IJITKM Volume 7 • Number 1 • December 2013 pp. 137-139 (ISSN 0973-4414)

Study on Two Step Clustering

Kanchan Chaudhary1 Dr. Anuj Sharma2
1. M.Tech. Scholar,Om Institute of Technology and Management, Hisar
2. Associate Professor, Om Institute of Technology and Management, Hisar
er.kanchan2787@gmail.com, anuj.k.er@gmail.com

ABSTRACT: A cluster analysis approach plays a Vital role in software alliance. Cluster analysis is the scheme for sorting out data into groups in a situation where no prior information about a grouping structure is available. In this paper we are studing about choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-means) andTwo Step Clustring by dividing the observations into homogeneous and distinct groups. so that observations within each group are similar to one another with respect to variables or attributes of interest and the groups themselves stand apart from one another.
Keywords: Two Step Clustring, K-medians, K-means++

I. INTRODUCTION

A cluster analysis approach plays a big role in software alliance. Cluster analysis is the scheme for sorting out data into clusters or groups in a situation where no prior information about a grouping structure is available. It divides data into groups (clusters) that are meaningful, useful or both. The clustering approach works like key of decision making and an effective inspiration method in generating ideas and obtaining solutions. The goal of a cluster analysis is to minimize and categorize the records for effective decision making.

Identifying groups of individuals or objects that are similar to each other but different from individuals in other groups can be intellectually satisfying, profitable, or sometimes both. Using your customer base, you may be able to form clusters of customers who have similar buying habits or demographics. You can take advantage of these similarities to target offers to subgroups that are most likely to be receptive to them. Based on scores on psychological inventories, you can cluster patients into subgroups that have similar response patterns. This may help you in targeting appropriate treatment and studying typologies of diseases. By analyzing the mineral contents of excavated materials, you can study their origins and spread. Although both cluster analysis and discriminate analysis classify objects (or cases) into categories, discriminate analysis requires you to know group membership for the cases used to derive the classification rule. The goal of cluster analysis is to identify the actual groups. For example, if you are interested in distinguishing between several disease groups using discriminate analysis, cases with known diagnoses must be available. Based on these cases, you derive a rule for classifying undiagnosed patients. In cluster analysis, you don't know who or what belongs in which group. You often don't even know the number of groups. The term cluster analysis does not identify a particular statistical method or model, as do discriminate analysis, factor analysis, and regression. You often don't have to make any assumptions about the underlying distribution of the data. Using cluster analysis, you can also form groups of related variables, similar to what you do in factor analysis. There are numerous ways you can sort cases into groups. Cluster Analysis depends on, among other things, the size of the data file. Methods commonly used for small data sets are impractical for data files with thousands of cases. SPSS has three different procedures that can be used to cluster data: hierarchical cluster analysis, k-means cluster, and two-step cluster. If you have a large data file (even 1,000 cases is large for clustering) or a mixture of continuous and categorical variables, you should use the SPSS two-step procedure. If you have a small data set and want to easily examine solutions with increasing numbers of clusters, you may want to use hierarchical clustering. If you know how many clusters you want and you have a moderately sized data set, you can use k-means clustering. You'll cluster three different sets of data using the three SPSS procedures. You'll use a hierarchical algorithm to cluster figure-skating judges in the 2002 Olympic Games. You'll use k-means clustering to study the metal composition of Roman pottery. Finally, you'll cluster the participants in the 2002 General Social Survey, using a two-stage clustering algorithm. You'll find homogenous clusters based on education, age, income, gender, and region of the country. You'll see how Internet use and television viewing varies across the clusters. In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, k-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

The optimization problem itself is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximative method is Lloyd's algorithm, often actually referred to as "k-means algorithm". It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-means).

Most k-means-type algorithms require the number of clusters - k - to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders in between of clusters (which is not surprising, as the algorithm optimized cluster centers, not cluster borders).

K-means has a number of interesting theoretical properties. On the one hand, it partitions the data space into a structure known as a Voronoi diagram. On the other hand, it is conceptually close to nearest neighbor classification, and as such is popular in machine learning.

