A Study On Point-Based Clustering Aggregation Using Data Fragments

Yamini Chalasani
Department of Computer Science & Engineering, V.R.Siddhartha Engineering College(Autonomous)
Affiliated to JNTUK, Vijayawada, Andhra Pradesh, India.

M.Vani Pujitha
Department of Computer Science & Engineering, V.R.Siddhartha Engineering College(Autonomous)
Affiliated to JNTUK, Vijayawada, Andhra Pradesh, India.

Abstract

Point-based clustering aggregation is applying aggregation algorithms to data points and then combining various clustering results. Applying clustering algorithms to data points increases the computational complexity and decreases the accuracy. Many existing clustering aggregation algorithms have a time complexity quadratic, cubic, or even exponential in the number of data points. Thus Data fragments are considered. A Data fragment is any subset of the data that is not split by any of the clustering results. Existing model gives high clustering error rate due to lack of preprocessing of outliers. Non spherical clusters will not be split by using distance metric. In this paper we made a study on agglomerative clustering aggregation algorithm used in different areas. In the proposed approach, data fragments are considered and Outlier detection techniques are employed for preprocessing of data. New clustering aggregation algorithm proposed includes the outlier detection technique and each disjoined set of fragments is clustered in parallel thus reducing the time complexity.

1. Introduction

Clustering is a hot research field in data mining. There are so many methods or algorithms designed for different type of data set on which data analysis action operates. Clustering is data analysis which assigns data objects into clusters so that similar objects are in the same cluster and dissimilar objects are in different clusters. Clustering aggregation provides a method for improving the clustering robustness by combining various clustering results. It determines the appropriate number of clusters. Clustering aggregation is applied in many different disciplines such as machine learning, pattern recognition, bioinformatics and information retrieval. Various applications of clustering aggregation are identifying the correct number of clusters, detecting outliers, improving clustering robustness and privacy preserving clustering. The algorithm that is mainly used for clustering aggregation is Agglomerative. Agglomerative algorithm performs in a bottom-up fashion, which initially takes each data points as a cluster and then repeatedly merges clusters until all data points have been merged into a single cluster [7]. Agglomerative algorithm creates a multilevel hierarchy tree, where clusters at one level are jointed as clusters at the next high level.

Ou Wu et al. [1] proposed that clustering aggregation algorithms can also be applied to data fragments instead of data points. A data fragment is any subset of the data that is not split by any of the clustering results. As the number of data fragments is much less than the number of data points the computational complexity decreases.

But this gives high clustering error rate due to lack of preprocessing of outliers. Thus we include outlier detection technique prior to applying aggregation algorithm on the data set. Agglomerative algorithm can be applied in parallel process for clustering data fragments which reduces the time complexity [4].

1.1 Basic agglomerative algorithm:

1. Place each node into a singleton cluster
2. Consider the pair of clusters with the smallest average distance
3. If average distance of the closest pair of clusters is less than 1/2 → merge them
4. If there are no two clusters with average distance smaller than 1/2 → stop

The paper is organised as follows: section 2 discusses about related work, section 3 about Proposed Work, section 4 about Conclusions and Future Work

2. Related work

Xi-xian niu et al. [2] proposed the Local Agglomerative Characteristic (LAC) algorithm which mainly focuses on the local agglomerative characteristic of the data objects. The Main idea of LAC clustering is that two objects have higher similarity if they have the k shared nearest neighbor and have the relative higher local agglomerative characteristic in local data objects area at the same time. Local characteristic is reflected by Local Average Distance (LAD) and Local Maximum Distance (LMD). Both LAD and LMD can reflect the local area data distribution characteristic. First, LMD is taken as local dynamic threshold, through simple compare and computation can get the LMD, but it is sensitive to local data point's distribution shape. For the limitation of LMD's representative of local data characteristics, second, LAD reflection of local data objects is checked out.

Advantages of this technique are it eliminates noisy points using LAD threshold. Proposed similarity measure is not only limited to consider the direct nearest neighbors, but also can take into the neighbor's local distributed feature. It is relatively resistant to noise and can handle clusters of arbitrary shapes and sizes, can deal with clusters of different density and natural distribution characteristics.

