OF-KNN Technique: An Approach for Chronic Kidney Disease Prediction

1R. Subhashini and 2M.K. Jeyakumar

1Noorul Islam Center for Higher Education, Kanyakumari, Tamil Nadu, India.
baskisubha24@gmail.com

2Additional Controller of Examinations, Noorul Islam Center for Higher Education, Kanyakumari, Tamil Nadu, India.

Abstract

Almost 10% of the population in the worldwide is affected with a major health problem chronic kidney disease. However, the prediction of chronic kidney diseases is evidently done using systematic and automatic methodologies. Among the methodologies, the machine learning is one of the very kind. The classifier in the machine learning algorithms can provide the class labels to the test samples with known features and unknown class. The existing works with machine learning algorithms fail to provide the accuracy of prediction to the needed extent. In order to fulfil the gap, this paper proposes a novel classification strategy to predict the chronic kidney disease from the medical dataset with environmental factors. The methodology introduced in this article is Optimal Fuzzy-K nearest neighbour technique. The optimum performance of fuzzy is obtained by tuning the membership functions utilizing the Bat optimization algorithm. Then the OF is utilised to measure the similarity in the KNN for the classification of disease. The performance of the proposed technique is analysed by the comparison with conventional methods. The accuracy is considered as the primary metric to evaluate the performance, and it is proved the proposed method provides better classification accuracy.

Key Words: Chronic kidney diseases, bat algorithm, fuzzy technique, K-Nearest Neighbour, training-testing samples, triangular membership function.
1. Introduction

Data mining is nothing but the main process of extracting information that's hidden in an informative database (Koh, H.C. and Tan, G., 2005). Predictive and descriptive are two preferred models of data mining which performs many tasks, estimation of values done by predictive model whereas descriptive model mainly identifies the relationship in data. Data mining have its application in several domains named image mining, text mining and sequential mining, etc., Extracting knowledge through data mining is done by the process called classification, clustering and association rules, these tasks comes under predictive and descriptive models. Classification is same as clustering where information is split into segments named classes (Ramya. S and Senthil Kumar, 2004).

Well known Mining applications in fields like e-business, marketing and most recently it has influence in health care (Cao, et al.,2008). Some of the benefits of data mining are false detection in health insurance, low-cost medical facilities and early detection of the cause of disease and appropriate care for it. It gives a chance for medical researchers in developing new medical policies, new drug recommendation systems and in developing separate health profiles for individuals (Yoo, I et al. 2012). Scientific analysis of data mining says that there is a different method from the nature of dataset from market-driven applications which leads to a detailed survey in data mining application in health care sector furnishing the types of data used and information details.

Data mining algorithms have a useful role in health care industries in the prediction cum diagnosis of diseases. These data mining applications relate to medical device industries and pharmaceutical industries including hospital management, with the main aim of finding useful and hidden information from the database. The process of knowledge discovery includes developing, understanding, selection, the creation of data and pre-processing of data. Data mining tool is found complex and time consuming instead, for advanced information findings use of a database is recommended information (Ranjan and Jayanthi, 2007).

In health care, data mining is a benefit in grouping patients with the similar disease or health issues so that the health care might give them effective treatments and to inform the hospitality information for patients. Recent technologies have to be established in health care sector for providing medical service in cost effective ways. These data mining techniques also analyse the factors that are responsible for the cause of disease this include type of food, living conditions, availability of pure water, etc. (Kandwal et al., 2009). Health care organization develops an extensive data which is found difficult to analyse and to make a proper decision regarding patient health, treatment costs and other details of hospitals; in this case, data mining tool is helpful (Sellappan et al.,2008). Artificial intelligence a branch Computer science involves in an
active and more efficient way in health sciences they are used in hospitals and clinics by helping patients and medical experts in making accurate decisions.

