EFFICIENCY OF DATA MINING TECHNIQUES FOR PREDICTING KIDNEY DISEASE

S. Gopika[#1], Dr.M.Vanitha[*2]
[#]Research Scholar, PG and Research Dept of Computer Science,
J.J College of Arts and Science (Autonomous) Pudukottai, Tamil Nadu
Email:gopikasrinivasan89@gmail.com
[*]Assistant Professor, Dept of Computer Applications, Alagappa University, Karaikudi, Tamil Nadu
Email: mvanitharavi@gmail.com

Abstract - Chronic kidney disease is an aging problem in the current growing population. Kidney disease surveillance and prediction is very important for patients to provide adequate and appropriate treatment at the right time. Data mining can extract interesting patterns for gigantic medical databases. Patients with kidney disease can be automatically analyzed from their disease data taking into account prior predictions. Though medical data is heterogeneous in nature including text, graphics and images, unwanted data can be removed to provide useful medical information on a patient. Medical data mining can detect disease patterns and predict severity of a patient's disease. Conformist theories are more pertinent than probabilistic theories for results as precise results and inferences become a necessity to save a patient's life. Fuzzy systems are generally used as they produce results based on mathematics, instead of probabilistic arbitrations like neural networks. The paper proposes new algorithm Improved Hybrid Fuzzy C-Means (IHFCM) which is an improvisation of FCM with Euclidean distances to predict kidney diseases in patients.

Keywords: Data Mining, Fuzzy System, Kidney Disease, Neural Network, Disease Prediction.

I. INTRODUCTION

Data mining is particularly useful in the medical field. Medical database consist of a large number of patients, diseases, hospitals, medical equipment and complex data. People's current lifestyle, working environment and eating habits lead to many diseases, one of which includes chronic kidney disease. Data mining can be used to extract, analyze and solve for knowledge, more in medical data sets. Data mining tools and techniques can be applied to processed data for aiding health professionals in appropriate decision making and thus improving the management of patients.

Chronic kidney diseases (CKD) are prevalent globally and need prompt detection and diagnosis. Researchers use data mining techniques to detect CKD in patients. Classification and clustering is a data mining technique used to predict group membership from data instances. Classification is comparable to clustering as it can segment information retrieval into distinct segments called classes. In predictions algorithms deal with a set of attributes with corresponding results, often called a training set of target or predictive attributes.

Chronic kidney disease (CKD) whose presence is global is the condition in which the kidney is damaged and toxic wastes are not filtered in the body, making it imperative to researches in this area. This proposes an improved hybrid FCM algorithm for detection of life-threatening CKD and demonstrates the empirical results of the study.

II. LITERATURE REVIEW

Pavithra N et al's work [1] showed a symbolic fuzzy C-means clustering algorithm with fuzzy knowledge in the structure. They improved attributes for recording higher accuracy without a structured knowledge test. The system was used to predict and diagnose patients with renal dysfunction. Their system could predict risk factors for renal failure, instead of severity based on medical tests. The FCM-related clustering algorithm was applied to the location of the disease in kidney disease patient files. Their adjustments on various FCM factors, changed strategies effectively correlated and found the best variety of clusters, discovering anomalies in traditional cases. The initial pre-processing information was based on repetitive records. In the classification phase, an FCM classification classified information on the risk level of a kidney disease.

Basma Boukenze et.al [2] pre-processed data with conversions and data mining methods to gain knowledge about the interaction between measurement parameters and the survival of a patient. Two data mining algorithms were used to form decision rules in extracting knowledge and predict the survival of patients. They explained the significance of exploring important parameters using data mining. Their new concept was implemented and tested using dialysis data collected from four different sites. Their method also reduced the cost and effort in selecting patients for clinical trials. The patients were selected based on predicted results and significant parameters found in their analysis.