Drawbacks are that LMD and LAD both can give representative in some degree, but LMD can show direction and shape information but not represent most data point's, and LAD can reflect most data objects relative distance but direction information lost.

Ying Peng Yongyi Ma et al. [3] proposed an algorithm for belief functions. The information carrier in Dempster-Shafer theory (DST) is belief function. Combination of belief functions is required for getting a fusion result [9], [10]. Combination is performed just on condition that belief functions are related to the same event. It is necessary to distinguish which belief functions are reporting on which event. Here agglomerative algorithm is used for clustering purpose. Belief distance is taken as the dissimilarity measure between two belief functions, so there is no need of transformation. And due to the utilization of agglomerative algorithm, there is no need to set cluster number in advance. After getting the hierarchical tree, cluster number by threshold value is determined. Agglomerative algorithm creates a multilevel hierarchy tree, where clusters at one level are jointed as clusters at the next high level.

Advantages of this technique are it overcomes the problem of indirect clustering for possible inequality of transformation. This approach allows constructing clusters within uncertain information.

Drawbacks are Clustering approach used in this system virtually based on comparison between two belief function, which may has problems of hidden conflict among beliefs in one cluster. Partitioning tree depends on the level wise threshold values which would take more time to construct.

Cheng-Hsien Tang et al. [4] proposed the Distributed Hierarchical Agglomerative clustering algorithm which divides the whole computation into several small tasks, distribute the tasks to message-passing processes, and merge the results to form a hierarchical cluster [11],[12].This clustering algorithm uses the reduced similarity matrix to sequentially create disjoined sets of highly related data items. Each disjoined set is clustered in parallel to form a hierarchical sub-tree. Finally, the algorithm computes the similarity scores among all the sub-trees to form a complete hierarchical tree. To justify whether a data item belongs to a disjoined set, the distance (similarity) between two disjoined sets are to be defined. The similarity matrix (distance matrix) is a matrix that stores the similarity scores of all pairs of data items. A naive computation strategy that can concurrently calculate the matrix is to process each row in parallel.

Advantages of this technique are it takes less time to construct clusters due to parallel process. Takes less overhead i.e., if one processor handles one row, the execution time should depend on the time for computing the last row because it has the most work to do. Parallel computing provides a good way to handle large data sets.

Drawbacks are the space complexity of a similarity matrix is O (n^2) given n data items. If outliers are out of interest, only a small portion of similarity matrix is used to construct a hierarchical tree. It suffers with data clusters size, shape and outliers.

Jaruloj Chongstitvataa et al. [5] proposed an Agglomerative clustering algorithm which uses the concept of Attraction and Distraction provides higher accuracy for iris and haberman data sets when compared to K-means algorithm. A cluster of data objects can form either a concave shape or a convex shape. This method uses the distance between clusters and the cluster size as parameters for clustering. Clusters of objects are formed by attraction and distraction. In this work, Euclidean distance is used as the measurement of dissimilarity between objects. The distance between a pair of clusters is measured by the distance between the closest pair of points in each cluster. Attraction indicates if two clusters can be merged, based on the number of similar objects between two clusters, compared to the size of the cluster. Distraction indicates if the merging of two clusters should be deferred, based on other possible merge. In this method, a cluster is considered too small to be a cluster by itself if it is smaller than the median of the size of all clusters. Each of these small clusters is merged with its nearest neighbor cluster [13], [14]. It is found that this algorithm yields better accuracy on some datasets.

Advantages of this technique are it overcomes the restriction of the cluster shape, the concepts of attraction and distraction is used in this system effectively. The overall accuracy of the proposed method is better than K-means algorithm.

Drawback is that it always performs for concave shape clusters and had Quadratic time complexity.

Rashid Naseem et al. [6] proposed an Agglomerative Clustering technique used for restructuring of the program using Binary Features [15], [16]. This uses the Complete Linkage (CL) algorithm with Jaccard similarity measure using binary features, to group the similar statements into a noncohesive structured program. Binary features just indicate the absence or presence of a feature. The correct translation of a program into group of statements makes cohesive procedures. A program that is incorrectly translated may result in more problems as compared to original one. The programs taken from source code written in structured languages are restructured. A similarity measure is applied to compute similarity between every pair of entities, resulting in a similarity matrix. After applying clustering algorithm, hierarchy of results obtained and can be shown with the help of tree like structure known as dendrogram.

