A STUDY OF CLUSTERING METHODS BASED ON CLASSIFICATION

Pankaj Dhumane
Sardar Patel Mahavidyalaya, Chandrapur
Corresponding author Email : pdhumane@rediffmail.com

Abstract:

I studied various classification methods including exclusive and inclusive, extrinsic and intrinsic, partitional and hierarchical classifications. Before that I tried to give the different definitions of clusters given by various researchers. Clustering is the process of putting similar data into groups.

Keywords:

Classification, cluster, groups,

Introduction

A cluster is comprised of a number of similar objects collected or grouped together. Everitt (1974) documents some of the definitions of a clusters are : "A cluster is a set of entities which are alike, and entities from different clusters are not alike." "A cluster is an aggregation of points in the test space such that the distance between any two points in the cluster is less than the distance between any point in the cluster and any point not in it." "Clusters may be described as connected regions of a multi-dimensional space containing a relatively high density of points, separated from other such regions by a region containing a relatively low density of points." The objects to be clustered are represented as points in the measurement space. It is easy to give a functional definition of a cluster, it is very difficult to give an operational definition of a cluster. This is due to the fact that objects can be grouped into clusters with different purposes in mind. Cluster analysis is one component of exploratory data analysis, which means sifting through data to make sense out of measurements by whatever means are available. The information gained about a set of data from a cluster analysis should prod one's creativity, suggest new experiments, and provide fresh insight into the subject matter. The modern digital computer makes all this possible. Cluster analysis is a tool for exploring

data and must be supplemented by techniques for visualizing data. The most direct visualization is a two-dimensional plot showing the objects to be clustered as points. Multivariate data cannot always be faithfully reproduced in two dimensions but when valid, such a representation is helpful in verifying the results of a clustering algorithm. Proximity Indices Clustering methods require that an index of proximity be established between pairs of patterns. This index can be computed from a pattern matrix, or can be formed from raw data. The data in some psychometric applications are collected as proximities. A proximity matrix [d(i,j)] accumulates the pairwise indices of proximity in a matrix in which each row and column represents a pattern. A proximity index is either a similarity or a dissimilarity. The more the ith and jth objects resemble one another, the larger a similarity index and the smaller the dissimilarity index. Anderberg (1973) provides a thorough review of measures of association and their interrelationships. A proximity index between the ith and kth patterns is denoted d(i,k) and must satisfy the following three properties: 1) a) For dissimilarity : d(i,j) = 0, for all i b) For similarity : d(i,j) ≥ max d(i,k), for all i 2) d(i,k) = d(k,i), for all i and k 3) d(i,k) ≥ 0, for all i and k. A clustering is a type of classification imposed on a finite set of objects. The relationship between objects is represented in a proximity matrix in which rows and columns correspond to objects. If the objects are characterized as patterns, or points in a d-dimensional metric space, the proximities can be distance between pairs of points, such as Euclidean distance. Unless a meaningful measure of distance, or proximity, between pairs of objects has been established, no meaningful cluster analysis is possible. The proximity matrix is the one and only input to a clustering algorithm.

Material and Method

Basic Clustering Techniques Exclusive versus nonexclusive - An exclusive classification is a partition of the set of objects. Each object belongs to exactly one subset, or cluster. Nonexclusive or overlapping, classification can assign

an object to several classes. Intrinsic versus extrinsic - An intrinsic classification uses only the proximity matrix to perform the classification. Intrinsic classification is called "unsupervised learning" in pattern recognition because no category labels denoting an a priori partition of the objects are used. Extrinsic classification uses category labels on the objects as well as the proximity matrix. The problem is then to establish a discriminant surface that separates the objects according to category. Hierarchical versus partitional - Exclusive, intrinsic classifications are subdivided into hierarchical and partitional classifications by the type of structure imposed on the data. A hierarchical classification is nested sequence of partitions, whereas a partitional classification is a single partition. Partitional and hierarchical clustering are the two major clustering techniques and there are many more techniques like Density-Based Clustering , Grid-Based Clustering , Model-Based Clustering, Categorial Data Clustering.

