected;
Step 2. Clustering algorithm is chosen;
Step 3. After clustering process the results are presented to be evaluated;
Step 4. Using an optimal criterion results are validated;
Step 5. Final obtained clusters are presented;
Step 6. Results are interpreted;
Step 7. Knowledge is extracted.

External quality criteria evaluate clustering based on specific intuition methods. Internal quality criteria are based on metrics that are based on initial data set and clustered data. There is a disadvantage and it is related to computational complexity. Relative quality criteria are not based on statistical methods.

4. Conclusions

In order to get knowledge clustering is a very good technique that can be used due to its principles, methods and optimal criteria which are applied.

Cluster analysis determines the right number of data clusters from a data set without specific rules and examples and that characteristic is an advantage when we deal with complex heterogeneous data.

Having enormous number of data sets from many fields clustering techniques are appropriate to be applied in order to extract real and accurate knowledge.

Cluster analysis algorithms can also be used for prediction of values as well as they provide natural groupings of data. Even if this is an obvious application, generally, clustering hasn’t been used too much for such purposes and we propose further research in this direction.

References

[1] Estivill-Castro V., Why So Many Clustering Algorithms: A Position Paper, *ACM SIGKDD Explorations Newsletter*, Vol. 4, Iss. 1, (June 2002), pp. 65-75, 2002.

[2] Jain A., Murty M. and Flynn P., Data Clustering: A Review, *ACM Computing Surveys* (CSUR), Vol. 21, Iss. 3 (Sep. 1999), pp. 264-323, 1999.

[3] Zhong, S. and Ghosh J., A Unified Framework for Model-based Clustering, *The Journal of Machine Learning Research*, Vol. 4, (Dec. 2003), pp. 1001-1037, 2003.

[4] Halkidi M., Batistakis Y. and Vazirgiannis M., On Clustering Validation Techniques, *Journal of Intelligent Information Systems*, 17:2/3, 107–145, 2001, Kluwer Academic Publishers, 2001.

[5] Jain A.K. and Dubes R.C., *Algorithms for Clustering Data*, Prentice Hall, Englewood Cliffs, 1988.

[6] Pedrycz W., *Knowledge-Based Clustering – From Data to Information Granules*, Wiley, New York, 316 pp., 2005.