AN ENSEMBLE MODEL FOR MEDICAL DATA USING DATAMINING ALGORITHMS

S Mythili[1], K Geethalakshmi[2]

Abstract

Clustering is the process of grouping data into clusters, where objects within each cluster have high similarity, but are dissimilar to the objects in other clusters. The K-means algorithm is used for clustering large sets of data. The accuracy of the K-means depends upon the selection of Centroids. The execution of the standard K-means algorithm need to reassign the data points a number of times, during every iteration of the loop. The hybrid approach that includes both K means algorithm and genetic algorithm yields good result in the process of clustering. In this study, we proposed an implementation of genetic algorithm which we investigate the quality of clustering technique compared with standard K-means clustering algorithm.

***Keywords** : Mutation, Genetic algorithm, K-Means Clustering, Cross Over*

I. Introduction

A cluster is an ordered list of objects, which have some common characteristics. Cluster analysis organizes data by abstracting underlying structure either as a grouping of individuals or as a hierarchy of groups [1].The distance between two clusters involves some or all elements of the two clusters. A similarity measure can be used to represent the similarity between the documents. Typical similarity generates values of 0 for documents exhibiting no agreement among the assigned indexed terms. The lowest possible input value of similarity required to join two objects in one cluster. Similarity between objects calculated by the function, represented in the form of a matrix is called a similarity matrix. The dissimilarity coefficient of two clusters is defined to be the distance between them. If the value of dissimilarity coefficient is smaller, the two clusters are more similar. First document or object of a cluster is defined as the initiator of that cluster i.e. every incoming object's similarity is compared with the initiator. The initiator is called the cluster seed.

1.1 Basic Clustering Steps:

Step1 : Preprocessing and feature selection

- An appropriate feature is chosen
- Preprocessing and feature extraction on the data items are done to measure the chosen feature set
- It requires a good domain knowledge and data analysis

Step 2 : Similarity measure

- Objects are grouped into various clusters
- The similar objects will be grouped in the same cluster
- The dissimilar objects are grouped in different clusters
- This is a function that takes two sets of data items as input
- The output will be a similarity measure between the two data items

[1] Assistant Professor, Department of Computer Applications, PSGR Krishnammal College for Women, Coimbatore.

[2] Head of the Department, Department of Computer Applications, PSGR Krishnammal College for Women, Coimbatore

Clustering Algorithms:

Divisive techniques are less common, and we will mention only one example before focusing on agglomerative techniques.

➢ **Simple Divisive Algorithm (Minimum Spanning Tree (MST))**

1) Compute a minimum spanning tree for the proximity graph.

2) Create a new cluster by breaking the link corresponding to the largest distance (smallest similarity).

3) Repeat step 2 until only singleton clusters remain.

This approach is the divisive version of the "single link" agglomerative technique that we will see shortly.

➢ **Agglomerative Hierarchical Clustering Algorithm**

Many hierarchical agglomerative techniques can be expressed by the following algorithm, which is known as the Lance-Williams algorithm.

➢ **Basic Agglomerative Hierarchical Clustering Algorithm**

1) Compute the proximity graph, if necessary. (Sometimes the proximity graph is all that is available.)

2) Merge the closest (most similar) two clusters.

3) Update the proximity matrix to reflect the proximity between the new cluster and the original clusters.

4) Repeat steps 3 and 4 until only a single cluster remains.

The key step of the previous algorithm is the calculation of the proximity between two clusters, and this is where the various agglomerative hierarchical techniques differ. Any of the cluster proximities that we discuss in this section can be viewed as a choice of different parameters (in the Lance-Williams formula) for the proximity between clusters Q and R, where R is formed by merging clusters A and B[8].

II. K– Means Clustering Algorithm

K-Means (KM) is one of the most popular methods used in data analysis due to its good computational performance. However, it is well known that KM might converge to a local optimum, and its result depends on the initialization process, which randomly generates the initial clustering. In other words, different runs of KM on the same input data might produce different results [2].

➢ **Basic Algorithm**

The K-means clustering technique is very simple and we immediately begin with a description of the basic algorithm. We elaborate in the following sections. Basic K-means Algorithm for finding K clusters[1].

