An Efficient Clustering Algorithm for Predicting Diseases from Hemogram Blood Test Samples

S. Vijayarani* and S. Sudha

Department of Computer Science, School of Computer Science and Engineering, Bharathiar University, Coimbatore-641046, Tamil Nadu, India; vijimohan_2000@yahoo.com, sudhas253@gmail.com

Abstract

This research work primarily focuses on predicting diseases from the hemogram blood test data set by using data mining techniques. In this research, a new clustering algorithm which is named as weight based k-means algorithm is developed for identifying the leukemia, inflammatory, bacterial or viral infection, HIV infection and pernicious anaemia diseases from the hemogram blood test samples data set. The newly proposed weight based k-means algorithm efficiency is evaluated with Fuzzy C-means and K-means clustering algorithms. These algorithms performances are evaluated by using the cluster accuracy, error rate and execution time. From the results investigation, it is known that the proposed weight based k-means algorithm performance is better than the other algorithms.

Keywords: Clustering, Disease Prediction, Fuzzy C-Means, K-Means, Weight based K-Means

1. Introduction

Nowadays, data mining techniques had a significant position in the field of healthcare and medical industry to analyze the medical data and finding patterns from those data. The primary objective of this analysis is to predict the diseases from the medical data sets. Many researchers are interested to do their research in this domain. In literature, many research papers are available which mainly focused on predicting diseases from health data sets using data mining techniques.

In 1, neural network techniques are used to perform the analysis process in the Cleveland heart disease data set. In order to train the heart disease dataset, authors used the neural network techniques, back propagation learning algorithm and feed forward neural network. The factors used are momentum and the variable learning rate. Number of nodes in the input layer is 13 and the output layer has four nodes which represents the class labels. With this classification, the data set is classified into four groups. After analyzing the results, it is known that the use of neural network techniques increased the efficiency.

Jabbar et al.3 discussed heart attack prediction using association rule mining technique. A new algorithm CBARBSN is proposed by the author. This algorithm has two significant steps; the first step has performed the conversion process, which is, converting the Cleveland heart data set into binary dataset. After the conversion process, the second step used the proposed CBARBSN algorithm. This algorithm is used for mining association rules using sequence numbers.

The paper entitled “Effective Diagnosis of Coronary Artery Disease Using the Rotation Forest Ensemble Method”4, described how to predict the cardiovascular heart disease using different data mining algorithms. Support vector machine, decision tree and artificial neural networks and RIPPER classifier are used for prediction of cardiovascular heart diseases. WEKA data mining tool is used for implementation. Sensitivity, Specificity, Error in classification and Classification accuracy are the per-

*Author for correspondence
performance parameters for comparing the performance of the above algorithms. Based on the results analysis, it is observed that the support vector machine provide minimum error and highest classification accuracy than other algorithms.

Sumathi et al., discussed the coronary heart disease prediction using enhanced k-means clustering algorithm. Two significant steps of this algorithm are, First, weighted ranking is used which helps to give the solution for selecting the initial centroids randomly. Second important step is, weight can be assigned to each attribute. In order to perform experimentation authors have used Heart dataset which is available in UCI machine learning repository. For assigning weights to different attributes more than 35 conditions are applied. From the experimental results analysis, the author provided the enhanced k-means algorithm developed the good quality clusters and it also maintained the consistency.

In the paper, “Effective Heart Disease Prediction System Using Frequent Feature Selection Method”, authors predicted heart diseases by using frequent feature selection method. In this method nonlinear integral and fuzzy measures are used which has improved the efficiency of the algorithm. With the help of patient’s medical profiles, author has forecasted a chance of patients may get a heart disease. Finally the paper proved that the execution time is decreased and the accuracy is improved by using the proposed algorithm.

Milan Kumari et al., applied rotation forest method to perform successful diagnosis of coronary artery disease which helps to make better clinical decision-making process. Authors have proposed Levenberg-Marquardt back propagation algorithm. The performance factors are classification accuracy, sensitivity, specificity and AUC and specificity.

In Ephzibah E.P., authors used naïve bayes classification technique for predicting the heart disease. From the patient’s medical profile information, the attributes age, gender, blood sugar and blood pressure are used to find the possibility of patients who may have heart problems. A web based questionnaire is developed to collect the information.

