Scientific Journal Impact Factor (SJIF) : 3.518
International Journal of Modern Trends in Engineering and Research
ISSN (ONLINE) : 2349 - 9745 ISSN (PRINT) : 2393 - 8161

A Hybrid Model for Medical Data Using Machine Learning Approaches

Dr.P.Sumathi [1] and Dr.V.Kathiresan[2]

Dept.of.Computer Science [1], Government Arts College, Coimbatore
Dept.of.Computer Applications [2], SNS College of Arts & Science, Cbe

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 using the Medical data set.

Keywords— Cross Over, Genetic Algorithm, K-Means Clustering, Mutation.

I. INTRODUCTION

The data mining techniques for healthcare management is becoming increasingly popular due to several reasons. One important factor is the huge amounts of data generated by healthcare transactions are too complex and voluminous to be processed and analyzed by traditional methods. The data consists of patient details, hospital resources, disease diagnosis details and equipment details. The large of data is a key resource to be processed and analyzed for knowledge extraction that enables support for cost-savings and can improve decision-making by discovering patterns and trends in large amounts of complex data. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain.

Hidden patterns or similarities and relationships in medical data like common treatment procedures provided to similar diseases or symptoms patterns that have occurred to previous patients. The problem of extracting hidden patterns in medial domain is becoming increasingly relevant today, as the records of patient's clinical trials and other attributes are available electronically. The purpose of finding patterns in medical databases is to identify those patients which share common attributes and hence constitute same risk group. These patterns, if accurately discovered, can be utilized for clinical diagnosis.

Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. Hidden patterns or similarities and relationships in medical data like common treatment procedures provided to similar diseases or symptoms patterns that have occurred to previous patients.

Three types of data mining operations can be performed on medical data. They are

1. Diagnosis: To recognize and classify patterns in multivariate patient attributes. Can be used to perform automated analysis of pathological signals (ECG, EEG, EMG) and medical images (mammograms, ultrasound, X-ray, CT and MRI).
2. Therapy: Selection from available treatment methods; based on effectiveness and suitability to patient. Examples include selecting best treatment plans using patient model

3. Prognosis: Predict future outcomes based on previous experience and present conditions. Examples include survival analysis for AIDS patients, determine cardiac surgical risk and breast cancer prognosis.

All the above operations can be efficiently performed using cluster analysis and classification. Clustering is an unsupervised learning technique that deals with finding a structure in a collection of unlabeled data. It is the process of organizing objects into groups whose members are similar in some 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. 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.

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 is given below [1].

Algorithm 1: K-Means Algorithm

1. Select K points as the initial centroids.
2. Assign all points to the closest centroid.
3. Re-compute the centroid of each cluster.
4. Repeat steps 2 and 3 until the centroids don't change.

The below given algorithm is fine in producing results. The major problem of this is that it produces different clusters for different sets of values of the initial centroids. Quality of the final clusters depends on the selection of the initial centroids. The K-Means algorithm is computationally expensive and requires time proportional to the product of the number of data items, number of clusters and the number of iterations.

Algorithm2: The K-Means Clustering algorithm

Input:

D = {d1, d2,.....,dn} //set of *n* data items.

k // Number of desired clusters

Output:

A set of *k* clusters.

Steps:

1. Arbitrarily choose *k* data-items from D as initial centroids;
2. Repeat

Assign each item d_i to the cluster which has the closest centroid;

Calculate new mean for each cluster; until convergence criteria is met.

Figure 1(a) shows the case when the cluster centers coincide with the circle centers. This is a global minimum. Figure 1(b) shows local minima.

