OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

An Efficient Clustering Method for Aggregation on Data Fragments

Srivaishnavi D., M. Kalaiarasu
PG Scholar (CSE), Sri Ramakrishna Engineering College, Coimbatore
Associate Professor, Dept. of IT(UG), Sri Ramakrishna Engineering College, Coimbatore

ABSTRACT:*Clustering is an important step in the process of data analysis with applications to numerous fields. Clustering ensembles, has emerged as a powerful technique for combining different clustering results to obtain a quality cluster. Existing clustering aggregation algorithms are applied directly to large number of data points. The algorithms are inefficient if the number of data points is large. This project defines an efficient approach for clustering aggregation based on data fragments. In fragment-based approach, a data fragment is any subset of the data. To increase the efficiency of the proposed approach, the clustering aggregation can be performed directly on data fragments under comparison measure and normalized mutual information measures for clustering aggregation, enhanced clustering aggregation algorithms are described. To show the minimal computational complexity. (Agglomerative, Furthest, and Local Search); nevertheless, which increases the accuracy.*

Key words: *Clustering aggregation, Fragment based approach, point based approach.*

I. Introduction

Clustering is a problem of partitioning data object into clusters that is object belongs to similar groups. Clustering becomes a problem of grouping objects together so that the quality measure is optimized .The goal of data clustering is to partition objects into disjoint clusters. Clustering is based on aggregation concepts.

The objective is to produce a single clustering that agrees m input clustering. Clustering aggregation is a optimization problem where there is a set of m clustering, To find the clustering that minimizes the total number of disagreement in the framework for problems related to clustering. It gives a natural clustering algorithm data. It detects outliers; clustering algorithm is used in machine learning, pattern recognition, bio informatics and information retrieval.

II. Related Works

Point-based clustering aggregation is applying aggregation algorithms to data points and then combining various clustering results. Apply clustering algorithms to data points increase the computational complexity and decrease the accuracy. Many existing clustering aggregation algorithms have a time complexity, decreases the efficiency, in the number of data points. Thus the 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 error rate due to lack of preprocessing of outliers. Non spherical clusters will not be split by using distance metric.

N.Nugen and R.Carunahave proposed each input clustering in effect votes whether two given data points should be in same clustering. Consensus clustering algorithms often generate better clustering, find a combined clustering unattainable by any single clustering algorithm; are less sensitive to noise, outliers or sample variations, and are able to integrate solutions from multiple distributed sources of data or attributes..

X.Z.Fern and C.E.Brodly, it proposes random projection for high dimensional data clustering Data reduction techniqueit does not use any “interestingness “criterion to optimize the projection. Random projections have special promise for high dimensional data clustering. High dimension poses two challenges; The drawback of random projection is that it is highly unstable and different random projection may lead to radically different clustering results.

Inclustering based similarity partitioning algorithm if two objects are in the same cluster then they are considered to be fully similar, and if not they are fully dissimilar. Define the similarity as a fraction of clustering’s in which two objects are in the same clusters.TheHyper graph partitioning algorithmis a direct approach to cluster ensembles that re-partitions the data using the given clusters which indicates the strong bonds .The hyper edge separator that partitions the hyper graph into k unconnected components of same size. The main idea of Meta clustering algorithm is to group and collapse related hyper edges and assign object to the

collapsed hyper edge in which it participates more strongly. The three methods require the constructions of edge between each pair of points; all their complexities are atleast O (N^2).

III. Proposed System

Fragment-based clustering aggregation is applying aggregation algorithms to data points and then combining various clustering results. Applying clustering algorithms to data points increases the time complexity and improves the quality of clustering. In this fragment –based approach, a data fragment is any subset of the subset of the data that is not split by any of the clustering results. The main aim of the fragment based clustering aggregation defines the problem of clustering aggregation and demonstrates the connection between clustering aggregation and correlation clustering.

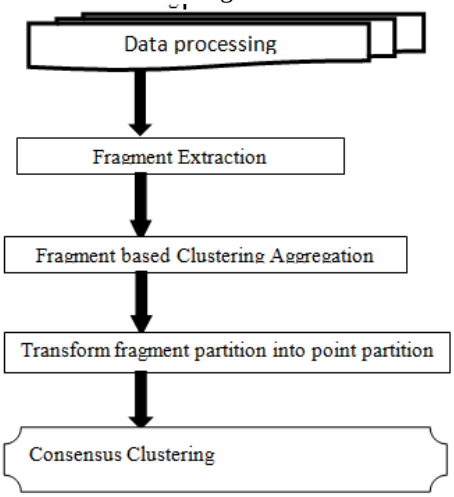

Fig.1. Architecture Diagram for Fragment Based Approach

Preprocess the datasets using the preprocessing technique then extract the data from the preprocessed data sets. The extracted data are fragmented using clustering aggregation. After these steps transform the fragment partition into point partitionwhich results the consensus clustering**.**