Ephzibah E.P. used only minimum number of attributes for heart disease diagnosis using fuzzy expert systems. Two approaches, genetic algorithm and fuzzy logic are combined to perform cost effective and effective heart disease diagnosis. In fuzzy approach, two models i.e. mamdani and takagi-sugeno fuzzy systems and genetic algorithms are used to calculate the cost. A set of all the selected features are considered as input and the value ‘1’ or ‘0’ is an output.

Ananthi Sheshasayee et al., conducted a performance analysis of two clustering algorithms namely K means and Fuzzy C means. In requirements engineering, clustering techniques reduced the cost of software development and maintenance. Authors have proved that the performance of fuzzy C means algorithm is good.

“Predicting the Heart Attack Symptoms using Biomedical Data Mining Techniques” paper has discussed the use of data mining techniques in biomedical domain. Data mining clustering (K-means and mean-clustering) and classification algorithms (KNN) are used for this prediction. The mean clustering algorithm minimizes the number of iterations. By comparing all these algorithms, the mean clustering algorithm is more efficient than other algorithms.

In, authors have analyzed the efficiency of apriori, WAC and naïve bayes classification algorithms. Naïve bayes, Apriori and WAC algorithms are used in this work. Performances of these algorithms are estimated by using the classification matrix which provides the occurrence of correct and incorrect prediction.

Ananthapadmanaban et al. discussed how data mining classification techniques are used for detection of eye disease diabetic retinopathy and they found that naïve bayes method is accurate. In, the authors discussed the tuberculosis disease classification and it has performed using hybrid technique genetic-neuro expert system. The paper “Retinal Identification System using Fourier-Mellin Transform and Fuzzy clustering” provided a new approach for retinal identification using fuzzy c means clustering algorithm. The authors concluded that the proposed technique has produced higher accuracy with minimum error rate.

2. Objectives of the Problem

The primary goal of this research work is to analyze the hemogram blood test samples to predict the diseases. Clustering algorithms namely K means and Fuzzy C Means are used for this analysis (Figure 1). In addition to this, a new weight based k-means algorithm is also proposed. Data mining K means clustering algorithm is enhanced by assigning weight to the data items. The performance factors used for finding an efficient algorithm are clustering accuracy, execution time and error rate.
Five diseases namely Leukemia, Inflammatory disease, bacterial or viral infection, HIV infection and pernicious anaemia are predicted from this hemogram blood test samples analysis.

3. Proposed Methodology

A Hemogram/Complete Blood Count (CBC) is possibly the commonest laboratory test conducted in any laboratory. As the name suggests, a hemogram/CBC analyses various blood components like RBC, WBC and platelets. These components are non-specific but very important indicators of general well being of an individual. Many of the parameters included in a hemogram do not pinpoint towards a specific diagnosis although careful analysis of a combination of several parameters included in a hemogram may help the clinician/hematologist to arrive at important conclusions. Sample required is EDTA whole blood and the entire range of tests can easily be conducted with just 2/3 ml of blood\textsuperscript{16}.

4. Clinical Significance

The definition of complete blood count is nothing but the number of cellular components in the blood. The complete blood count test, the physicians have considered three major cells. The three important blood cells are RBC-Red Blood Cell, WBC-White Blood Cell and Thrombocytes or Platelets.

RBCs are the most general kind of blood cell. The most important task of RBC is to deliver oxygen via blood flow to body tissues through the circulatory system. Red blood cells are also called as RBCs, red cells, erythrocytes, haematids and red blood corpuscles. The RBC count is below the normal range it leads to develop several blood diseases, i.e. anemia, hemolysis, polycythemias, microangiopathic diseases and hemolytic transfusion reaction. Some of the important anemic diseases are sickle cell disease, iron deficiency anemia, hereditary spherocytosis, pernicious anemia, aplastic anemia, pure red cell aplastia and thalassemia.

Hemoglobin is a protein molecule and it is found in the red blood cells. The core task performed by hemoglobin is the transmission of oxygen to the body tissues from the lungs and takes carbon dioxide from the body tissues. Hemoglobin test is a common blood test which helps to determine how much hemoglobin is in our body. The persons whose hemoglobin level is lower than normal they may suffer with diseases like anemia, chronic kidney diseases, digestive tract bleeding and bone marrow. Whether the persons whose hemoglobin level is higher than normal hemoglobin they may have diseases like congenital hear disease, cor pulmonale, pulmonary fibrosis, and bone marrow.

