Pawan et al. International Journal of Institutional & Industrial Research ISSN: 2456-1274, Vol. 1, Issue 3, September-December 2016, pp.6-10

GSA Optimized Agglomerative Clustering

Pawan[1], Parveen Khanchi[2]

[1]M.Tech Student, Department of ECE, BIMT, Chidana, India

[2]Head of the Department of ECE, BIMT, Chidana, India

Abstract: **In statistic and data mining, agglomerative clustering is well known for its efficiency in clustering large data sets. The aim is to group data points into clusters such that similar items are lumped together in the same cluster. In general, given a set of objects together with their attributes, the goal is to divide the objects into clusters such that objects lying in one cluster should be as close as possible to each other's (homogeneity) and objects lying in different clusters are further apart from each other. However, there exist some flaws in classical agglomerative clustering algorithm. According to the method, first, the algorithm is sensitive to selecting initial threshold level and on the other hand, the agglomerative clustering is NP hard problem in selecting the optimum threshold level so that maximum F-measure or correct assignment of data to right clusters can be obtained.**

In this paper, to solving the agglomerative clustering problem, we provide optimizing threshold level in clustering to decide the number of clusters, which in this algorithm we consider the issue of how to derive an optimization model to the maximum accuracy which is measured in terms of F-measure. We introduce the optimization algorithm named Gravitational Search Algorithm (GSA) to optimize k-means algorithm to guarantee the result of clustering is more accurate than clustering by basic clustering algorithms. F-measure is used to compare the performance of both algorithms.

Keywords: **Optimized Agglomerative Clustering (OAC and hierarchical clustering..**

1. Introduction

The history of extraction of patterns from data is centuries old. The earlier method which has been used is Bayes' theorem (1700s) and regression analysis (1800s). [1] In the field of computer technology, using the ever growing power of computers, we develop an essential tool for working with data. Such as, it is being able to work with increasing size of the datasets and complexity. And also an urgent need to further refine the automatic data processing, which has been aided by other discoveries in computer science, means that our ability for data collection storage and manipulation of data has been increased.As definition, Data mining or important part of Knowledge Discovery in Database (KDD), used to discover the most important information throughout the data, is a powerful new technology. Across a myriad variety of fields, data are being collected and of course, there is an urgent need to computational technology which is able to handle the challenges posed by these new types of data sets.

The field of Data mining grows up in order to extract useful information from the rapidly growing volumes of data. It scours information within the data that queries and reports can't effectively reveal.

This process contains a series of transformation steps, from data pre-processing to data mining results. [1]

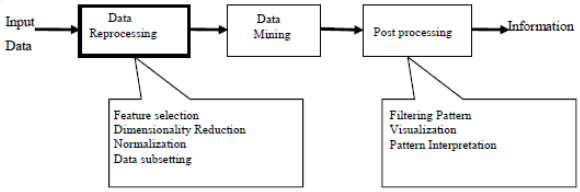

Figure 1.1 The overall steps of the process of Knowledge Discovery in Database (KDD)

There are challenges in traditional data analysis techniques and always new types of datasets. In order to cope with these new challenges, researchers have been developing more efficient and scalable tools that can more easily handle diverse types of data. In particular, data mining draws upon ideas such as:

1- Sampling ,estimating and hypothesis testing from statistic

2- Search algorithms, modelling techniques and learning theories from artificial intelligence, pattern recognition and machine learning.

And also data mining has been adopting from other areas, such as: optimization, evolutionary computing, information theory, signal processing, visualization and information retrieval. [6] Agglomerative clustering is also an important method for data mining clustering which is gaining attention these days. In Agglomerative clustering algorithm, a cut height parameter is required to determine the dissimilarity threshold at which clusters are allowed to be merged together. This parameter greatly influences the clustering accuracy, as measured by the Rand index, of the final clusters

produced. For instance, using a very high cut height or dissimilarity threshold would result in most data being included in one giant cluster since a weak measure of similarity is enforced during the merging process. So an optimum selection of threshold level is necessary. In our work we work towards optimising this threshold height.

2. Proposed Work

2.1 Hierarchical Clustering

The set of given data objects are partitioned in form of a tree like structure or nested clusters in hierarchical clustering. The hierarchical methods can be classified into two types.