Advantages of this technique are this approach has the added benefit that it is very simple to understand and implement. It uses binary features, just to indicate the presence and absence of features. Effectively identifies the program reconstruction and program structures and tokens information. This helps to translate a non cohesive procedure into cohesive procedures.

Drawbacks are it does not suit for non structure programs. It does not handle special characters and new keywords informations. It does not work for non binary features.

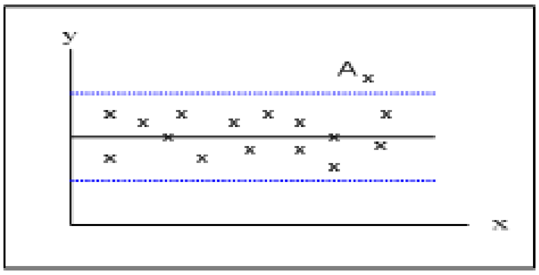
3. Proposed work

Existing Fragment based Agglomerative clustering algorithms does not concentrate on outliers. “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”. Accuracy of outlier detection depends on how good the clustering algorithm captures the structure of clusters. A set of many abnormal data objects that are similar to each other would be recognized as a cluster rather than as noise/outliers. Clustering algorithms are optimized to find clusters rather than outliers. This increases the cluster error rate and decreases the accuracy. Thus in the proposed approach we employ Statistical Control Chart based outlier detection Technique [17] to identify and remove outliers.

Control Chart Technique (CCT): The purpose of a control chart is to detect any unwanted changes in the process. These changes will be signaled by abnormal (outlier) points on the graph. Basically, control chart consists of three basic components:

1) A centre line, usually the mathematical average of all the samples plotted.
2) Upper and lower control limits that define the constraints of common cause variations.
3) Performance data plotted over time.

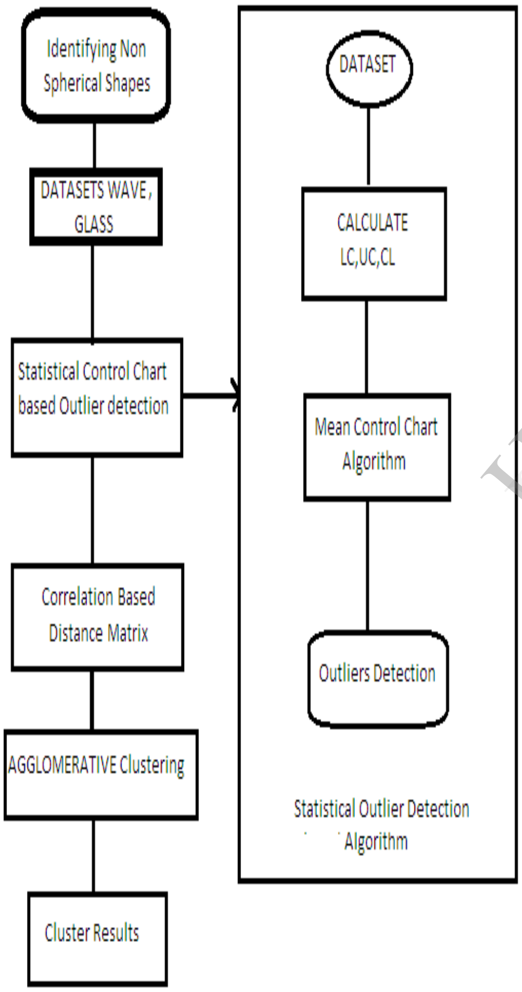
Figure 1 shows an example of how this technique works. Here A is an outlier.



“Figure 1: Example for detecting outliers.”

After detecting and deleting the outliers by using the above technique the data fragments can be clustered by using agglomerative clustering algorithm. Thus we can obtain the better clustering results having less cluster error rate.

Figure 2 shows our proposed approach. Non spherical shapes can be identified by plotting the data points on the graph. Later the outliers in the data set are removed by applying control chart based outlier detection as described above. Then distance matrix can be developed for data fragments [1]. Finally Agglomerative clustering algorithm can be applied in parallel process to cluster disjoint subsets of fragments for getting better clustering results [4].