White blood cells are also termed as leukocytes, cells of the immune system. WBC’s are in charge for shielding our body from infectious diseases. White blood cells are categorized into five different types; they are basophils, eosinophils, lymphocytes, monocytes and neutrophils. Low white blood cell count causes bone marrow deficiency, lupus, liver or spleen disease, viral illness and rigorous bacterial infections. High white blood cell count gives anemia, infections, inflammatory disease, leukemia, relentless mental or physical stress and tissue damage. Platelets are small blood cell that helps to protect our body to form clots which stops bleeding. A platelet count specifies the amount of platelets in a volume of blood.
5. **Major Parameters**

The following chart indicates the definitions of the terms that appear in a standard hemogram report (Table 1). It depicts the details about the clinical significance of the findings\(^\text{16}\).

Table 1. Hemogram Report Parameters

| Parameters                          |
|------------------------------------|
| RBC Count                          |
| RDW- Red Blood Cell Distribution Width |
| Hemoglobin                         |
| PCV - Hematocrit or Packed Cell Volume |
| MCH - Mean Corpuscular Hemoglobin  |
| MCHC - Mean Corpuscular Hemoglobin Concentration |
| MCV - Mean Corpuscular Volume      |
| Total White Blood Cells            |
| Lymphocytes                        |
| Neutrophils                        |
| Basophils                          |
| Monocytes                          |
| Platelet Count                     |

6. **Clustering Algorithms**

Among many data mining techniques, clustering technique is considered as one of the efficient and popular data mining techniques and it clusters or groups the data items based on their similarity. Several similarity and dissimilarity measures are applied to find the relationships and patterns which exist in those data items. Many types of clustering algorithms are available; they are, hierarchical algorithms, partitioning algorithms, density based, grid based and distance based algorithms.

6.1 **K-means Clustering Algorithm**

In data mining clustering, the most popular and traditional algorithm is k-means clustering algorithm\(^\text{17}\). It has the ability to handle massive data and cluster those data efficiently and quickly. A simple and straightforward iterative method is used to partition the data set into k-number of clusters. The characteristics of this approach are given below:

- For each and every k cluster, D\(_j\) is represented by the mean (or weighted average) d\(_j\) of its objects, the centroid.
- To find the centroids of the object the clusters are iteratively recomputed.

The working of k-means algorithm is described here. The first step of this algorithm is the selection of data objects, i.e. random selection of k number of data objects and these are considered as centroid or cluster mean. Remaining data objects are assigned to any one of these k clusters in which distance are most similar. The next step is to perform calculation for the new mean value for all the clusters. This process is continued until the termination has occurred.

K-means clustering has two significant phases. During the first phase, the individual centroids are defined to k clusters. In the second phase the data items whose values nearest to the centroids are clustered together. Similarity and distance measures are used for finding the similarity between the data items. Euclidean distance is generally considered to determine the distance between data points and the centroids\(^\text{18}\) (Table 2).

Table 2. K-Means Clustering Algorithm

1. Randomly pick \(k\) data objects from the database D and consider these data objects are cluster centers for initial clusters;
2. **Repeat**
   2.1 Consider the remaining objects and (re) assign these objects to clusters based on the similarity.
   2.2 Next step is the cluster mean updation process, here new mean value is calculated based on the newly entered data objects in the clusters.
3. **Until** there is no change;

6.2 **Fuzzy C-Means Clustering Algorithm**

The working behavior of fuzzy c-means algorithm is, the membership is assigned to every data item which corresponds to every cluster center. This has performed to calculate the distance between the data point and the
cluster center. Sum of membership of every data point must be equal to one (Table 3).

### 6.3 Weight based K-means Algorithm

The prime objective of the proposed algorithm is to defeat the difficulty in k-means clustering algorithm that is sensitive to initial condition. Selecting the different initial condition may attain different cluster results. The algorithm may be caught in the local optimum. For the same medical dataset, different cluster may lead to different sets of rule for building a classifier model. The accuracy of the system again depends on the local optimum. To improve the accuracy, efficiency and consistency in the cluster members, we proposed weight based algorithm. In this work, we considered hemogram blood dataset to predict the various risk level based on the invasive and noninvasive data. Weight range is assigned by the physician to all the data items for developing the proposed system.