• Agglomerative and

• Divisive

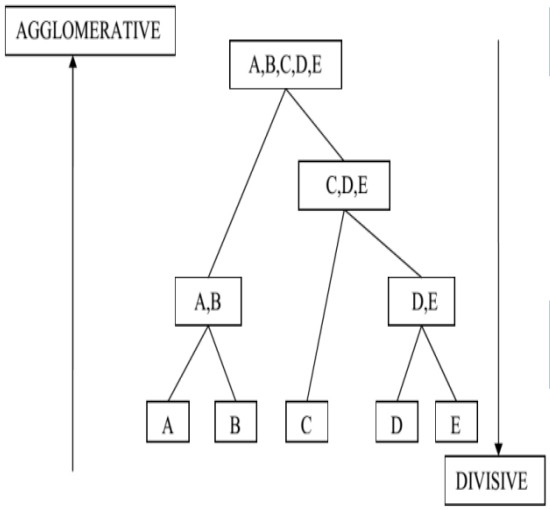

Figure 2.1: Agglomerative and Divisive clustering

In agglomerative method also known as bottom-up approach, each object forms a separate group. It successively merges the groups close to one another by checking the similarity function, until all the groups are merged into one, that's until the top most level of hierarchy is reached or until a termination condition holds. In divisive clustering also known as top-down approach, initially all the objects are grouped into a single cluster which can also be called as parent. In each successive iteration, a cluster is split up into smaller clusters, until eventually each object is in one cluster or until a termination condition holds.

2.1.1 Agglomerative Method

This method begins by treating each object as an individual cluster and then proceeds by merging two nearest clusters. The distance between any two clusters m and n is defined by a metric $D_{m,n}$. Metrics can be single-link, complete-link and group average etc. A general class of metrics was given by Lance and Williams [21]. If $D_{k,ij}$ be the distance between cluster k and the union of cluster i and cluster j, then:

$$D_{k,ij} = \alpha_i D_{k,i} + \alpha_j D_{k,j} + \beta D_{i,j} + \gamma |D_{k,i} - D_{k,j}|$$

The agglomerative method is as follows:

• Consider each object to be an atomic cluster. The (n x n) distance matrix represents the distance between all possible pairs of clusters.

• Find the smallest element in the matrix. This corresponds to the pair of clusters that are most similar. Merge these two clusters, say m and n, together.

• Measure the distances between the newly formed cluster and the other remaining clusters using a distance function. Delete the row and column of m and overwrite row and column of cluster n with the new values.

• If the current number of clusters is more than k then go to step 2; otherwise stop.

The merging process can continue until all the objects are in one cluster. The advantages of hierarchical methods are that they are easy to implement computationally. They are able to tackle larger datasets than the k-medoids method and we can run the algorithm without providing the input k (the number of clusters to be formed). The drawbacks of agglomerative method are:

• The algorithm has O(n3) time complexity. Even though the order of the distance matrix decreases with each iteration, the cost of Step2 on iteration k is O((n - k)2), and we are guaranteed (n - k) iterations before we get to k;

• The clusters produced are heavily dependent on the metric Di, j. Different metrics can produce different clusters. For instance, the complete-link metric tends to produce spherical clusters, whereas the single-link metric produces elongated clusters [21].

2.2 Problem Description

The agglomerative clustering is the unsupervised approach which makes cluster of same kind of data to make data analysis easier. This clustering algorithm is described in previous chapter in section 3.1. We are targeting the bottom approach of agglomerative clustering in which every node starts with its own cluster and based on similarity value, other nodes are combined with its cluster. This similarity measure is calculated in terms of Euclidean distance between nodes. These steps are repeated for all nodes and clusters are dependent upon the minimum distance between nodes. Once all nodes are clustered depending upon their distance, number of clusters are selected using a threshold level. Classical agglomerative

clustering approach selects the threshold level to divide the cluster tree into specific clusters numbers by using Euclidean distance. The accuracy of this clustering algorithm depends upon the threshold level. For example if actual cluster number is 4 but threshold level divides the cluster tree into 5 clusters then false clustering will be high. So to improve the accuracy and F-measure, it should be set at optimal position. To fulfill this purpose evolutionary optimization algorithms have been used which uses minimum distance concept to cluster. Particle swarm optimization, bacterial foraging optimization, genetic algorithm etc are used earlier for data clustering purpose. These evolutionary techniques can be categorized as global optimization and local optimization techniques. As all such kind of algorithms look for local minimum position for which cost function has minimum value but local optimization algorithms like genetic algorithms (GA), ant colony optimization (ACO), particle swarm optimization (PSO) etc, sometimes jumps over the local minimum point whereas global optimization techniques like gravitational search algorithm (GSA) which came into existence in 2013 has no such issue, it checks for all iteration values for local minimum point. In case of multi objective functions, global optimization algorithms perform well. But these suffers from a drawback of speed. Iteration speeds of global techniques are less than local. So these take long time to process.

So in our work we have used a gravitational search algorithm technique for data clustering which used agglomerative clustering objective function for data clustering.