1. Select K points as the initial centroids.

2. Assign all points to the closest centroid.

3. Recompute the centroid of each cluster.

4. Repeat steps 2 and 3 until the centroids don't change.

Figure (a) shows the case when the cluster centers coincide with the circle centers. This is a global minimum. Figure (b) shows local minima.

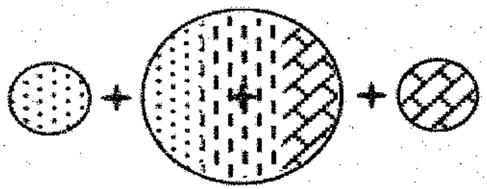

Figure (a). A globally minimal clustering solution

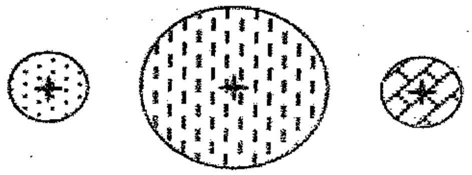

Figure (b). A locally minimal clustering solution.

The idea of K-Means is to choose random cluster centers, one for each cluster. These centers are preferred to be as far as possible from each other. Starting points affect the clustering process and results [3]

After that, each point will be taken into consideration to calculate similarity with all cluster centers through a distance measure, and it will be assigned to the most similar cluster, the nearest cluster center. When this assignment process is over, a new center will be calculated for each cluster using the points in it[9][10]. For each cluster, the mean value will be calculated for the coordinates of all the points in that cluster and set as the coordinates of the new center.

Once we have these *k* new centroids or center points, the assignment process must start over. As a result of this loop we may notice that the *k* centroids change their locations step by step until no more changes are made.

When the centroids do not move any more or no more errors exist in the clusters, we call the clustering has reached a minima. Finally, this algorithm aims at minimizing an objective function, which is in this case a squared error function.

Time and Space Complexity

Since only the vectors are stored, the space requirements are basically O(mn), where m is the number of points and n is the number of attributes. The time requirements are O(I*K*m*n), where I is the number of iterations required for convergence. I is typically small (5-10) and can be easily bounded as most changes occur in the first few iterations [7]. Thus, K- means is linear in m, the number of points, and is efficient, as well as simple, as long as the number of clusters is significantly less than m.

Choosing initial centroids

Choosing the proper initial centroids is the key step of the basic K-means procedure. It is easy and efficient to choose initial centroids randomly, but the results are often poor[5].

We start with a very simple example of three clusters and 16 points. Figure (a) indicates the "natural" clustering those results when the initial centroids are "well" distributed. Figure (b) indicates a "less natural" clustering that happens when the initial centroids are poorly chosen.

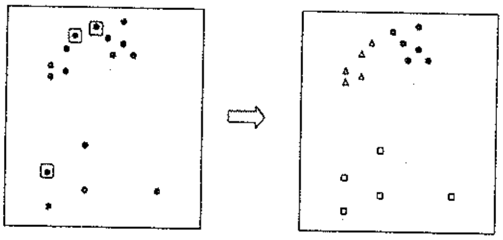

Figure (a): Good starting centroids and a 'natural' clustering

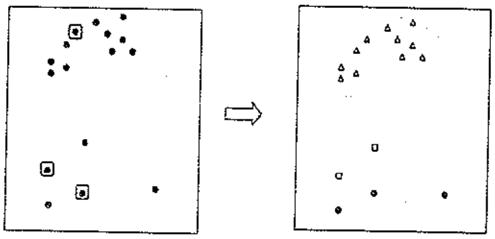

Figure (b): Bad starting centroids and a "less natural" clustering.

Drawback of K-Means Clustering

One drawback of KM is that it is sensitive to the initially selected points, and so it does not always produce the same output.

III. Genetic Algorithms

Genetic Algorithms are used to determine the best initialization of clusters as well as optimization of initial parameters. Genetic Algorithms attempt to incorporate the ideas of natural evolution [4].In general they start with an initial population, and then a new population is created based on the notion of survival of the fittest. Typically fitness is the measure for how good this population is and can be calculated depending on the nature of the application, where a distance measure is the most common. Then a process called crossover is done over the new population where substrings from selected pairs are swapped.