The first step of this algorithm calculates the sum of the distance from origin to each weighted attribute of a data point in a data set. Weighted data points are calculated using the formula given below:

\[
U_j = \sum_{i=1}^{n} W_i X_i
\]

Where \( W_i \) represents the weightage of \( X_i \) attribute. For \( n \) data points, \( n \) numbers of \( U \) values are calculated. Based on the sorted distances, the original data points of hemogram blood data sets are sorted and the sorted data points are grouped into k-number of equal sets. For each and every group, middle points are considered as the initial centroids. These initial centroids direct the consistent unique clustering results. The following algorithm describes the process for calculating the unique initial centroids. This method gives the consistency in the selection of initial centroids (Table 4).

### 7. Experimental Results

Experiments are carried out on hemogram blood dataset which are collected from Kovai Scan Center. The real dataset is collected to predict the diseases and also to find the efficiency of the algorithm. This dataset has 524 instances and thirteen attributes. In this research work,
five types of diseases namely Leukemia, Inflammatory Disease, Bacterial or Viral Infection, HIV Infections and Pernicious Anaemia are predicted. For implementation Matlab R2010a is used. The performances of the clustering algorithms are analyzed by using the factors clustering accuracy, execution time and error rate.

7.1 Time Factor

The execution time factor describes the amount of time needed for predicting the diseases. This helps to find out the time requirement of existing and the proposed clustering algorithms. Table 5 shows that the time taken by each algorithm for predicting the diseases. From this, the proposed Weight Based K means clustering algorithm has taken less time when compared to other algorithms (Figure 2).

Table 4. Weight based K-Means Clustering Algorithm

| DISEASES               | K MEANS (in msec) | FUZZY C MEANS (in msec) | Weight based K MEANS (in msec) |
|------------------------|-------------------|-------------------------|-------------------------------|
| Leukemia               | 2160              | 3080                    | 617                           |
| Inflammatory Disease   | 2060              | 2200                    | 618                           |
| Bacterial Or Viral Infection | 2640          | 2700                    | 624                           |
| HIV Infection          | 750               | 800                     | 731                           |
| Pernicious Anaemia     | 710               | 900                     | 692                           |

Figure 2. Time Factor.
7.2 Clustering Accuracy

Table 6 shows the clustering accuracy. The clustering accuracy is nothing but, we have to verify whether all the data items are grouped or clustered correctly to their respective groups or not. It could be seen that the clustering accuracy of weight based k means algorithm is better when compared to k means and fuzzy c means (Figure 3).

| DISEASES                | K MEANS (%) | FUZZY C MEANS (%) | WEIGHTED K MEANS (%) |
|-------------------------|-------------|-------------------|----------------------|
| Leukemia                | 78          | 75                | 85                   |
| Inflammatory Disease    | 85          | 80                | 90                   |
| Bacterial Or Viral Infection | 80      | 74                | 98                   |
| HIV Infection           | 88          | 88                | 93                   |
| Pernicious Anaemia      | 88          | 83                | 94                   |

7.3 Error Rate

This performance measure is used to identify the percentage of data items which has placed incorrectly in the set of clusters. By analyzing this measure, the proposed Weight Based K means clustering algorithm attains least error rate when compared to k means and fuzzy c means algorithm (Figure 4). This is represented in Table 7.

Table 7. Error rate

| DISEASES                | K MEANS (%) | FUZZY C MEANS (%) | WEIGHTED K MEANS (%) |
|-------------------------|-------------|-------------------|----------------------|
| Leukemia                | 22          | 25                | 15                   |
| Inflammatory Disease    | 15          | 20                | 10                   |
| Bacterial Or Viral Infection | 20      | 26                | 12                   |
| HIV Infection           | 12          | 12                | 7                    |
| Pernicious Anaemia      | 12          | 17                | 6                    |

Figure 3. Clustering Accuracy.