IV. Implementation Details

In Fragment based extraction partitions the given data sets into a set of clusters. Each partition divides the data set into three clusters. Large data sets are split into one sentences and then extracted into a one word.The fragments can be represented as

$F=\{F_1,\ldots..F_2,\ldots.Fz\}$

Z-Represent the number of fragments

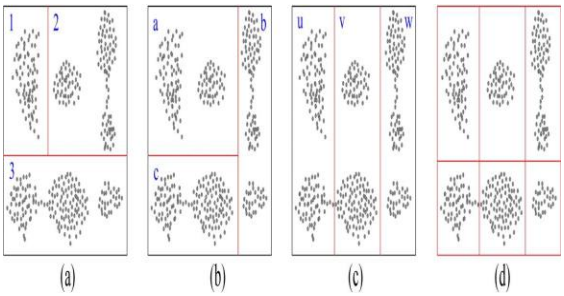

Fig.2.Sample Extraction for Data Fragment

(a), (b), and (c) are three input partitions; (d) shows the union of all the three solutions. Each data subset, enclosed by red lines and red borders in (d), is a data fragment. The number of fragments is six.

The clustering aggregation can be directly achieved on data fragments when either comparision measure or the NMI measure is used. This provides a theoretical basis to guarantee that the fragment based clustering aggregation is feasible.

Clustering Algorithm

4.1. F-Agglomerative

The Agglomerative algorithm is a standard bottom-up algorithm for the correlation clustering problem. It starts by placing every node into a singleton cluster. It then proceeds by considering the pair of clusters with the smallest average distance. The average distance between two clusters is defined as the average weight of the edges between the two clusters. If the average distance of the closest pair of clusters is less than 0.5 then the two clusters are merged into a single cluster. If there are no two clusters with average distance smaller than 0.5, then no merging of current clusters can lead to a solution with improved cost of class.

The average distance between two clusters

$$avg\_dis(\pi ik, \pi jl) \sum_{u \hat{I} \pi ik} \sum_{v \hat{I} \pi jl} DMP(u,v) / (|\pi ik| \, |\pi jl|) \quad \ldots..1.1$$

Where,

u,v-Edges of data point

Avg-Average

D-Distance

Since the proposed system is performed on fragments, the average distance between two clusters can be rewritten as then no merging of current clusters can lead to a solution with improved cost of class.

Algorithm 1

Input: X={x1,….,xi,….Xn},∏={π_1,….π_i,….π_H}

Output: Consensus clustering

Steps:

1) Extract data fragments based on input partitions
2) Calculate the distance matrix for fragments (DMF)
3) Place each fragment in a single cluster.
4) Calculate the average distance between each pair of clusters using formula, and choose the smallest average distance is below 0.5.otherwise, go to the next step.
5) Transform the obtained fragment partition into a data point partition, and return the new data point partition.

4.2. F-Furthest

Furthest algorithm is a top-down algorithm that's work on the correlation clustering problem. It is inspired by the first traversal algorithm, it uses centers to partition the graph in a top-down fashion.

Its starts by placing all nodes into a single cluster .Its find the pair of nodes that are furthest apart and places them into different clusters. Then the nodes become the center of clusters.

The DMP distance is largest, and then makes each point the center of a new cluster. The remaining nodes are assigned to one or other of two clusters according to their distances to the two clusters centers. The procedure is repeated iteratively: at each step, the furthest data point from the existing centers is choosen and taken as a new center of singleton cluster; the nodes are assigned to the new cluster that leads to the least moving cost defined by formula. If the total moving cost is not smaller than 0, the algorithm stops their work.it moves the fragments instead of data points at each step .Equation and can be rewritten for fragment as

$$cost(F\varsigma, \pi_{ik,} \to \pi_{il}) = cost(F\varsigma, \pi_{il}) - cost(F\varsigma, \pi_{ik}) \quad ..1.2$$

Where

Cost-To calculate the value for data points

Fζ-Total number of fragment

DMP-Distance matrix point

Then, the cost of moving Fζfrom cluster πikto πil is

Algorithm 2

Steps of F-Furthest

Input: X={x1… xi…xN}, ∏={π1,…,πi,….πh}

Output: **Consensus clustering**

Steps:

1) Extract data fragments based on the input partitions.
2) Calculate the distance matrix for fragments.
3) Place the entire fragment in a single cluster, then find the pair of fragments whose distance (DMF distance) is furthest, and make each of them the center of a singleton cluster.