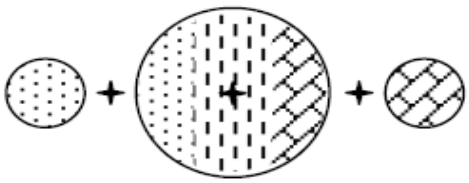

Fig.1 (a) A globally minimal clustering solution

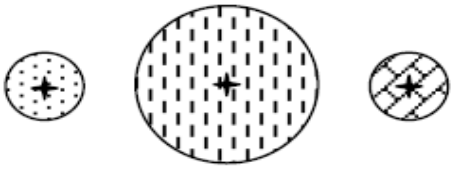

Fig. 1(b). A locally minimal clustering solution

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. Initial points affect the clustering procedure 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 procedure must begin 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 minimize the objective 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 amount of clusters is considerably 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 2(a) indicates the "natural" clustering those results when the initial centroids are "well" distributed. Figure 2(b) indicates a "less natural" clustering that happens when the initial centroids are poorly chosen.

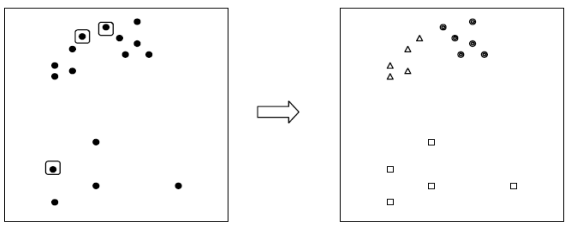

Fig. 2(a) Good starting centroids and a 'natural' clustering.

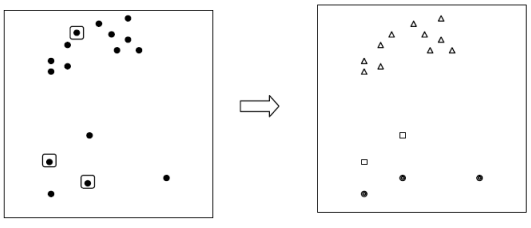

Fig. 2(b) Bad starting centroids and a "less Natural" clustering

III. GENETIC ALGORITHM

Genetic Algorithms (GA) 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 [10]. Then a process called crossover is done over the new population where substrings from selected pairs are swapped.

Algorithm 3: 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.

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).

Algorithm 4: Advanced Genetic Algorithm

Yan Wang et al developed an advanced genetic algorithm for complex value encoding [11]. The proposed improved genetic algorithm is developed with simple modification of Yan Wang et al algorithm . Then the new algorithm is combined with K-Means and makes the selection process of centroids. The algorithm is as follows:

1. In the beginning, two populations with the size of N chromosomes $(\rho_1, \rho_2 \ldots \ldots \rho_m)$ and $(\theta_1, \theta_2 \ldots \ldots \theta_m)$ were created randomly by system, which and indicate the modulus and angle of

complex of allele respectively. The chromosomes' length is m. $(\rho_k \in \left[0, \frac{b_k - a_k}{2}\right], \theta_k \in [0, 2\pi], K - 1, 2 \ldots \ldots m)$. 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 $k = 1, 2 \ldots m$

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

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 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 as noted in the table-I. Then the efficiency in terms of time and accuracy of the advanced genetic algorithm is also compared with existing algorithm such as K-Means as given in the table-II and in the fig.3 and fig.4. 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 .

Finally, the average value is calculated for each algorithm by using different datasets. The datasets used for this work are Diabetes, Bupa and Breast Cancer datasets. Three real-life datasets were obtained from UCI Machine learning runs and results were noted repository. (http://archive.ics.uci.edu/ml/datasets.html).

TABLE 1

CHARACTERISTICS OF DATA SETS USED DURING EXPERIMENTATION

S.No	Dataset	Type	No. of Instances	No. of Dimensions	No. of Classes
1	Bupa	Multivariate	150	4	2
2	Breast Cancer	Multivariate	699	32	2
3	Diabetes	Multivariate	768	8	2

TABLE II

RESULTS OF THE PROPOSED METHODOLOGY

ALGORITHM	DATA SETS					
	Bupa		Breast Cancer		Diabetes	
	Execution Time (sec)	Accuracy (%)	Execution Time (sec)	Accuracy (%)	Execution Time (sec)	Accuracy (%)
K - Means Algorithm	0.4957	15.1980	0.3718	26.0211	0.4318	14.9280
Genetic Algorithm (Proposed)	0.4776	16.9031	0.2989	26.8303	0.3889	16.8023