Figure 4. Error rate.
8. Conclusion

This research paper discussed about the disease prediction using hemogram blood dataset. The data was collected from blood test results of 524 patients who had been examined by hematologists, at the Kovai Scan Center in Coimbatore. We have predicted five types of diseases using hemogram blood samples. The diseases are Leukemia, Inflammatory disease, bacterial or viral infection, HIV infection and Pernicious anaemia. In this research, K means clustering algorithm, fuzzy c means clustering and the weight based k means clustering algorithms are compared by using the performance factors namely time, clustering accuracy and error rate. The proposed weight based k means clustering algorithm has performed well when compared with other algorithms. The proposed weight based k means clustering algorithm attains highest accuracy in less time. In future this work is to be extended by using other data mining algorithms and optimization algorithms for predicting other diseases from the hemogram blood test data set by using reduced attributes.

9. References

1. Rani U. Analysis of heart diseases dataset using neural network approach. IJDKP. 2011 Sep; 1(5): 1–8.
2. Soman, Diwakar S, Ajay V. Insight into Data mining theory and practice. Prentice Hall of India; 2008.
3. Jabbar MA, Chandra P, Deekshatult BL. Cluster based association rule mining for heart attack prediction. JTAT. 2011 Oct; 32(2):196–201.
4. Karabulut EM, Ibrikci T. Effective diagnosis of coronary artery disease using the rotation forest ensemble method. 2011 Sep 13; 3011–8.
5. Sumathi, Kirubakaran. Enhanced Weighted K-Means Clustering Based Risk Level Prediction for Coronary Heart Disease. European Journal of Scientific Research. 2012; 71(4):490–500.
6. Sarumathi, Nithya NS. Effective heart disease prediction system using frequent feature selection method. International Journal of Communications and Engineering. 2012 Mar; 1(1).
7. Kumari M, Godara S. Comparative study of data mining classification methods in cardiovascular disease prediction. IJCST. 2011 Jun; 2(2):304–8.
8. Subbalakshmi G, Ramesh K, Rao MC. Decision support in heart disease prediction system using Naïve Bayes. Indian Journal of Computer Science and Engineering. 2011 Apr-May; 2(2).
9. Ephzibah EP. A hybrid genetic-fuzzy expert system for effective heart disease diagnosis. ACITY CCIS. 2011; 198:115–21.
10. Sheshasayee A, Sharmila P. Comparative study of fuzzy C means and K means algorithm for requirements clustering. Indian Journal of Science and Technology. 2014 Jun; 7(6):853–7.
11. Jaya Rama Krishniah VV, Chandra Sekar DV, Rao RH. Predicting the heart attack symptoms using biomedical data mining techniques. The International Journal of Computer Science & Applications. 2012 May; 1(3).
12. Aditya Sundar N, Pushpa Latha P, Rama Chandra M. Performance analysis of classification data mining techniques over heart disease data base. International Journal of Engineering Science and Advanced Technology. 2(3):470–8.
13. Ananthapadmanaban KR, Parthiban G. Prediction of chances - diabetic retinopathy using data mining classification techniques. Indian Journal of Science and Technology. 2014 Oct; 7(10):1498–503.
14. Geetha PV, Lukshmi RA, Venkatesan P. Tuberculosis disease classification using genetic-neuro expert system. Indian Journal of Science and Technology. 2014 Apr; 7(4):421–5.
15. Hadi J, Hamid T. Retinal identification system using Fourier-mellin transform and fuzzy clustering. Indian Journal of Science and Technology. 2014 Sep; 7(9):1289–96.
16. Health Screen. A Magazine for Pre-patient Care. 2007 Jul; 3(31).
17. Abdul Nazeer KA, Sebastian MP. Improving the accuracy and efficiency of the K-means clustering algorithm. Proceedings of World Congress on Engineering, International Conference on Data Mining and Knowledge Engineering; 2009. p. 1.
18. Wu X, Kumar V, Quinlan JR, Ghosh J, Yang Q, Motoda H, McLachlan GJ, Ng A, Liu B, Yu PS, Zhou Z-H, Steinbach M, Hand DJ, Steinberg D. Top 10 algorithms in data mining. Knowledge Information Systems. 2008; 14:1–37. doi: 10.1007/s10115-007-0114-2.
19. Bezdek, JC, Pattern Recognition with Fuzzy Objective Function Algorithms. New York: Plenum Press; 1981.