Summary of Cluster Analysis of Economic Data

The paper discusses cluster analysis algorithms, particularly in the context of economic data. Cluster analysis aims at identifying groups of similar objects based on selected variables, overall cluster analysis is a powerful tool in multivariate data exploration. The fundamental concept in cluster analysis is similarity. Similarity is frequently assessed by calculating the distance between two objects. When dealing with quantitative variables, a widely used measure is the Euclidean distance. The Euclidean distance formula is expressed as:

Here, represents the Euclidean distance between objects I and j, m denotes the number of variables (such as economic indicators), and , are the values for the -th variable of objects I and j respectively. The formula calculates the square root of the sum of squared differences between corresponding variable values of the two objects, providing a quantitative measure of their dissimilarity/similarity.

Recent developments in cluster analysis focus on challenges like clustering large datasets, handling categorical variables, and incorporating fuzzy clustering. Issues such as outlier detection and cluster number determination remain relevant. The resurgence of interest, especially in connection with evolving data mining techniques, has led to the proposal of new algorithms. The article explores the evolution of different clustering methods.

The author then goes on to present different clustering algorithms starting with the classical ones, namely Hierarchical cluster analysis and K-clustering. The first algorithm discussed, and arguably the most widely used in the context of economic data, is agglomerative hierarchical cluster analysis, this method utilizes a proximity matrix, which includes the similarity evaluation for all pairs of objects, accommodating various measures for different variable types. It offers advantages such as graphical output in the form of a dendrogram, especially useful for smaller datasets. The article demonstrates the application of hierarchical cluster analysis on EU countries based on economic activity rates, providing insights into country groupings. Additionally, the paper explores clustering nominal variables. The basic measure when dealing with nominal variables is simple matching coefficients or in other word the overlap measure. The similarity between vectors xi and xj, denoted as sij, is calculated by comparing values in the ith and jth rows for all variables. The relationship for the lth variable is expressed as slij. If xil = xjl, then slij = 1; otherwise, slij = 0. The overall similarity, sij, is computed as the arithmetic mean.

The article also motions that additionally hierarchical cluster analysis can directly utilize a proximity matrix to assess relationships among all pairs of variables, facilitating the identification of similarity in variables and groups of variables within a dendrogram.

Additionally, the paper explores the challenges and solutions related to clustering nominal variables, introducing measures like Eskin, OF, IOF, and Lin to address mismatch weights and varying category numbers, since the overlap measure does not consider different numbers of categories for individual variables.

The next algorithm discussed is k-clustering, here objects are divided into a specified number (k) of clusters, and various approaches can be classified from different perspectives. The primary classifications include hard and fuzzy clustering. Hard clustering assigns each object exclusively to one cluster, generating a membership matrix with ones (object assigned to cluster) and zeroes (object not assigned). In fuzzy clustering, membership degrees are calculated for all cluster-object pairs, introducing uncertainty expressions in cluster analysis.

Another classification distinguishes k-centroid and k-medoids clustering. In k-centroid, the cluster center is represented by a vector of variable characteristics, while k-medoids represents the center by a selected object from the input matrix. One widely used k-centroid technique is the k-means algorithm, which minimizes an objective function involving Euclidean distances.

The variables , taking values in the set {0, 1}, represent the assignment of object vectors to clusters, where 1 signifies the assignment. The term corresponds to the Euclidean distance between the jth object and the center (a vector of means) of the hth cluster. The following conditions must be satisfied:

Overall, the objective function represents the Hard C-Means (HCM) algorithm's goal to minimize the sum of squared Euclidean distances in the process of clustering. In the equation, k signifies the number of clusters, n denotes the total number of objects, is a binary variable indicating whether object i belongs to cluster ℎ and represents the squared Euclidean distance between the jth object and the center of the hth cluster. The objective is to find optimal assignments and cluster centers that minimize the total within-cluster sum of squares. The conditions ensure that each object is precisely assigned to one cluster and that each cluster has at least one assigned object, essential for the coherence of the clustering results.

The text notes that K-centroid clustering is advantageous for large datasets due to its scalability but is prone to instability, yielding different results for various orders of object vectors. The outcome is influenced by the initialization method for determining initial centroids, and while these methods seek optimal solutions, the achieved optimality may be local rather than global. Despite these challenges, k-clustering methods play a crucial role in exploratory data analysis.