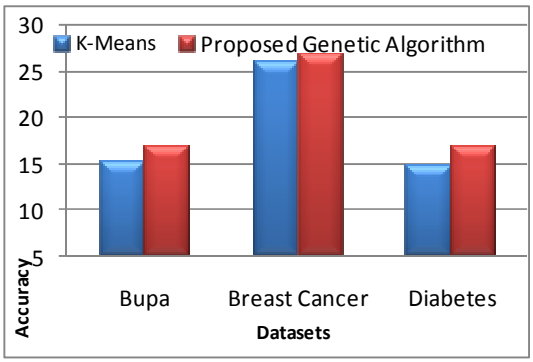

Fig. 3 : Accuracy Value for the Proposed GA

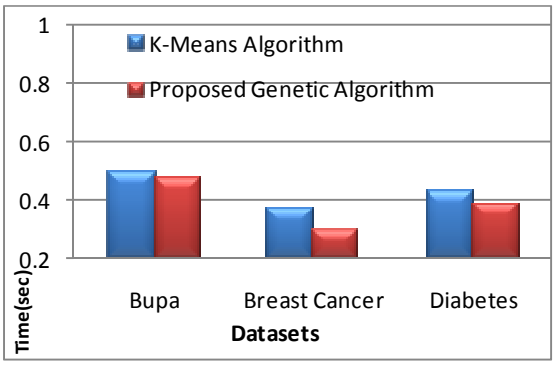

Fig. 4: Execution Time for the Datasets

After conducting 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.

ACKNOWLEDGMENT

This work is funded by the University Grants Commission as a part of the Minor Research Project titled Early Prediction of Human Diseases Using Machine Learning Techniques Based on Medical Data.

REFERENCES

[1] Anna D. Peterson, Arka P. Ghosh and Ranjan Maitra, "A systematic evaluation of different methods for initializing the K-Means clustering algorithm", IEEE transactions on knowledge and data engineering, 2010, pp.522-537.
[2] Ayhan Demiriz, Bennett.K, “Semi-Supervised Clustering Using Genetic Algorithms”, Artificial Neural Networks in Engineering (ANNIE), 1999, pp.809-814.
[3] Bashar Al-Shboul, and Sung-Hyon Myaeng” Initializing K-Means using Genetic Algorithms”, World Academy of Science, Engineering and Technology 54, June 2009, Issue 30, pp.114.
[4] Cheng Min-Yuan and Huang Kuo-Yu “K-Means clustering and Chaos Genetic Algorithm for Nonlinear Optimization”, Information and Computational Technology, 2009, pp.520-526.
[5] Dharmendra K Roy and Lokesh K Sharma, "Genetic K-Means clustering algorithm for mixed numeric and categorical data sets”, International Journal of Artificial Intelligence and applications (IJAIA), Vol.1, No.2, April 2010.
[6] Fang-Xiang Wu1, Anthony J. Kusalik and W. J. Zhang, "Genetic Weighted K-Means for Large-Scale Clustering Problems”, Association for the Advancement of Artificial Intelligence (AAAI) Press, 2005.
[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, pp.1793-8163.
[8] Hao-jun Sun, Lang-huan Xiong, "Genetic Algorithm-based High-dimensional Data Clustering Technique", in Proc. Sixth International IEEE Conference on Fuzzy Systems and Knowledge Discovery, Tianjin, August 2009, Vol.1, pp.485-489.
[9] Ian H. Witten; Eibe Frank, "Data Mining: Practical Machine Learning Tools and Techniques, 2nd Edition", Morgan Kaufmann, San Francisco, 2005.
[10] Indrajit Saha and Anirban Mukhopadhyay “Genetic Algorithm and Simulated Annealing based Approaches to Categorical Data Clustering”, Proceedings of the International MultiConference of Engineers and Computer Scientists (IMECS), Hong Kong,Vol.1, March 2008, pp.19-21.

[11] Yan Wang et al., "A novel quantum swarm evolutionary algorithm and its applications", 2007,Vol.70, Issues 4-6, pp.633-640.