II. CLUSTERING APPROACH

Cluster analysis is a tentative data analysis tool for solving classification problems. Its objective is to sort cases (people, things, events, etc) into groups, or clusters, so that the degree of friendship is strong between members of the same cluster and weak between members of different clusters [9]. Each cluster thus describes, in terms of the data composed, the class to which its members fit in; and this picture may be abstracted through use from the particular to the general class or type. Cluster approach is thus a tool of analysis. The results of cluster analysis may contribute to the definition of a formal classification scheme, such as taxonomy for related animals, insects or plants; or suggest statistical models with which to describe populations; or indicate rules for assigning new cases to classes for recognition and problem-solving purposes. Cluster analysis is a course of identification and categorization of subsets of objects. Partitioning or clustering techniques are used in many areas for a wide spectrum of problems. Among the areas in which cluster analysis is used are graph theory, business area analysis, information architecture, information retrieval, resource allocation, image processing, software testing, galaxy studies, chip design, pattern recognition, economics, statistics and biology.

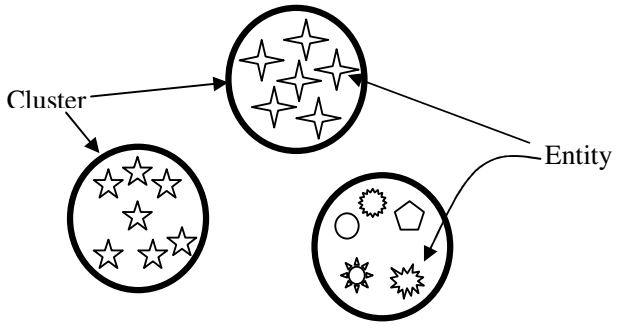

Figure 1.1 Entities having very similar properties are in same cluster and others are in different.

The figure 1.1 shows how a cluster can be formed. The goal of clustering methods is to dig out an existing 'natural' cluster structure. However, different methods may come up with different clustering. So a particular algorithm may force a structure rather than find an existing one. It might even be the case that an algorithm 'finds' a structure while there really is no natural structure in the data. In the figure above we have shown how the homogeneous data is collected into similar clusters.

Cluster analysis is a multivariate statistical technique for grouping cases of data based on the similarity of responses to several variables/subjects. The purpose of cluster analysis is to place subjects/objects into groups, or clusters, suggested by the data, such that objects in a given cluster are homogenous in some sense, and objects in different clusters are dissimilar to a great extent. In cluster analysis, the groups are not predefined but are rather suggested on the basis of the data. The cluster analysis can also be used to summarize data rather than to find observed clusters. This process is sometimes called dissection.

Clustering techniques have been applied to a wide variety of research problems. Hartigan (1975) provides an excellent summary of the many published studies reporting the results of cluster analyses. For example, in the field of medicine, clustering diseases, cures for diseases, or symptoms of diseases can lead to very useful taxonomies. In the field of psychiatry, the correct diagnosis of clusters of symptoms such as paranoia, schizophrenia, etc. is essential for successful therapy. In archeology, researchers have attempted to establish taxonomies of stone tools, funeral objects, etc. by applying cluster analytic techniques. In general, whenever one needs to classify a "mountain" of information into manageable meaningful piles, cluster analysis is of great utility.

Two-Step clustering method is used for clustering process. Two-Step Cluster is an algorithm primarily designed to analyze large datasets. The algorithm groups the observations in clusters. The procedure uses an agglomerative hierarchical clustering method. Compared to classical methods of cluster analysis, Two-Step is capable of handling both continuous and categorical variables and attributes, and it requires only one data pass in the procedure Moreover, the method can automatically determine the optimal number of clusters. In the first step of the procedure, you pre-cluster the records into many small sub-clusters. Then, cluster the sub-clusters from the pre-cluster step into the desired number of clusters. The number of industries increases themselves, but they are facing a lot of problems to satisfy the customers, as we have noted customers move from one company to different by porting system. To retain in the market a company need new schemes for the customer regularly. Our proposed methodology will be a great milestone the era of telecommunication.The results gathered from running a simulation are consistently accurate and scalable in performance.

2.1 SOME DEFINITIONS OF CLUSTER STUDY

*Statistically*Cluster analysis is used to classify data into clusters using neural network algorithms. It is a general approach to multivariate problems whose aim is to determine whether the individuals fall into groups or clusters.

*Business Definition*for Cluster Analysis refers that it is a statistical method used to analyze complex data and identify groupings that share common features. Cluster analysis is a form of multivariate analysis that attempts to explain variability in a set of data. It involves finding unifying elements that enable identification of groups or clusters displaying common characteristics. It could be used, for example, to analyze results of attitude research and delineate groups of respondents that share certain attitudes.

*Dental Dictionary*suggested cluster analysis as a complex statistical technique of data analysis of numeric scale scores, produces clusters of variables allied to one another.