Goldberg's Pseudo-code of Genetic Algorithm

```
Begin
1.    T=0
2.    Initialize population P(t)
3.    Compute fitness P(t)
4.    T = t+1
5.    If termination criterion achieved go to step 10
6.    Select P(t) from P(t-1)
7.    Crossover P(t)
8.    Mutate P(t)
9.    Go to step 3
10.   Output best and stop
End
```

Where't' represents the generation number, and P stands for population. The first population is initialized by coding it into a specific type of representation then assigned to a cluster. Fitness is calculated in the evaluation step. Selection process chooses individuals from population for the process of crossover. Recombination (or crossover) is done by exchanging a part (or some parts) between the chosen individuals, which is dependent on the type of crossover (Single point, Two points, Uniform, etc)[6]. Mutation is done by replacing few points among randomly chosen individuals. Then fitness has to be recalculated to be the basis for the next cycle.

Advanced Genetic Algorithm

Yan Wang et al developed an advanced genetic algorithm for complex value encoding[10].The algorithm is as follows:

1. In the beginning, two populations with the size of N chromosomes $(\rho_1, \rho_2 \dots \dots \rho_m)$ and $(\theta_1, \theta_2 \dots \dots \theta_m)$ were created randomly by system, which and indicate the modulus and angle of complex of allele respectively. The chromosomes' length is m.

   $(\rho_k \epsilon \left[0, \frac{b_k - a_k}{2}\right], \theta_k \epsilon [0, 2\pi], K = 1,2 \dots \dots m)$. The $2 * N$ chromosomes contained the initial population with N chromosomes. Then the variable x_k which corresponded by allele can be expressed as follows:

   $$x_k = \rho_k cos\theta_k + \frac{a_k + b_k}{2}$$

   Where $k = 1, 2 \dots m$

2. Evaluates the fitness of each individual in that population;

3. If the pre-specified the termination criteria are reached, then stop;
4. Select the best-fit individuals for reproduction;
5. Breed new individuals through crossover and mutation operations to give birth to offspring; Go back to step 2.

The proposed improved genetic algorithm is developed with simple modification of Yan Wang et al algorithm. Then the new algorithm is combines with K-Means and makes the selection process of centroids.

IV. Proposed Methodology

Initial starting points generated by K-means make the clustering results reach the local optima. The better results of K-means clustering can be achieved by computing more than one time. However, it is difficult to decide the computation limit, which can give the better result. In this paper, we propose a new approach to optimize the initial centroids for K-means. It utilizes all the clustering results of K-means in certain times.Then, the result by combining with Hierarchical algorithm in order to determine the initial centroids for K-means. The experimental results show how effective the proposed method to improve the clustering results by K-means. The following are the advantages of hybrid approach (combination of K-Means and genetic algorithms).

Advantages of hybrid approach

1. Embedded flexibility regarding a level of granularity.
2. Easy of handling of any forms of similarity or distance.
3. Consequence applicability to any attributes types.
4. More versatile.
5. It converges fast given a good initialization.
6. It is robust to noisy data.
7. It can accept the desired number of clusters as input.

Thus, the proposed advanced genetic algorithm initiates the process of K-Means. This algorithm accuracy is thoroughly checked with different datasets. The experimental analyses are discussed in next chapter.

V. Experimental Results

The proposed advanced genetic algorithm is executed with different data sets. Then the efficiency and accuracy of the advanced genetic algorithm is also compared with existing algorithms such as K means. K Means algorithm is sensitive to the initially selected points. Hence it does not always produce the same output. So, we are selecting the initial points randomly. To reduce the effect of randomness, we have to run the algorithm many times before taking an average values for all runs, or at least take the median value. So, the K-Means algorithm is executed for 50 times and the readings are noted. Similarly the proposed and existing genetic algorithms were also executed for 50 runs and results were noted. Finally, the average value is calculated for each algorithm by using different datasets.The datasets used for this work are Bupa and Breast Cancer datasets.

