

Parallel Clustering Algorithms for Categorical and Mixed Data

NGUYEN Thi Minh Hai

School of Information Science,

Japan Advanced Institute of Science and Technology

February 13, 2003

Keywords: clustering, categorical data, mixed data, very large databases, parallel algorithm.

1. Introduction

Clustering is a fundamental and important technique in many research fields such as image processing, pattern recognition, machine learning, etc. Clustering is used to group a set of unlabeled data objects into smaller groups of similar objects to simplify data management and data mining for useful knowledge. Each group, called a cluster, is a collection of data objects satisfying the condition that objects within one cluster are similar to one another and dissimilar to objects in other clusters.

From a practical viewpoint, it should be possible to apply clustering methods to various types of databases. In general, data types are classified into two types: categorical data and mixed data. Mixed data is the data containing both numeric and categorical attributes. Until recently, the clustering task for categorical and mixed data has been widely studied. However, it is unable to perform arithmetic operations and order comparison on categorical values. Thus, it is difficult to find suitable measures for categorical data and mixed data. A number of researches tried to convert categorical data to numerical data and used normal similarity measure of numeric data for the converted data. However, these conversions may cause the loss of sensitive information of data, which lead to bad clustering results.

The original k-means is a well-known algorithm for clustering tasks, but it is designed to work primarily on numeric databases. This prevents the algorithm from being directly applied on the categorical data which appears in many data mining applications. Many clustering algorithms were extended from the k-means paradigm in order to deal with categorical data and mixed data. Unfortunately, these algorithms fail into either stopping at local optimization or depending on the initial selected groups. Besides, most of the remainder of the above categorical algorithms is very time consuming when choosing hierarchical approach for clustering task. This prevents the algorithms from processing with large data sets in acceptable time. Recently, a clustering algorithm for categorical and mixed data named k-sets was proposed without drawback that cluster results depend on random selection of initial points (modes) as of the clusters of previous clustering algorithms. The k-sets can also overcome stopping at local optimization by its method to construct k cores of k

clusters. Unfortunately, the k-sets does not always achieve high accuracy clustering result and also can not deal with very large databases either.

Besides of dealing with complex structure data, it is also important for the clustering algorithm to deal with rapidly increasing size of data. Because the recent databases are very large with gigabytes and much larger in the future, a serial clustering algorithm seems to be impossible to process them in acceptable time. Furthermore, when the size of database becomes very large, data can not be stored in a single memory. As a result, the recent development of high performance technologies is hopeful to enable clustering with complex and huge databases. Most of the recent parallel algorithms were implemented to deal with only numeric data therefore it is still needed to develop parallel algorithms to deal with complex data.

In summary, clustering task for categorical data and mixed data is very necessary for managing, analyzing real databases. In addition, the ability of dealing with huge databases of clustering algorithm is not less important. The effective and efficient clustering algorithms for large database with categorical attributes and mixed attributes are extremely essential.

2. Parallel HAC Algorithms for Categorical Data and Mixed Data

We proposed a High Accuracy Clustering algorithm for categorical and mixed data (HAC algorithm) to achieve good clustering results. In addition, to deal with the problem of very huge database, two parallel approaches for the HAC algorithm are also presented in this thesis.

2.1. High Accuracy Clustering Algorithm for Categorical and Mixed Data based on a new Grouping Approach

The k-sets algorithm is quite good compared with the other clustering algorithms. However, it can not always achieve high accuracy clustering results. Thus, in this thesis, we propose an improving approach for the re-partition step of k-sets to aim at achieving higher quality in clustering results.

When trying to assign remaining data objects to clusters (re-partition step), the original k-sets finds the closest cluster for each object of the N-ESC sets. These N-ESC set are the sets such that the objects in the same N-ESC set are more similar than the objects in different N-ESC sets according to the similarity measure of the k-sets. As a result, after re-partition step, the objects which are similar to each other, i.e. objects belong to the same N-ESC set, may be separated into different clusters. Then, the quality of clustering task will be decreased. To overcome the drawback, we try to assign all objects of the same built N-ESC set to the same cluster. This is hoped to increase the accuracy of clustering results.

2.2. Parallel HAC Algorithms for Categorical and Mixed data

An effective and efficient clustering algorithm is the one capable in dealing with complex and huge databases. Parallelization is one useful tool which helps the clustering algorithm can save much processing time. These facts encourage us to apply parallel techniques to our HAC algorithm to deal with complex and large databases in an acceptable time.

HAC algorithm is an algorithm clustering database into k groups. There are several parameters have to be provided. To achieve high accuracy in the results, the HAC algorithm will considers all possible values of the input parameters, i.e. that parameters will be determined automatically by the algorithm after trying all their possible values. These trials make the algorithm a long time for processing. We then employ parallel techniques to upgrade HAC to Single-level Parallel HAC algorithm and Multi-level Parallel HAC algorithm. The first one just considers the trials of one input parameter while the second considers the trials of more than one parameter, which is much more effective than the first one in case of dealing with mixed data.

3. Performances

In our thesis, the performance of the HAC algorithm was calculated to evaluate the clustering result of the algorithm. The experiments were done on the same databases used in related works, for a sound comparison. In addition, we did experiments to evaluate performance of our proposed parallel algorithms with various data sets on Cray T3e. We also discussed in detail the average execution time, the speed-up, the efficiency and losing time for parallel tasks.

Our experiment results showed that the HAC algorithm always get highest accuracy compare with the previous algorithms on the same databases and accuracy calculating method. In additions, the two proposed parallel algorithms based on the HAC algorithm help to decrease significantly processing time and can achieve greater than 90 times in speed-up on 120 processors.

4. Conclusions

We have discussed the related researches to determine the motivation and goals of this work. In the thesis, we have already proposed a new clustering algorithm for categorical and mixed data based on improvement of the re-partition step in the k -sets algorithm. The proposed serial algorithm named High Accuracy Clustering Algorithm for Categorical and Mixed Data based on a new Grouping Approach (HAC algorithm). This algorithm determines k clusters by selecting k largest groups on whole data objects. It avoids the limit of local optimization and it does not depend on the data input sequence. Our HAC algorithm was shown as the good clustering method for categorical and mixed data by showing highest accuracy compared with the previous researches on the same databases. We also have proposed two parallel algorithms based on HAC algorithm to cluster very huge databases. The proposed algorithms were shown as scalable parallel algorithms by spending on much smaller time than the serial one. They also are effective parallel algorithms with high speed up on our experiments. In summary, our proposed parallel algorithms are able to deal with complex and huge databases. In the future, we intend to develop the suitable measure to increase effectiveness and efficiency of the HAC algorithm as well as to develop a more scalable parallel clustering algorithm for complex data based on logical of k -means algorithms.