THE INTERNATIONAL JOURNAL OF SCIENCE & TECHNOLEDGE

A Novel Method for Prediction of Human Diseases Using Machine Learning Algorithms

Dr. P. Sumathi
Assistant Professor, Government Arts College, Coimbatore, Tamil Nadu, India
Dr. V. Kathiresan
Assistant Professor & Head, Department of Computer Applications,
Dr. SNS Rajalakshmi College of Arts & Science, Coimbatore, Tamil Nadu, India

Abstract:
Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. In the past few decades, healthcare industries have become increasingly data intensive and with the advancements with digital technology and low cost storage media have led this growth in size, complexity and quantity of data collected. Medical data mining has great potential for exploring the hidden patterns in the data sets of the medical domain. Hidden patterns are similarities and relationships in medical data like common treatment procedures provided to similar diseases or symptoms patterns that have occurred to previous patient. These patterns, if accurately discovered, can be utilized for clinical diagnosis. This research work proposes the use of data clustering for this purpose.

Keywords: *Clusters, cross over, mutation, patterns, intuitionistic fuzzy c-means, point symmetry distance measure, genetic algorithm.*

1. Introduction

The successful application of data mining in highly visible fields like e-business, marketing and retail have led to the popularity of its use in knowledge discovery in databases (KDD) in order industries and sectors. Among these sectors that are just discovering data mining are the fields of medicine and public health. Advances in medical information technology have enables healthcare industries to automatically collect huge amount of data through clinical laboratory examinations.

Data is a great asset to meet long-term goals of any organization and can help to improve cross-patient analysis and patient-doctor relationship management. It can also benefit healthcare providers like hospitals, clinics and physicians and patients, for example, by identifying effective treatments and best practices. Cross-patient analysis of such data has attracted much interest because it might reveal underlying relationships between the course of examination results and diseases, which might be commonly observed on many patients. This analysis can improve patient care and diagnosis.

Recent advances in software and technological breakthroughs in hardware had made the storage and accessing of huge amount of data economical. It is now possible to operate on large datasets in a reasonable time to perform exhaustive searches and find brute force solutions. In spite these advanced techniques, knowledge discovery from these huge databases is still a big challenge, especially in the field of medicine [1]. Knowledge discovery using data mining techniques is a perfect solution of these situations.

The main challenge now is to find techniques that bridge the two fields, data mining and medical science, for an efficient and successful knowledge discovery. The eventual goal of this data mining effort is to identify factors that will improve the quality and cost effectiveness of patient care.

Usage of 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, and disease diagnosis details 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 patterns.

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 diagnosis.

Cross-patient analysis of such data has attracted much The successful application of data mining in highly visible fields like e-business, marketing and retail have led to the popularity of its use in knowledge discovery in databases (KDD) in order industries and sectors. Among these sectors that are just discovering data mining are the fields of medicine and public health. Advances in medical information technology have enables healthcare industries to automatically collect huge amount of data through clinical laboratory examinations.

Data is a great asset to meet long-term goals of any organization and can help to improve cross-patient analysis and patient-doctor relationship management. It can also benefit healthcare providers like hospitals, clinics and physicians and patients, for example, by identifying effective treatments and best practices. Cross-patient analysis of such data has attracted much interest because it might reveal underlying relationships between the course of examination results and diseases, which might be commonly observed on many patients. This analysis can improve patient care and diagnosis.

Recent advances in software and technological breakthroughs in hardware had made the storage and accessing of huge amount of data economical. It is now possible to operate on large datasets in a reasonable time to perform exhaustive searches and find brute force solutions. In spite these advanced techniques, knowledge discovery from these huge databases is still a big challenge, especially in the field of medicine. Knowledge discovery using data mining techniques is a perfect solution of these situations.

The main challenge now is to find techniques that bridge the two fields, data mining and medical science, for an efficient and successful knowledge discovery. The eventual goal of this data mining effort is to identify factors that will improve the quality and cost effectiveness of patient care.

Usage of 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, and disease diagnosis details 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 patterns. 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 [2]. 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 diagnosis. The successful application of data mining in highly visible fields like e-business, marketing and retail have led to the popularity of its use in knowledge discovery in databases (KDD) in order industries and sectors. Among these sectors that are just discovering data mining are the fields of medicine and public health. Advances in medical information technology have enables healthcare industries to automatically collect huge amount of data through clinical laboratory examinations.

