Proposal for a Hybrid Data Clustering Algorithm with Cohesion-Based Self Merging

ISSN 2319-9725

B. Geetha
Associate Professor
Dr.SNS. Rajalakshmi college of Arts and Science

Abstract: This work studies and compares different clustering algorithm and different similarity measure. The objective of the study is to combine the ideas of partitioned clustering algorithm and hierarchical clustering algorithm and to propose a new hybrid algorithm. A new similarity measure called cohesion is proposed using cohesion a two-phase clustering algorithm is designed, called Cohesion-Based Self-Merging (CSM) which runs in time linear to the size of input data set. The performance studies of this work show that the proposed similarity measure cohesion is more effective than others. With cohesion algorithm CSM is not only very robust to the existence of outliers, but also lead to better clustering results than most prior methods while incurring much shortest method.

Keywords: Cohesion-Based Self-Merging (CSM), single-link algorithm, complete-link algorithm, data mining, hierarchical clustering, outliers.
1. **Introduction:**

The requirements for problem solving are increasing in day to day life because of the fast development of computer technologies. The problems were different in different applications. To solve this type of real world problem in computer technologies variety of solutions were driven. There are varieties of data in the problem. To handle this different type of data to solve the problem as fast as possible with in finite time is very difficult. Several approaches are available to analyze the data.

Cluster analysis is one of the common approaches. Once the data is analyzed, the problem can be easily solved. Problem solving techniques consists of several steps. Cluster analysis has been widely used in information retrieval, text and web mining, pattern recognition, image segmentation and software reverse engineering.

Clustering involves partitioning a given data set into several groups based on some similarity / dissimilarity measurements. The main task of this thesis is to study the comparison of similarity measurements of hierarchical clustering algorithm and partitioned clustering algorithm and to propose a new clustering algorithm using cohesion-based self-merging. This new efficient approach is to provide better clustering results than by prior methods.

2. **Review Of Literature:**

Due to the usefulness of data mining in many application including selection of marketing, decision support, managing of business, the significant of data mining has attracted significantly in research.

In cluster algorithm, there are two types of attributes are associated with input data are numerical attributes and categorical attributes. In the exploratory data analysis, the cluster algorithm is an important technical and most used in research problems.

The errors in the data or could be correct data may represented as outliers that are merely from remaining data.
3. Method:

Depending on the strategies used in clustering, clustering methods can be grouped into one of two categories are hierarchical methods or partitioning method.

In hierarchical structure, different clustering results can be obtained for different similarity requirements. The single-link and complete–link algorithm are most existing algorithm of hierarchical clustering algorithm with time complexity of $O(n^2 \log n)$.

In single-link algorithm, the two closet points of two clusters are used the distance between where as in the complete-link algorithm uses two outer most points. The non-supporting of some mechanism in single-link algorithm and complete-link algorithm the CURE algorithm is proposed.

To measure the similarity of two clusters, Cohesion based self merging algorithm is proposed. It is basically different superior from other measures.

4. Report:

In this paper, a new similarity measure is proposed, cohesion, to measure the intercluster distances. By using cohesion, a two-phase clustering algorithm, CSM are proposed whose time complexity is linear to the size of the input data set. Combining the features of partitioned and hierarchical algorithms, algorithm CSM is able to not only resist outliers, but also lead to high quality clustering results while incurring a much shorter execution time than other algorithms.

The time and the space complexities of CSM are also analyzed. The cohesion-based clustering is very robust and possesses excellent tolerance to outliers in various workloads shown in the performance studies. More importantly, algorithm CSM is shown to be able to cluster the data sets of arbitrary shapes very efficiently and to provide better clustering results than those by prior methods. The CSM algorithm can be modified by combining any clustering algorithm in future.
References:

15. Margaret H Dunham, “Data Mining Introductory and Advanced topics”, 2003


Web Sites:

1. APPROXIMATION TO K-MEANS-TYPE CLUSTERING – By YU WEI

2. A Tutorial on Clustering Algorithms
   www.elet.polimi.it/upload/matteucc/Clustering/tutorial_html

3. An Introduction to Cluster Analysis for Data Mining
   www.cs.umn.edu/~han/dmclass/cluster_survey_10_02_00.pdf


5. Cluster analysis - From Wikipedia, the free encyclopedia by
   http://en.wikipedia.org/wiki/Cluster_analysis