4) Assign the rest of the fragments to one of the two clusters in order to achieve the minimum moving cost.
5) Choose the furthest fragment from the existing centers and make the fragment the center of a new singleton cluster.

4.3. F-Local search

Local search algorithm is an application of a local search heuristic to the problem of correlation clustering. The algorithm starts with some clustering of the nodes. The aim of the F-local search algorithm is to find a partition of a data set such that the data points with low DMP distances the values is less than 0.5,the data points are kept together.

Its start with initial clustering is given. The algorithm goes through the all the data point by considering a local movement of a point from a cluster to one another or creating a new singleton cluster with this node. For each local movement, calculate the moving cost by formula .If one data points yield a negative moving cost, it improves the goodness of the current clustering. Repeatedly the node is placed in the cluster that yields minimum moving cost .The algorithm iterates until all moving cots are not smaller than 0, Based on local search ,the steps of F-Local search are detailed in Algorithm 3.

The process is iterated until there is no move that can improve the cost .

4.4.Comparision and Analysis

Three algorithms are used to evaluate the performance analysis .The input to the algorithm are datasets which are evaluate using the following formula**.**

Entropy formula

$$E\chi(\pi\iota) = \sum_{\varphi=1}^{\pi\iota} (|\ \pi\iota\varphi\ |\ -\mu\iota\varphi) / N \quad \ldots..1.3$$

Where
E-Entropy
P-Partition of data sets
N-Number of data sets

Average entropy formula

$$AE(\pi i) = \sum_{j=1}^{pi} |\ \pi ij\ | /N \times \left(\sum_{k=1}^{|c|} -mij/\ |\ \pi ij\ |\ log2mij/\ |\ \pi ij\ | \right) \quad \ldots.1.4$$

Where
AE-Average Entropy
N-Total number of data points

The above formula computes the entropy values which are compared to plot out the performance. Based on the entropy values generated, the agglomerative algorithm has the best performance among the three algorithms

V. Results and Discussion

A graph is plotted to represent the sum of entropy values for three different algorithms. It is observed that the value of the sum of entropy is least for agglomerative compared to other two algorithms.

Table 5.1. Comparision of algorithms with respect to datasets

Data sets	F-Agglomerative	F-local search	F-Furthest search
Glass	**30.84**	**37.38**	**32.3**
Hayes	**30.27**	**37.38**	**30.37**
yeast	**34.51**	**35.48**	**37.62**

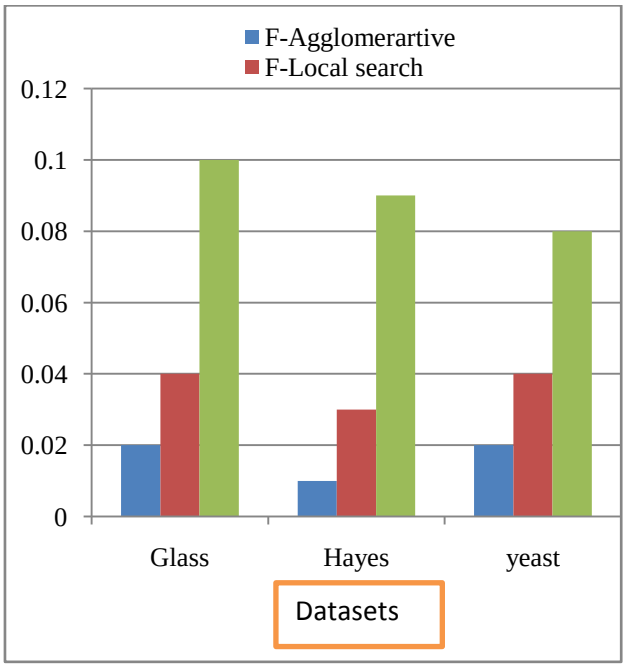

Fig .3. Datasets results using entropy

Table.5.3.Table Comparison of algorithm with respect to data sets

Datasets	F-agglomerative	F-Local search	F-furthest search
Glass	**30.84**	**37.38**	**32.3**
Hayes	**30.27**	**37.38**	**30.37**
yeast	**34.51**	**37.62**	**71.6**