Identifying Non Spherical Shapes
DATASETS WAVE, GLASS
Statistical Control Chart based Outlier detection
Correlation Based Distance Matrix
AGGLOMERATIVE Clustering
Cluster Results
DATASET
CALCULATE LC,UC,CL
Mean Control Chart Algorithm
Outliers Detection
Statistical Outlier Detection Algorithm

"Figure 2: Proposed method for finding clusters with minimised clustering error rate."

4. Conclusion and future work

According to the rapidly changing technology new clustering algorithms are needed to decrease clustering error rate and increase the accuracy. Existing data mining clustering algorithms are very time consuming and they generate incorrect clusters, hence we extend fragment based agglomerative algorithm to detect outliers and remove them by adding outlier detection technique called CCT and then applying parallel agglomerative clustering algorithm to decrease time complexity. This approach can be applied to non spherical clusters also. Our future work can concentrate on testing this technique for various data sets and check for accuracy. Distance based outlier detection techniques can also be employed.

References

[1] Ou Wu, Member, IEEE,WeimingHu, Senior Member, IEEE, Stephen J. Maybank, Senior Member, IEEE, Mingliang Zhu, and Bing Li " Efficient Clustering Aggregation Based on Data Fragments".

[2] Xi-xian Niu, Kui-he Yang, Dong Fu , "Local Agglomerative Characteristics based Clustering Algorithm", Hebei University of Science and Technology Shijiazhuang.

[3] Ying Peng Yongyi Ma, Huairong Shen "Clustering Belief Functions Using Agglomerative Algorithm", Academy of Equipment Command & Technology Beijing, 101416, China.

[4] Cheng-Hsien Tang, An-Ching Huang, Meng-Feng Tsai, Wei-Jen Wang "An Efficient Distributed Hierarchical-Clustering Algorithm for Large Scale Data", National Central University, Taiwan.

[5] Jaruloj Chongstitvataa and Wanwara Thubtimdang "Clustering by Attraction and Distraction", Chulalongkom University, Bangkok 10330 THAILAND .

[6] Rashid Naseem, Adeel Ahmed, Sajid Ullah Khan, Muhammad Saqib, Masood Habib "Program Restructuring Using Agglomerative Clustering Technique Based on Binary Features".

[7] Aristides Giones ,"Clustering Aggregation" ,Yahoo! Research Labs, Barcelona Heikki Mannila University of Helsinki and Helsinki University of Technology.

[8] Qian Weining, Zhou Aoying, Analyzing Popular Clustering Algorithms from Different Viewpoints. Journal of Software,2002,13(8):1382~1394.

[9] G. Shafer, "A Mathematical Theory of Evidence," Princeton: Princeton University Press.

[10] A. P. Dempster, “Upper and lower probabilities induced by a multivalued mapping,” Annals of Mathematical Statistics, vol. 38, pp. 325–339.

[11] X. Li, “Parallel algorithms for hierarchical clustering and cluster validity,” IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 12, pp. 1088–1092, 1990.

[12] V. Olman, F. Mao, H. Wu, and Y. Xu, “Parallel clustering algorithm for large data sets with applications in bioinformatics,” *IEEE/ACM* Transactions on Computational Biology and Bioinformatics, vol. 6, pp. 344–352, 2009.

[13] J. B. Macqueen, "Some Methods for classification and Analysis of Multivariate Observations," Proceedings of 5th Berkeley Symposium on Mathematical Statistics and Probability, Berkeley: University of California Press,pp. 281-297.

[14] R. C. Tryon, Cluster analysis. New York, McGraw-Hill.

[15] R. Xu and D. Wunsch, “Survey of clustering algorithms,” IEEE Transactionson neural networks, vol. 16, no. 3, pp. 645–678, 2005.

[16] L. Rokach, “A survey of Clustering Algorithms,” Data Mining andKnowledge Discovery Handbook, pp. 269–298, 2010.

[17] Zuriana Abu Bakar, “A Comparative Study for Outlier Detection Techniques in Data Mining”.