2.3 Process

We are working on data clustering using optimization algorithm based on moments of celestial bodies, which are totally separate fields. The terms used in optimization algorithm have their technical counterpart as per applications. In our application we are optimizing the position of cluster head amongst the data so that rest data settle down near to those cluster heads and get classified. These cluster heads are agents in GSA. Their positions are changing in algorithm but in their technical counterpart positions of cluster heads are updating. Table 2.1 shows the technical counter part of evolutionary algorithm for our application.

Table 2.1: Technical and evolutionary techniques equivalent terms

	Evolutionary algorithm Term	Technical Term
1	agents	Cutoff level (threshold value in cluster tree)
2	Search dimension	Total number of co-ordinates of all cluster heads
3	Position of agents	Cutoff level value

The performance of the proposed algorithm is compared with the classical agglomerative algorithm and F-measure is used as the comparison parameter. The objective function is the soul of optimization algorithms.

Pseudo steps for the algorithm are as:

Load input clustering data
Find out the number of attributes in the data
For ii=1: number of attributes
Inputattribute=inputdata(attributes)
Choose number of cluster heads
Initialize all variables and steps in GSA optimization algorithm
For 1: maximum iterations
Pass each initial agent positions to objective function
Calculate the minimum F-measure using equation:

$$\mathrm{F-measure} = \frac{2 \times \mathrm{Precision} \times \mathrm{recall}}{\mathrm{Precision} + \mathrm{recall}}$$

Update the agents' positions as described in section 3.2
Call objective function again

Take out the minimum value of the fitness function and that agents' position will serve as best position till now
Update the velocity of agents particles as in section 3.2

Iterations end
Assign the data of attribute to corresponding cluster which will be nearest to them
Repeat all above steps for all attributes in the data
Calculate the f-measure
end

In our work we have calculated the F-measure for each new position of cluster head and all other nodes.

3. Results

This work classifies the data using optimized agglomerative clustering approach. It is unsupervised classification algorithm. The evolutionary algorithms which gravitational search algorithm (GSA) is implemented in MATLAB tool.To evaluate the performance of proposed algorithm, six datasets have been used. These six datasets are iris, thyroid, wine, Contraceptive Method Choice (CMC), liver disorder and glass, are collected from the weblink ftp://ftp.ics.uci.edu./pub/machine-learning-databases/. The six datasets used in this paper is described in Table 3.1.

Table 3.1: Input Datasets

	Dataset Name	attributes	classes	instances
1	Liver disorder	7	2	345
2	Iris	4	3	150
3	Wine	13	3	178
4	Glass	9	6	214
5	Thyroid	5	6	215
6	Contraceptive Method Choice (CMC)	10	3	1473

We have used GSA to optimize the clustering algorithm and results have been compared with classical agglomerative clustering in terms of F-measure which is discussed in chapter 4. Since each data sets have various number of attributes so data clustering will be executed for each attribute separately. Combine results of all attributes are used to calculate the minimum, maximum, standard deviation and average of F-measure are calculated. F-measure nearest to 1 guarantees good data clustering. Optimization in data clustering used here provides different F-measure in every execution because cut off level is initialized randomly every time. For each attributes total minimum distance by optimization in each execution is plotted. A good optimization is guaranteed when total minimum F-measure is increased in each iteration and finally settles to a maximum value and stick to that till last iteration. An example for it for iris data is shown in figure 3.1.