Data is a great asset to meet long-term goals of any organization and can help to improve cross-patient analysis and patient-doctor relationship management. It can also benefit healthcare providers like hospitals, clinics and physicians and patients, for example, by identifying effective treatments and best practices [7]. Cross-patient analysis of such data has attracted much interest because it might reveal underlying relationships between the course of examination results and diseases, which might be commonly observed on many patients. This analysis can improve patient care and diagnosis. Recent advances in software and technological breakthroughs in hardware had made the storage and accessing of huge amount of data economical. It is now possible to operate on large datasets in a reasonable time to perform exhaustive searches and find brute force solutions. In spite these advanced techniques, knowledge discovery from these huge databases is still a big challenge, especially in the field of medicine [3].

Knowledge discovery using data mining techniques is a perfect solution of these situations. The main challenge now is to find techniques that bridge the two fields, data mining and medical science, for an efficient and successful knowledge discovery. The eventual goal of this data mining effort is to identify factors that will improve the quality and cost effectiveness of patient care.

Usage of 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, and disease diagnosis details 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 patterns [9]. 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 diagnosis.

2. Intuitionistic Fuzzy C-means with Point Symmetry Distance Measure (IFCM-PS)

Intuitionistic fuzzy c-means only takes into account the distance between objects and centroids. Hence, could not efficiently cluster non-spherically separable data. We are proposing a robust method by incorporating the distance metric which also considers the variation of points within the cluster into conventional IFCM algorithm to regularize the distance variation in each cluster. IFCM-σ minimizes the objective function as: [11]

$$J_{IFCM-ps} = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{*m} \,.\, \hat{d}_{ki}^{2} + \sum_{i=1}^{c} \pi_i^{*} e^{1-\pi_i^{*}} \tag{1}$$

Where

$$\hat{d}_{ki}^{2} = \frac{\|x_k - v_i\|^2}{ps_i}$$

And PSi is the weighted mean distance of cluster i and is given by

$$ps_i = \left\{ \frac{\sum_{k=1}^{n} u_{ki}^{*m} \cdot \|x_k - v_i\|^2}{\sum_{k=1}^{n} u_{ki}^{*m}} \right\}^{1/2} \tag{2}$$

Here $u_{ik}^{*} = u_{ik} + \pi_{ik}$, where u_{ik}^{*}, denotes the IFCM-PS membership and u_{ik}^{*}, denotes the FCM-PS membership of the kth data in ith class.

3. 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. [12]

3.1. Algorithm 1: 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) [5][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. 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).

3.2. Algorithm 2: Proposed Genetic Algorithm

Yan Wang et al developed an advanced genetic algorithm for complex value encoding. The proposed improved genetic algorithm is developed with simple modification of Yan Wang et al. algorithm [9]. Then the new algorithm [12] 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 $(\rho_1, \rho_2 \dots \dots \rho_m)$ 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 $k = 1, 2 \dots m$

1. Evaluates the fitness of each individual in that population;
2. If pre-specified, the termination criteria are reached, then stop;
3. Select the best-fit individuals for reproduction;

Breed new individuals through crossover and mutation operations to give birth to offspring; Go back to step 2.

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.

4. Comparison of IFCM-PS with Advanced Genetic Algorithm

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 IFCM-PS as given in the Table-I and in the fig 1.and fig 2. So, the IFCM-PS algorithm is executed for 50 times and the readings are noted. Similarly, the hybrid algorithm was also executed for 50 times. Finally, the average value is calculated for each algorithm by using different datasets. The datasets used for this work are Diabetes, Heart disease and Breast Cancer datasets. Three real-life datasets were obtained from UCI Machine learning runs and results were noted repository.

ALGORITHM	**DATA SETS**					
	Heart Diseases		**Breast Cancer**		**Diabetes**	
	Execution Time (sec)	**Accuracy (%)**	**Execution Time (sec)**	**Accuracy (%)**	**Execution Time (sec)**	**Accuracy (%)**
IFCM-PS	0.4856	13.9980	0.3201	25.9211	0.4118	15.8673
Genetic Algorithm (Proposed)	0.4776	16.9031	0.2989	26.8303	0.3889	16.8023