Result and Discussion

A huge collection of clustering algorithms is available now a days. New clustering programs continue to appear in the scientific literature. However, most of these algorithms are based on the following two popular clustering techniques: iterative square-error partitional clustering and agglomerative hierarchical clustering. Hierarchical techniques organize the data in a nested sequence of groups. Square-error partitional algorithms attempt to obtain that partition which minimizes within-cluster or maximizes between-cluster scatter.

Conclusion

All the above classifications are done to derive the most important clustering techniques that are partitional and hierarchical. These two techniques are most popular but for few problems there is need of different clustering techniques.

Reference

AGRAWALA, A.K. , MOHR, J.M. , and BRYANT, R.M. (1976), "An approach to the workload characterization problem" Computer 9 (June), 18-32.

AHUJA N. (1982), "Dot pattern processing using voronoi neighbourhoods." IEEE Transactions on pattern Analysis and Machine Intelligence PAMI 4, 336-343.

ANDERBERG, M.R. (1973), Cluster Analysis for Applications, Academic Press, Inc. , New York.

BACKER, E. (1978) Cluster Analysis by Optimal Decomposition of Induced Fuzzy Sets. Delft University Press, Delft. The Netherlands.

BACKER, E., and JAIN, A.K. (1981), "A clustering performance measure based on fuzzy set decomposition" IEEE Transactions on Pattern Analysis and Machine Intelligence PAMI 3, 66-75.

BAILEY, T.A. and DUBES, R.C. (1982), "Cluster validity profiles." Pattern Recognition 15, 61-83

BEZDEK, J.C. (1974), "Cluster validity with fuzzy sets", Journal of Cybernetics 3, 58-73.

BINDER, D.A. (1978) "Bayesian Cluster Analysis", Biometrika 65, 31-38.

COLEMAN, G.B. and ANDREWS, H.C. (1979), "Image segmentation by clustering", Proceedings of the IEEE 67, 773-785.

COOLEY, W.W. and LOHNES, P.R. (1971), Multivariate Data Analysis. John Wiley & Sons, Inc., New York.

DUBES,R. (1968). Theory of Applied Probability, Prentice-Hall, Inc. Englewood

Cliffs, N.J. DUBES, R. and JAIN A.K. (1976) "Clustering techniques : the user's dilemma", Pattern Recognition 8, 247-260.

DUBES, R. and JAIN A.K. (1979) "Validity studies in clustering methodologies", Pattern Recognition 11, 235-254.

DUBES, R. and JAIN A.K. (1980) "Clustering methodologies in exploratory data analysis", In Advances in Computers, Vol.19 , Academic Press, Inc., New York, pp. 113-215.

DUBES, R., and ZENG, G. (1987), "A test for spatial homogeneity in cluster analysis." Journal of Classification 4, 33-56.

EDWARDS, A.W.F. and CAVALLI-SFORZA, L.L. (1965), "A method for cluster analysis", Biometrics 21, 362-375.

EVERITT, B.S. (1974) Cluster Analysis, John Wiley & Sons, Inc. New York.

GOWER, J.C. (1967). "A comparison of some methods of cluster analysis", Biometrics 23, 623-637.

JAIN, A.K. (1987) "Advances in Statistical Pattern Recognition", Pattern Recognition Theory and Applications, Springer Verlag, NewYork, pp. 1-19.

JAIN, A.K. (1985) "Experiments in texture analysis using spatial filtering", Proceedings IEEE Workshop on Languages for Automation, Palme De Mallorca, June pp. 66-70.

JAIN, A.K. (1986) "Cluster analysis" In handbook of Pattern Recognition and Image Processing, Academic Press, Inc. New York pp. 35-57.

NEYMAN, J. and SCOTT, E.L. (1972) "Processes of clustering and applications", In Stochastic Point Processes : Statistical Analysis, Theory and Applications, John Wiley & Sons Inc. New York, pp. 646-681.

RAO, M.R. (1971) "Cluster analysis and mathematical programming", Journal of the American Statistical Association 66, 622-626.