Neha Sharma et al [3], detected and predicted kidney diseases as a prelude to proper treatment to patients. The system was used for detection in patients with kidney disease and the results of their IF-THEN rules predicted the presence of a disease. Their technique used two fuzzy systems and a neural network called a neural blur system, based on the result of the input data set obtained. Their system was a combination of fuzzy systems that produced results using accurate mathematical calculations, instead of probabilistic based classifications. Generally results based on mathematics tend to have higher accuracies. Their work was able to obtain useful data along with optimizations in results.

Veenita Kunwar et al [4]. In their study predicted chronic kidney disease (CKD) using naive Bayesian classification and artificial neural network (ANN). Their results showed that naive Bayesian produced accurate results than artificial neural networks. It was also observed that classification algorithms were widely used for investigation and identification of CKDs.

Swathi Baby P et al [5] demonstrated that data mining methods could be effectively used in medical applications. Their study collected data from patients affected with kidney diseases. The results showed data mining's applicability in a variety of medical applications. K-means (KM) algorithm can determine number of clusters in large data sets. Their study analyzed tree AD, J48, star K, Bayesian sensible, random forest and tree - based ADT naive Bayesian on J48 Kidney Disease Data Se and noted that the techniques provide statistical analysis on the use of algorithms to predict kidney diseases in patients.

III. PROBLEM FORMULATION

Probability theory cannot be used to obtain the results in prediction of kidney diseases as it involves the patient's life and the exact results are a necessity. Statistical methods, Bayesian classification or association rule based predictions cannot be used to predict CKD as the results obtained may be less accurate. Predicting disease can save a patient's life and if detected early can help proper cure of the disease. Thus a need to evolve CKD prediction with new techniques.

IV. PROPOSED WORK

Diseased kidneys are increasing in an aging population making it imperative to monitoring or prediction diseased kidneys. General predictions are based on a set of if then rules on kidney datasets. Erroneous predictions of CKD can lead to loss of life. The proposed a new technique IHFCM is used for predicting and detecting kidney disease in a patient data set.

V. METHODOLOGY

A. Fuzzy Model

Fuzzy grouping is based on generation of graphs for each pattern within the group. Fuzzy modeling can match human reasoning models and manage data. The main advantages of fuzzy logic include its simplicity and flexibility. Fuzzy logic can handle inaccurate and incomplete data where traditional statistical models may fail. A fuzzy system can be any model of a complex nonlinear function and provides transparency with explanation on rules. These rules can be potential clinical guidelines.

B. Fuzzy C Means

The fuzzy c-means (FCM) algorithm is a traditional and classical image segmentation algorithm. It is a method that allows clustering, where data may belong to two or more clusters The FCM algorithm focuses on minimizing the value of an objective function that measures the quality of the partitioning a dataset into clusters. It produces an optimal partition by minimizing the weights within a group sum of squared error objective function. It is frequently used in pattern recognition. The fuzzy C-means algorithm is listed below in Figure 1

Fuzzy C Means Algorithm

Input: Feature extracted CT scan kidney segmented image

Output: given image has Kidney Disease or not kidney disease

Step 1: Set the number of clusters

Step 2: Set the fuzzicification parameter, image size and ending condition.

Step 3: Initialize randomly the fuzzy cluster and conditions.

Step 4: Set the loop condition initialize by 0

Step 5: Modify the segmented matrix $M=\{M_{ij}\}$ using Euclidean Distance

Step 6: Modify the Cluster conditions using fuzzy membership function (MF)

Step 7: If $(MAX|MF_{new} - MF_{old}| < \text{End Condition})$ then Stop

Step 8: otherwise increment Loop condition +1 and go to step 5.

Fig. 1 – FCM Algorithm

Where MF= [MF_1, MF2… MF_C] are membership function of cluster condition. At the end point, a defuzzification process takes place to convert the fuzzy image to crisp segmented image. The disadvantages of FCM includes a priori specification of the number of clusters, Euclidean distance measures may unequally weigh factors and takes more number of iterations to produce better results.