ALGORITHM	DATA SETS			
	DATASET 1 (BUPA)		DATASET 2 (BREAST CANCER)	
	Time	Average Error Rate	Time	Average Error Rate
K - Means	0.1657	16.9031	0.0718	26.8303
Genetic	0.4895	15.0094	0.314	26.1343

Table 1: Result comparison between different approaches

After conducted experiments on various datasets, it is clearly depicted that the proposed genetic algorithm works. The various analyses can be seen so far. From that we can conclude that the average error rate for k means is higher than other algorithms. The main aim for developing genetic algorithm is to improve the accuracy of K Means in the process of selecting centre points of the clusters. As per our results, we are getting good accuracy during execution of genetic algorithm for initialization process of K Means. Whatever K means is very fast in computation, the error rate is high. Due to the help of genetic algorithm with K means, the average error rate is reduced gradually.

VI. CONCLUSION

The K- means algorithm is widely used for clustering large sets of data. But the standard algorithm does not guarantees good results. The accuracy of the K-means depends upon the selection of centroids. Moreover, the execution of the standard K-means algorithm need to reassign the data points a number of times, during every iteration of the loop.

The genetic algorithm improves the accuracy and efficiency of the K means initialization process. Our experimental evaluation scheme was used to provide a common base of performance assessment and comparison with other methods. The proposed genetic algorithm was then compared with existing K means algorithm. The results of this comparison show that the GA can achieve better results for the solutions in a faster time from the execution of algorithm on the four data sets; we find that improved algorithm work well and yield meaningful and good results in the terms of clustering techniques.

The hybrid approach that includes both K means algorithm and genetic algorithm yields good result in the process of clustering. However, the experimental results shows that accuracy in clustering process, the execution time is little more than standard K means algorithm.

REFERENCES

1. Anna D. Peterson, Arka P. Ghosh and Ranjan Maitra 2007..A systematic evaluation of different methods for initializing the K-means clustering algorithm, IEEE transactions on knowledge and data engineering.12[1];234-241.

2. Ayhan Demiriz, Bennett.K, 1999.Semi-Supervised Clustering Using Genetic Algorithms", Artificial Neural Networks in Engineering (ANNIE), 21[2]:34-39

3. Bashar Al-Shboul, and Sung-Hyon Myaeng" Initializing K-Means using Genetic Algorithms", World Academy of Science, Engineering and Technology 54 2009.

4. Cheng Min-Yuan and Huang Kuo-Yu "K-means clustering and Chaos Genetic Algorithm for Nonlinear Optimization", Information and Computational Technology.

5. Dharmendra K Roy and Lokesh K Sharma" Genetic K-Means clustering algorithm for mixed numeric and categorical datasets", International Journal of Artificial Intelligence and applications(IJAIA), Vol.1, No.2, April2010

6. Fang-Xiang Wu1, Anthony J. Kusalik and W. J. Zhang" Genetic Weighted K-means for Large-Scale Clustering Problems". 7[2]: 443-449.

7. First A. S.Siva Sathya , Second B. Philomina Simon, Member IACSIT ",A Document Retrieval System with Combination Terms Using Genetic Algorithm", International Journal of Computer and Electrical Engineering, Vol. 2, No. 1, February, 2010 1793-8163.

8. Hao-jun SUN, Lang-huan XIONG, Genetic Algorithm-based High-dimensional Data Clustering Technique, 2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery, 2009.

9. Ian H. Witten; Eibe Frank (2005). "Data Mining: Practical machine learning tools and techniques, 2nd Edition". Morgan Kaufmann, San Francisco.

10. Indrajit Saha and Anirban Mukhopadhyay "Genetic Algorithm and Simulated Annealing based Approaches to Categorical Data Clustering", Proceedings of the International Multi Conference of Engineers and Computer Scientists 2008 Vol I 19-21 March, 2008.

Author's Biography

Mythili S, Assistant Professor, Department of Computer Applications in PSGR Krishnammal College for Women, Coimbatore. She received her B.Sc., in 2002 and MCA in 2005 from Periyar University, Coimbatore. She obtained her M.Phil. in the area of Data mining from Karpagam University, Coimbatore in 2013. Her research interest lies in the area of Data mining. She has around 5 years of Academic and 1 year of Industry experience.

Geethalakshmi K, Assistant Professor and Head of the Department, BCA, PSGR Krishnammal College for Women, Coimbatore. She completed her MCA and M.Phil..,. Her research interest lies in the area of Image Processing. She has around 8 years of Academic experience.