Table 1: Comparative Results of IFCM-PS and Advanced Genetic Algorithm

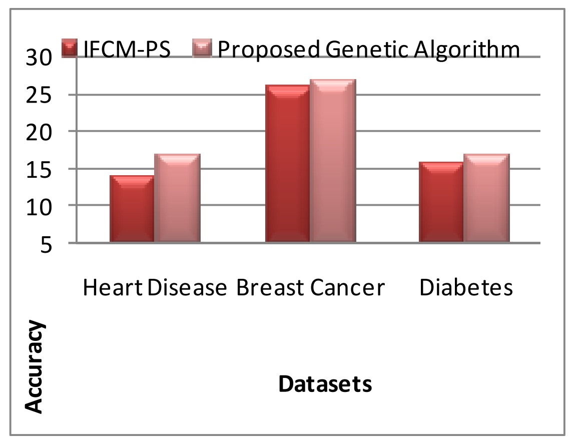

Figure 1: Accuracy Value

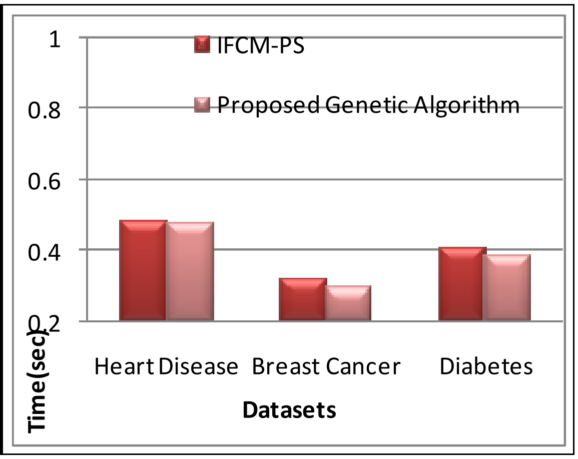

Figure 2: Execution Time for the Datasets

After conducting experiments on various datasets, it is clearly shown 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.

5. Conclusion

The hybrid approach that includes both K-Means algorithm and genetic algorithm yields good result in the process of clustering when compared with IFCM-PS approach. The genetic algorithm shows the improvement in the accuracy and efficiency of the K-Means initialization process. The 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.

6. 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.

7. References

i. Kaur, Prabhjot, A. K. Soni, and Anjana Gosai, "Robust Intuitionistic Fuzzy C-means clustering for linearly and nonlinearly separable data", In Image information Processing (ICIIP), 2011 International Conference on, pp. 1-6.

ii. Tripathy, B. K., Anurag Tripathy, K. Govindarajulu, and Rohan Bhargav, "On Kernel Based Rough Intuitionistic Fuzzy C-means Algorithm and a Comparative Analysis", In Advanced Computing, Networking and Informatics, 2014,Vol. 1, pp. 349-359.

iii. Lin, K, "A Novel Evolutionary Kernel Intuitionistic Fuzzy C-means Clustering Algorithm", 2013, pp.1-1.

iv. Huang, Ching-Wen, Kuo-Ping Lin, Ming-Chang Wu, Kuo-Chen Hung, Gia-Shie Liu, and Chih-Hung Jen, "Intuitionistic fuzzy c-means clustering algorithm with neighborhood attraction in segmenting medical image." Soft Computing, 2014, pp.1-12.

v. 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.

vi. Ayhan Demiriz, Bennett.K, “Semi-Supervised Clustering Using Genetic Algorithms”, Artificial Neural Networks in Engineering (ANNIE), 1999, pp.809- 814.

vii. 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.

viii. 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.

ix. Yan Wang et al.. “An Improved Genetic Algorithm Based Complex-valued Encoding” IJCSNS International Journal of Computer Science and Network Security, Vol.10 No.6, June 2010.

x. TarikArici, SaitCelebi, Ali S. Aydin, Talha T. Temiz. "Robust gesture recognition using feature pre-processing and weighted dynamic time warping." Multimedia Tools and Applications (2013), pp.1-18.

xi. Sumathi P and Kathiresan V "An Intuitionistic Fuzzy C-Means Clustering Algorithm with Point Symmetry Distance Measure", IFRSA International Journal of Data Warehousing & Mining, Vol. 4, issue.4, Nov. 2014. ISSN (PRINT) : 2249-7161, ISSN (ONLINE) : 2249-2186.

xii. Sumathi P and Kathiresan V "A Hybrid Model for Medical Data Using Machine Learning Approaches ", International Journal of Modern Trends in Engineering and Research, Vol. 3, issue 2, Feb. 2016. ISSN (PRINT) : 2393-8161, ISSN (ONLINE) : 2349-9745.