C. Kernel Fuzzy C Means Clustering Algorithm (KFCM)

FCM algorithm with added kernel information is KFCM and can overcome FCM disadvantages. Kernel maps nonlinear input data into high dimensional features possibly with infinite dimensionality. It uses the dot products in the kernel space and expresses it as a Mercer kernel. The distance in the kernel need not be explicitly computed as it can be replaced by a function or kernel trick. The Kernel fuzzy C-means algorithm is listed below in Figure 2

Kernel Fuzzy C Means Algorithm

Input: Feature extracted CT scan kidney segmented image

Output: given image has Kidney Disease or not kidney disease

Step 1: Set the number of clusters

Step 2: Set the fuzzicification parameter, image size and ending condition.

Step 3: Initialize randomly the fuzzy cluster and conditions.

Step 4: Set the loop condition initialize by 0

Step 5: Initialize inner product kernel function.

$$|\varphi(X_k) - \varphi(V_i)| = K(x_k, x_k) + K(v_i, v_i) - 2K(x_k, v_i)$$

Step 6: Minimize the Constraints of kernel function

Step 7: Modify the segmented matrix $M=\{M_{ij}\}$ using Euclidean Distance (d)

Step 8: Modify the Cluster conditions using fuzzy membership function (MF)

Step 9: If $(MAX|MF_{new} - MF_{old}| < \text{End Condition})$ then Stop

Fig. 2 – KFCM Algorithm

Where MF= [MF_1, MF2… MF_C] are membership function of cluster condition. At the end point, a defuzzification process takes place to convert the fuzzy image to crisp segmented image.

D. Proposed Improved Hybrid Fuzzy C Means Clustering Algorithm (IHFCM)

The fuzzy c-means is introduced by Ruspini and then extended by Dunn and Bezdek and is widely used as clustering analysis, pattern recognition and image processing in Fuzzy C Means Clustering Algorithm (FCM). It is based on the K-means and the basic idea of FCM that each data point belongs to the membership in the degree of poor clustering, and K means that each data point belongs to a particular group or not. So FCM uses fuzzy partitioning so that when you can belong to multiple groups, the members are between 0 and 1. However, through the degree of data provided by the degree of membership, FCM still uses the cost function to try to split the data set. When minimized. It makes the matrix member having a U element value between 0 and 1. The algorithm works iteratively through the preceding two conditions until the no more improvement is noticed. In a batch mode operation, FCM determines the cluster centers i, c and the membership matrix U using the following steps:

Input: Feature extracted CT scan kidney segmented image

Output: given image has Kidney Disease or not kidney disease

Step 1: Set the number of clusters

Step 2: Set the Fuzzification parameter, image size and ending condition.

Step 3: Initialize randomly the fuzzy cluster and conditions.

Step 4: Set the loop condition initialize by 0

Step 5: Calculate the weighted fuzzy factor using Euclidean distance measure.

$$W_{ij} = (1 / d_{ij}+1)$$

Step 6: Modify the segmented matrix M= $\{M_{ij}\}$ using Euclidean Distance (d)

Step 7: Modify the Cluster conditions using fuzzy membership function (MF)

Step 8: If ($MAX|MF_{new} - MF_{old}| <$ End Condition) then Stop

Step 9: otherwise increment Loop condition +1 and go to step 5.

Where MF= [MF_1, MF2… MF_C] are membership function of cluster condition. At the end point, a defuzzification process takes place to convert the fuzzy image to crisp segmented image.

IHFCM can be applied to Identifying a disease in a patient's dataset and even be used for Drug Activity Prediction.

VI. EXPERIMENTAL RESULTS

This work is done on MATLAB which can manipulate matrices, product functions and data, implement algorithms, create user interfaces, and interact with programs written in other languages. The experimental IHFC is worked on MATLAB. The data set is extracted from the reference point UCI library machine. In the UCI machine learning library they are in the machine learning community used in the machine learning algorithm to conduct an empirical analysis of the field of database theory and data generation. The document was created by David Aha in the 1987 FTP file and other graduate students at the University of California, Irvine. Since then, it has been widely used by students, educators and researchers from major sources of data collection machines around the world.