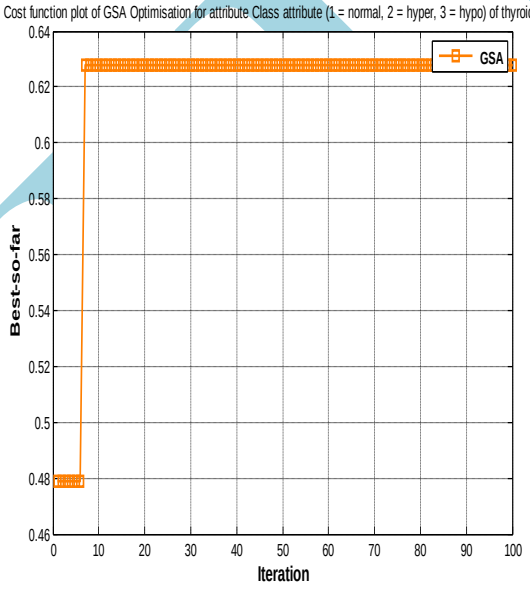

Figure 3.1(a): fitness function value plotted for 1st attribute

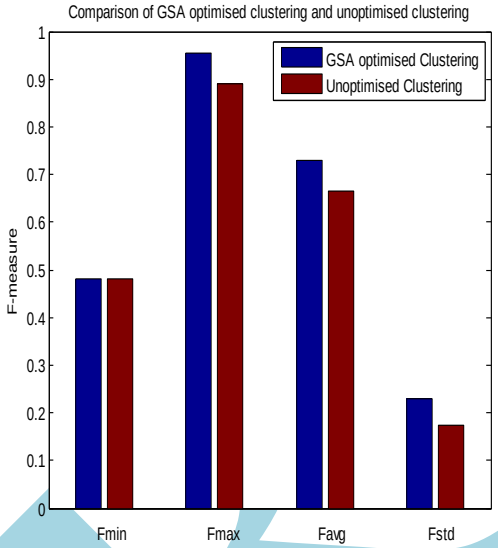

figure 3.2: comparison of F measure for proposed algorithm and conventional agglomerative clustering

In above figure each data set is settled to a maximum F-measure after some number of iterations. For this data set clustering is done for petal length and width and sepal length and width which are four attributes of iris data. Same process is done for all types of data loaded.

The comparison with conventional agglomerative clustering in terms of F -measure is shown in figure 3.2 along with dendogram of clusters in figure 3.3. Processing is done all data mentioned in table 3.1 and results in terms of F-measure are shown. Higher value of f-measure is indication of good clustering.

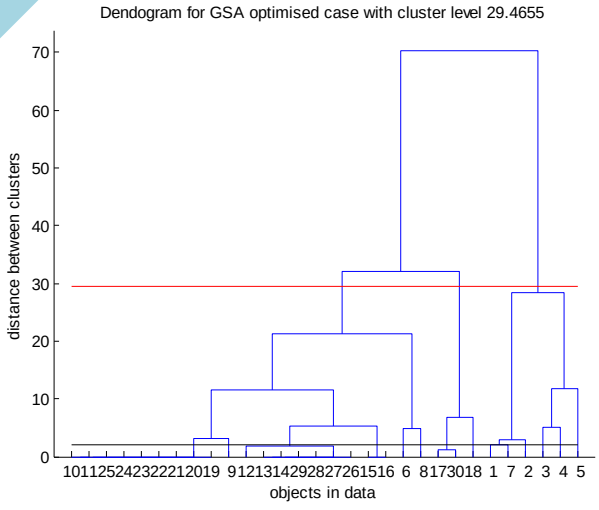

figure 3.3: Dendogram plot after proposed clustering approach

A comparison graph of maximum F -measure is shown in figure 3.4 below.

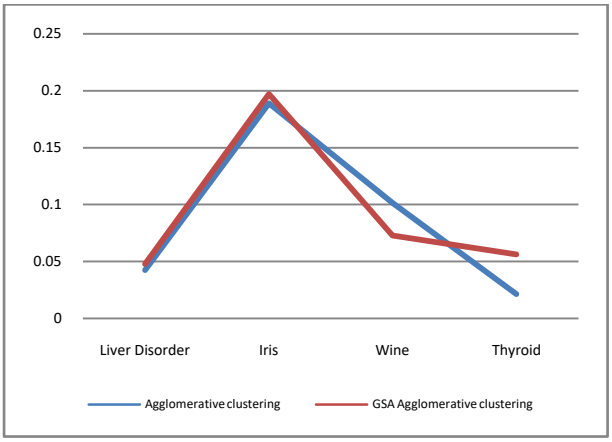

Figure 3.4: maximum F-measure comparison of F – measure

Above figure shows that proposed data clustering algorithm have higher F-measure for most of the data.

4. Conclusion

In this work, the problem was to solve the agglomerative clustering problem by introducing a clustering technique – Gravitational search Algorithm in agglomerative clustering which is an optimization algorithm to tune the cutoff level of clustering tree. The problem in clustering as we notice is because of the cut off level in aglomerative method which is based on the Euclidean distance between nodes. Sometimes we have a poor clustering (some clusters don't have any member). The goal is clustering in the best behaviour, which should be to group similar data points as much as possible. But with classical agglomerative clustering this rarely is the case.

To optimize the clustering, we propose an algorithm. In this data clustering method concept of maximizing the F measure between every cluster head and other data points. Gravitational search algorithm (GSA) is used. Cut off position optimized by GSA serves as input to the clustering algorithm. Position of level is initialized randomly in this also but later on it changes the position as per tuning method of respective optimization technique. Comparison of results with classical agglomerative algorithm is done in terms of F-measure. Standard deviation comparison if it, as in figure 3.4 shows improvement by proposed algorithm.