*Geography Dictionary*defined cluster analysis as a type of multivariate analysis which aims to group a set of variables or individuals into classes, so that the objects in each class are like each other as possible and as unlike the other classes as possible, as defined by a designated list of characteristics and indicators. In social geography, the technique can be used to create classifications of, for example, urban areas by type. In general, the classification process begins by drawing up a table of correlation coefficients of dissimilarity between each pair of objects. From here, the objects can be combined into larger and larger groups, or broken down into smaller and smaller ones.

*Sports Science and Medicine*cluster analysis is a technique used to differentiate between subgroups within a single collection of information made about a group, people, or object.

2.2 OBJECTIVES OF CLUSTERING APPROACH

Classifying data into natural groupings on the basis of similar or related properties.

Developing a more homogeneous group of items from a large list of dissimilar items.

In abstract terms we can put together automatic tagging as the grouping of large amounts of things in groups (modules) in such a way that the things in one group are closely related compared to the relationships between things in different groups. In cluster analysis such groups are called clusters. Clusters define in a similar mode as "continuous regions of space containing a relatively high density of points, separated from other such regions by regions containing a relatively low density of points". This is a very general definition which appeals to our intuition. The goal of clustering methods is to extract an existing 'natural' cluster structure. However, different methods may come up with different clustering. So a particular algorithm may impose a structure rather than find an existing one. It might even be the case that an algorithm 'finds' a structure while there really is no natural structure in the data. Random hypotheses (performing the algorithm on random data sets, which have no structure) can be used to check on this phenomenon. When the entities are pieces of software (systems) imposing a structure need not be a problem. It can actually be turned into an advantage as imposing a structure on those pieces is exactly what we want to achieve. By selecting an appropriate method we can steer the clustering and make it suit to our ideas of a good cataloging of modules.

The objective of cluster analysis is to assign observations to groups (clusters) so that observations within each group are similar to one another with respect to variables or attributes of interest and the groups themselves stand apart from one another. In their words, the objective is to divide the observations into homogeneous and distinct groups. In contrast to the classification problem where each observation is known to belong to one of a number of groups and the objective is to predict the group to which a new observation belongs, cluster analysis seeks to discover the number and composition of the groups. There are a number of clustering methods. One method, for example, begins with as many groups as there are observations, and then systematically merges observations to reduce the number of groups by one, two… until a single group containing all observations is formed. Another method begins with a given number of groups and an arbitrary assignment of the observations to the groups, and then reassigns the observations one by one so that ultimately each observation belongs to the nearest group. Cluster analysis is also used to group variables into homogeneous and distinct groups. This approach is used, for example, in revising a questionnaire on the basis of responses received to a draft of the questionnaire. The grouping of the questions by means of cluster analysis helps to identify redundant questions and reduce their number, thus improving the chances of a good response rate is happened.

REFERENCES

[1] Shih M., Jheng J. And Lai L., "A Two-Step Method for Clustering Mixed Categorical and Numeric Data", Tamkang Journal of Science and Engineering, Vol. 13, Issue 1, pp 11-19, 2010

[2] Johnson A. J., Johnson H. C., Devadoss S. And Foltz J., "Strategic Group Analysis of U.S. Food Business Using the Two-Step Clustering Method", International Food and Agribusiness Management Review Vol. 14, Issue 2, pp 83-102, 2011.

[3] Strauch M. et al., "A Two-Step Clustering for 3-D Gene Expression Data Reveals the Main Features of the Arabidopsis Stress Response", Journal of Integrative Bioinformatics, 2007.

[4] Schiopu D., "Applying TwoStep Cluster Analysis for Identifying Bank Customers' Profile", Petroleum-Gas University of Ploiesti, Romania, Vol. 62, Issue 3, pp 66-75, 2010.

[5] Jai Bhagwan and Ashish Oberoi, "Software Modules Clustering: An Effective Approach for Reusability", Journal of Information Engineering and Applications, Vol. 1, Issue 4, 2011.

[6] Czibula I. G., Serban G., "Hierarchical Clustering for Software System Restructuring", Babes Bolyai University, Romania, 2007.

[7] Wiggerts T. A., "Using Clustering Algorithms in Legacy Systems Remodularization", IEEE, pp 33-47, 1997.

[8] Sembiring R. W. et al., "A Comparative Agglomerative Hierarchical Clustering Method to Cluster Implemented Course", Journal of Computing, Vol. 2, Issue 12, pp 1-6, 2010.