A. Fuzzification Score

The algorithm calculates the fuzzy C meaning as the diffuse score for each value in the corresponding table of the contents of the query that is entered as a score. The higher the score, the more similar the string. A score of 1.0 or 0.9 means that the fuzzy score results in a highly risky clustering. 0.0% means that the corresponding symptoms have a risk level that is less affected or is not at risk. The user can enter the minimum and highest possible risk factors that are set to contact the doctor and the base, the individual gives each query score, FCM is divided into two categories with the lowest and highest levels found again with the result with the given range of values Find the minimum and maximum scores given to their limits. Thus, FCM can provide three low-risk scores for finding high-risk results with fuzzy scores, fuzzy average sub-risk and cluster-based results.

B. Results

The performance of FCM is evaluated by statistical measures like sensitivity, specificity and accuracy to illustrate the normal life style score. These metrics also enumerate how the test was good and consistent. Sensitivity evaluates the normal life style score correctly at detecting a disease positively. Specificity measures how the proportion of patients without disease can be correctly ruled out. The objective function of IHFCM is depicted in Figure 3. The comparative performance of the algorithms is listed in table 1 and Figure 4.

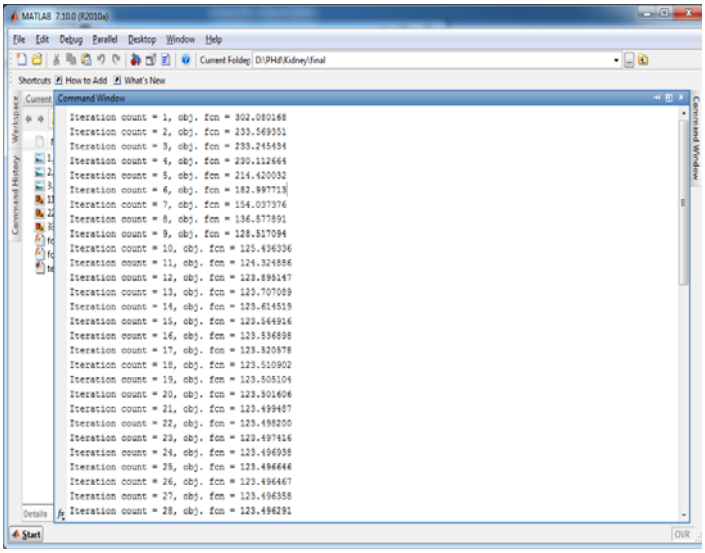

Fig. 3. Objective function of IHFCM

TABLE I. Performance of FCM, KFCM and IHFCM

Clustering Technique	*Type*	*No of Patients*	*Sensitivity*	*Specificity*	*Accuracy*
FCM	*High*	*200*	*90.697*	*23.255*	*96*
KFCM	*High*	*200*	*91.111*	*24.39*	*96*
IHFCM	*High*	*200*	*95.744*	*27.027*	*96*