REFERENCES

[1]. Kriti Srivastava, R. Shah, D. Valia, and H. Swaminarayan,"Data Mining Using Hierarchical Agglomerative Clustering Algorithm in Distributed Cloud Computing Environment", *International Journal of Computer Theory and Engineering, Vol. 5, No. 3, June 2013*

[2]. Marek Lipczak, Evangelos Milios,"Agglomerative Genetic Algorithm for Clustering in Social Networks", Genetic and Evolutionary Computation Conference, GECCO 2009, Proceedings, Montreal, Québec, Canada, July 8-12, 2009.

[3]. Singaravelu.S, A.Sherin and S.Savitha,"Agglomerative Fuzzy K-Means Clustering Algorithm",A Journal of Nehru Arts and Science College (NASC), Vol 1 (2013)

[4]. R. Krishnamoorthy and S. SreedharKumar, "New optimized agglomerative clustering algorithm using multilevel threshold for finding optimum number of clusters on large data set," *Emerging Trends in Science, Engineering and Technology (INCOSET), 2012 International Conference on*, Tiruchirappalli, Tamilnadu, India, 2012, pp. 121-129.

[5]. Jake M Drew and Tyler Moore,"Optimized combined – clustering methods for finding replicated criminal websites", *EURASIP Journal on Information Security* (2014).

[6]. Sanjay Tiwari, Mahainder Kumar Rao," Optimization In Association Rule Mining Using Distance Weight Vector And Genetic Algorithm" International Journal of Advanced Technology & Engineering Research (IJATER), Volume 4, Issue 1, Jan. 2014.

[7]. Poonam Sehrawat, Manju," Association Rule Mining Using Particle Swarm Optimization", International Journal of Innovations & Advancement in Computer Science, Volume 2, Issue 1 January 2014

[8]. R.Jensi and G.Wiselin Jiji," Hybrid Data Clustering Approach Using K-Means And Flower Pollination Algorithm", Advanced Computational Intelligence: An International Journal (ACII), Vol.2, No.2, April 2015

[9]. Khalid Raza," Clustering analysis of cancerous microarray data", Journal of Chemical and Pharmaceutical Research, 2014, 6(9)

[10]. P. Ramachandran, N.Girija," Early Detection and Prevention of Cancer using Data Mining Techniques", International Journal of Computer Applications, Volume 97– No.13, July 2014.

[11]. Hlaudi Daniel Masethe, Mosima Anna Masethe," Prediction of Heart Disease using Classification Algorithms", Proceedings of the World Congress on Engineering and Computer Science 2014 Vol II

[12]. P.Kalyani," Medical Data Set Analysis Ñ A Enhanced Clustering Approach" International Journal of Latest Research in Science and Technology, Volume 3, Issue 1: Page No.102-105 ,January-February 2014

[13]. Ibrahim M. El-Hasnony, Hazem M. El Bakry, Ahmed A. Saleh,"Data Mining Techniques for Medical Applications: A Survey", Mathematical Methods in Science and Mechanics,2014

[14]. Sundararajan S, Dr. Karthikeyan S," An Hybrid Technique for Data Clustering Using Genetic Algorithm with Particle Swarm Optimization", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 4, Issue 12, December 2014

[15]. Sundararajan S., And Karthikeyan S," An Efficient Hybrid Approach For Data Clustering Using Dynamic K-Means Algorithm And Firefly Algorithm", ARPN Journal Of Engineering And Applied Sciences, Vol. 9, No. 8, August 2014

[16]. Sandeep Rana, Sanjay Jasola, Rajesh Kumar," A hybrid sequential approach for data clustering using K-Means and particle swarm optimization algorithm", International Journal of Engineering, Science and Technology, Vol. 2, No. 6, 2010

[17]. Sandeep U. Mane, Pankaj G. Gaikwad," Hybrid Particle Swarm Optimization (HPSO) for Data Clustering", International Journal of Computer Applications (0975 8887) Volume 97 - No. 19, July 2014

[18]. T. Niknam, M. Nayeripour and B.Bahmani Firouzi," Application of a New Hybrid optimization Algorithm on Cluster Analysis", International Journal of Computer, Electrical, Automation, Control and Information Engineering Vol:2, No:10, 2008

[19]. Amin Rostami and Maryam Lashkari," Extended Pso Algorithm For Improvement Problems K-Means Clustering Algorithm", International Journal of Managing Information Technology (IJMIT) Vol.6, No.3, August 2014

[20]. M. Bhanu Sridhar1, Y. Srinivas2, M. H. M. Krishna Prasad,"Software Reuse in Cardiology Related Medical Database Using K-Means Clustering Technique", Journal of Software Engineering and Applications, 2012.