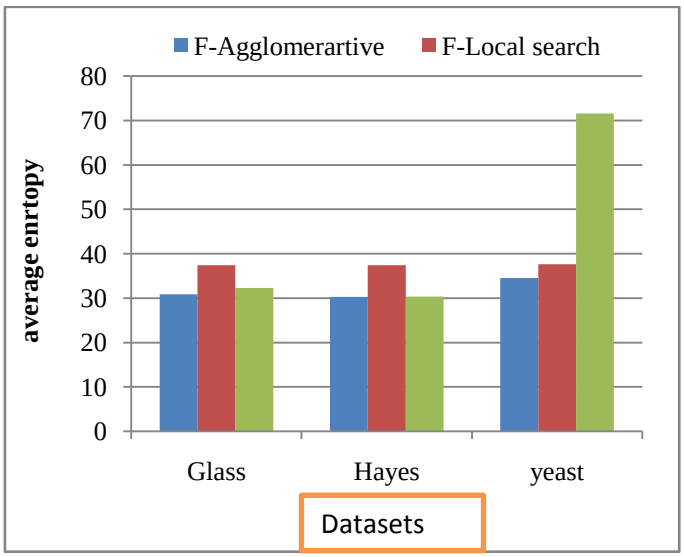

Fig.4. Datasets results using entropy

Another graph is plotted to represents the sum of average entropy for three different algorithm using the same algorithm. The results show that the agglomerative has the least value. Compared to other algorithms using different datasets.

Results on the magic sets

From the above table the lowest running time is F-agglomerative algorithm. Compared to other two algorithms F-agglomerative is efficient one.

VI. Conclusion

Proposed system has defined data fragments and proposed a new fragment-based clustering aggregation approach. This approach is based on the proposition that in an optimal partition each fragment is a subset of a cluster. Existing point-based algorithms can be brought into our proposed approach. As the number of data fragments is usually far smaller than the number of data points, clustering aggregation based on data fragments is very likely to have a low error rate than directly on data points. To demonstrate the efficiency of the proposed approach, three new clustering aggregation algorithms are presented, namely-Agglomerative, F-Furthest, and F-Local Search (based on three existing point-based clustering aggregation algorithms ones). Experiments were on three public data sets. The results show that the three new algorithms outperform the original clustering aggregation algorithms in terms of running time without sacrificing effectiveness.

VII. Future Enhancement

In our future work, we aim to solve the following problems:

The numbers of fragments increases when there are missing values or points with unknown clusters in the original clustering results. This leads to a lower efficiency.

Clustering validity measurement technique such as Davies Bouldin Index technique is used. In this Davies Bouldin Index technique, we determine the most similar cluster among the clusters has same similarity. If the fragment has similar cluster more than one clusters means then this technique is calculated the most similar cluster. The Davies – Bouldin index is based on similarity measure of clusters (Rij) whose bases are the dispersion measure of a cluster (si) and the cluster dissimilarity measure (dij).

REFERENCES

[1]. Ou Wu, *Member, IEEE*, Weiming Hu, *Senior Member, IEEE*,Stephen J. Maybank, *Senior Member, IEEE*, Mingliang Zhu, and Bing Li,vol. 42, no 3, june* 2012

[2]. J. Wu, H. Xiong, and J. Chen, "A data distribution view of clustering algorithms," in *Encyclopedia of Data Warehousing and Mining*, vol. 1,J. Wang, Ed., 2nd ed. Hershey, PA: IGI Global, pp. 374–381,2008.

[3]. Gionis.A, Mannila.H, and Tsaparas.P, 'Clustering aggregation', *ACM Trans. Knowl. Discov. Data*, vol. 1, no. 1, pp. 1–30, Mar. 2007

[4]. Z. Zhou and W. Tang, "Cluster ensemble," *Knowledge-Based Systems*,vol. 19, no. 1, pp. 77–83, Mar. 2006.

[5]. Ailon.N, Charikar.M, and Newman. A, 'Aggregating Inconsistent Information: Ranking and Clustering', in *Proc. STOC*, New York, pp. 684–693,2005.

[6]. N. Nguyen and R. Caruana, "Consensus clustering," in *Proc. IEEE ICDM*,, pp. 607–612,2005.

[7]. M. Meil˘a, "Comparing clustering's—An axiomatic view," in *Proc. ICML*, pp. 577–584, 2005.

[8]. A. P. Topchy, M. H. C. Law, A. K. Jain, and A. L. Fred, "Analysis of consensus partition in cluster ensemble," in *Proc. IEEE ICDM*, pp. 225–232,2004.

[9]. X. Z. Fern and C. E. Brodley, "Random projection for high dimensional data clustering: A cluster ensemble approach," in *Proc. ICML*,pp. 186–93,2003.

[10]. Fred.A.L.N and Jain.A.K, 'Robust data clustering', in *Proc. IEEECVPR*, vol. 2, pp. II-128–II-133,2002.

[11]. D. R. Cox and P. A.W. Lewis, *The Statistical Analysis of Series of Events*. London: Methuen, 1966.

[12]. [Online].Available: http://archive.ics.uci.edu/ml/