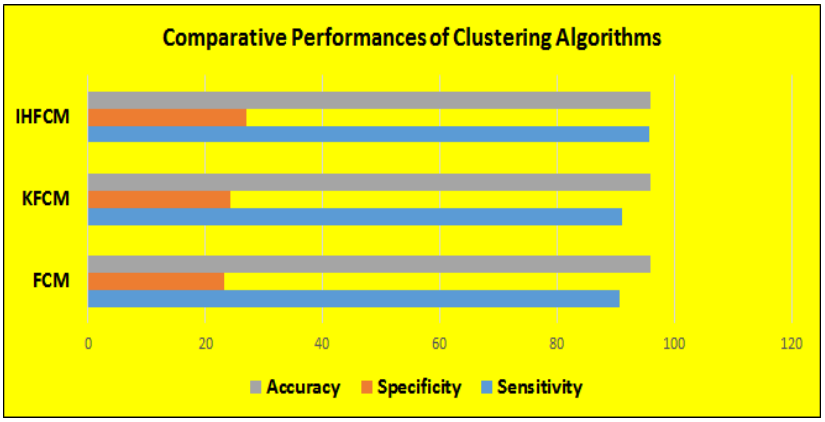

Fig. 4 – Comparative performances of FCM, KFCM and IHFCM on multiple parameters

VII. CONCLUSION

The proposed IHFCM is an extension of FCM and is applied for locating kidney disorders in patient records. The paper demonstrates that correct adjustment to FCM can help build a new strategy for discovering unusual and traditional cases. Initial pre-processing of IHFCM is deleting duplicate records. Results of clustering which obtained from 300 patients showed that FCM based clustering algorithms achieve higher accuracy than most existing algorithms. The proposed IHFCM's performance has been proved clearly in terms of accuracy.

REFERENCES

[1] Pavithra N, Dr. R. Shanmugavadivu, “Efficient Early Risk Factor Analysis of Kidney Disorder Using Data mining Technique”, International Journal of Innovative Research in Computer and Communication Engineering, (2017)
[2] Basma Boukenze, Hajar Mousannif and Abdelkrim Haqiq, “Performance of Data Mining Techniques to Predict In Healthcare Case Study: Chronic Kidney Failure Disease”, International Journal of Database Management Systems (IJDMS), (2016)
[3] Neha Sharma, Er. Rohit Kumar Verma, “Prediction of Kidney Disease by using Data Minining Techniques”, Prediction of Kidney Disease by using Data Minining Techniques, (2016)
[4] Veenita Kunwar, Khushboo Chandel, A. Sai Sabitha, and Abhay Bansal, “Chronic Kidney Disease Analysis Using Data Mining Classification Techniques”, IEEE, (2016).
[5] Swathi Baby P and Panduranga Vital T, “Statistical Analysis and Predicting Kidney Diseases using Machine Learning Algorithms”, International Journal of Engineering Research & Technology (IJERT), (2015)

[6] Dr. S. Vijayarani1, Mr. Dhayanand, "Data Mining Classification Algorithms for Kidney Disease Prediction", International Journal on Cybernetics & Informatics (IJCI) (2015).
[7] Jayalakshmi and Santhakumaran, "Improved Gradient Descent Back Propagation Neural Networks for Diagnoses of Type II Diabetes Mellitus", Global Journal of Computer Science and Technology, (2010)
[8] Koushal Kumar and Abhishek, "Artificial Neural Networks for Diagnosis of Kidney Stones Disease", International Journal Information Technology and Computer Science, (2012)
[9] Rajalakshmi, Neelamegam and Bharathi, "Diagnosis and Classification of Level of Kidney function Using Associative Neural Network and Polynomial Neural Network", Research Journal of Pharmaceutical, Biological and Chemical Sciences, (2013).
[10] Vijayarani and Dhayanand, "Kidney disease prediction using Support Vector Machine and Artificial Neural Network algorithms", International Journal of Computing and Business Research (IJCBR), (2015).
[11] Shweta Kharya, "Using data mining techniques for diagnosis and prognosis of cancer disease", International Journal of Computer Science, Engineering and Information Technology (IJCSEIT), (2012)
[12] Abhinandan Dubey, "A Classification of CKD Cases Using Multi Variate K-Means Clustering", International Journal of Scientific and Research Publications, (2015).
[13] Yashpal Singh and Alok Singh Chauhan," Neural Networks in Data Mining", Journal of Theoretical and Applied Information Technology, (2005)
[14] Baylor, Konukseven and Koku," Control of a Differentially Driven Mobile Robot Using Radial Basis Function Based Neural Networks", World Scientific and Engineering Academy and Society Transactions on Systems and Control, (2008).
[15] Shubham Bind, Arvind Kumar Tiwari and Anil Kumar Sahani, "A Survey of Machine Learning Based Approaches for Parkinson Disease Prediction", International Journal of Computer Science and Information Technologies, (2015).