On the other hand, k-medoids, while more stable, might be less suitable for large datasets. The hard k-medoids algorithm, also known as PAM (Partitioning Around Medoids), minimizes an objective function to determine cluster assignments, where mh represents the medoid of the hth cluster.

The expression represents the objective function for the k-means algorithm, where k denotes the number of clusters, h and i are indices iterating over clusters and data points, respectively. The term taking values in {0, 1}, signifies the assignment of data points to clusters, with 1 indicating the assignment. The expression ‖x\_i-m\_i ‖ corresponds to the Euclidean distance between the ith data point and the center (mean) of the hth cluster. The objective is to minimize this function, determining optimal cluster assignments by iteratively updating cluster centroids based on the data points' distances. In contrast to the k-medoids objective function, which uses the actual data point as the center (medoid) of the cluster, the k-means objective function utilizes the mean, making it sensitive to outliers and noise but allowing for efficient application to large datasets.

Next the article discusses fuzzy cluster analysis, fuzzy cluster analysis is implemented to address the inherent uncertainty and imprecision present in real-world data. Unlike traditional hard clustering methods that assign objects strictly to one cluster, fuzzy clustering allows for a more nuanced representation of membership by assigning degrees of belongingness to multiple clusters. Among various algorithms for fuzzy clustering, the fuzzy k-means (FCM) algorithm, also known as fuzzy c-means, stands out. FCM minimizes the objective function, involving membership degrees (uih) with values between 0 and 1, and a fuzzifier parameter (q) to control the degree of fuzziness.

In FCM the fuzziness parameter q controls the sensitivity of these membership degrees to changes in distances between objects and cluster centroids, q is a weighting exponent, and its value is typically set greater than 1. Commonly, q is set to 2 in FCM, known as the Euclidean norm case. When q is 2, the algorithm tends to be more sensitive to the Euclidean distances between objects and cluster centroids. Higher values of q (greater than 2) lead to increased fuzziness, allowing objects to have more evenly distributed membership degrees across multiple clusters. Lower values of q (less than 2) result in sharper membership assignments, making objects more distinctly belong to one cluster. Overall, the parameter q controls the trade-off between soft and hard assignments in fuzzy clustering.

The author mentions encounter challenges with sensitivity to noise and outliers, as exemplified when objects are equidistant from cluster centroids, resulting in equal membership degrees. To address this, alternative algorithms like the possibilistic k-means (PCM) are introduced, minimizing an objective function that incorporates membership degrees a fuzzifier (q), and specific conditions. Furthermore, rough set theory offers an alternative to hard clustering, with the rough k-means (RCM) algorithm defining clusters based on lower and upper approximations. The RCM algorithm introduces two values characterizing the membership of an object, hence providing more nuanced assignments. Combining rough and fuzzy concepts yields the rough-fuzzy k-means (RFCM) algorithm. Additionally, modifications such as the shadowed k-means (SCM) incorporate user-defined thresholds for dynamic cluster evaluation based on original data. These diverse approaches highlight the ongoing efforts to enhance clustering techniques by addressing their inherent limitations and tailoring them to various data characteristics.

Finally, the article delves into other approaches, than hierarchical clustering and K-clustering, which have been proposed to address specific challenges in clustering, such as handling large data sets and incorporating categorical variables. The two-step cluster analysis, exemplified in IBM SPSS Statistics, utilizes the BIRCH algorithm to cluster large data sets with both quantitative and qualitative variables. This method first organizes data into cluster features (CFs) and then applies hierarchical clustering to these CFs. However, it is sensitive to the order of objects. Two-step cluster analysis allows users to employ either Euclidean distance for quantitative data or log-likelihood distance for a combination of quantitative and qualitative variables. This method has been applied to cluster households based on material deprivation indicators, illustrating its effectiveness in handling large data sets with mixed variable types. Alternative techniques, such as CLARA for large applications and BCLUST for high-dimensional data in R, offer additional solutions to clustering challenges. Approaches for clustering categorical data are summarized, and when dealing with mixed-type variables, cluster ensembles, like CLUE for R, provide a viable solution by combining individual cluster solutions for different variable types.