One of the methodologies for development in health care is gathering the input data and presenting the output Information, and artificial intelligence uses Healthcare knowledge, data mining here it is considered as a core step for information recovery in health care (Mariscal et al., 2010). Discovering of knowledge in data mining takes health care to an improved level of service quality and also is the main reason by guiding medical practitioners in reducing the adverse drug effect in treatments and also in the reduction of treatment expenses (Venkatadri M and Reddy LC, 2011). This technology is very much used as a user related approach to data. Discovering of knowledge in the database is said as unknown extraction of information from a data source (Yang, H. et al., 2012).

Data mining provides a tool for decision making especially for individual patients this tool identifies valid, useful and understandable data and maintains a high confidential prediction for people, a standard statistical tool has is used for survival (Fayyad et al., 1996).

Kidney disease is increasing, and its prevailing seeks public attention, the high cost of treatment for this illness and the negligence may lead to cardiovascular disease (Levey et al., 2009).

Chronic kidney disease (CKD) is said to affect kidney and also causes damage to the kidney and fails to purify the blood. Kidneys functions decrease up to half its working efficiency it is said as the chronic renal failure. The advanced stage of this is End stage renal disease (ESRD) results in very severe malfunction of the kidney, in this case, the function of the organ is reduced a bit, and the only possibility for survival is either transplantation or dialysis of the kidney.

Haemodialysis and peritoneal dialysis are the dialysis treatments for this disease, haemodialysis did for patients in a clinical environment and peritoneal dialysis performed at the patient's home itself (Zadeh et al., 2006). Kidney disease such as Acute Nephritic Syndrome, Chronic Kidney Disease, Acute Renal Failure and Chronic Glomerulonephritis can be predicted using data mining algorithms such as Support Vector Machine (SVM), artificial neural network (ANN), and Naive Bayes. These algorithms consist of a set of attributes which produce the outcome as a predictive quality and these predominantly focuses on chronic kidney disease which is a life-threatening illness (Mandi et al., 2014).

This paper is further organised as follows: Section 2 represents the related work, Section 3 provides the motivation and section 4 is the proposed methodology. Further section 5 representing the modelling of Fuzzy technique based on Bat optimization, section 6 is the result and discussion, and finally, section 7 provides the conclusion.
2. Related Work

Some of the most recent research related to the chronic kidney disease identification using various techniques of data mining is listed below:

Jams hid Norouzi et al., (Norouzi et al.2016) have proposed an adaptive neuro-fuzzy inference system (ANFIS) for predicting the renal failure timeframe of CKD based on real clinical data, methods used was clinical study records up to 10-year data were collected from newly diagnosed CKD patients. Glomerular filtration rate with a threshold value (15 ccs/kg/min/1. 73 m2) used as a marker for renal failure. Inputs of models such as Variables of age, sex, weight, underlying diseases, diastolic blood pressure, creatinine, calcium, phosphorus, uric acid, and GFR were initially selected for the predicting model. Weight, Diastolic blood pressure, diabetes mellitus as underlying disease including current glomerular comparison is made between the real and predicted values, Adaptive Neuro-fuzzy inference system (ANFIS) is a learning based system based on the neural networks concepts which accurately estimates GFR variation at long future periods.

Christina L Marino et al., (Marino, C.L. 2014) have proposed chronic kidney disease and degenerative joint disease both considered in cats primary study was to determine the prevalence of CKD in two cohorts of cats finding four main age groups namely (RS group) and cats recruited for degenerative joint disease studies (DJD group), the prevalence of chronic kidney disease in these both groups was higher. CKD was common in cats between 1 and 15 years of age, concurrence between CKD and DJD in cats of all ages have been found, this study offers the idea of a relationship and causal commonality between CKD and DJD because of their age groups and lifestyles.

Veenita Kunwar et al., (Kunwar V. et al.2016) have proposed data mining a new trend for obtaining diagnostic results. Health care industries collect a Large amount of unmined data in finding out undiscovered pieces of information for diagnosis and decision making. Data mining techniques like clustering, classification, association, analysis, regression, etc. are explained using naive Bayes and artificial neural network. Prediction of chronic kidney disease using classification techniques by the implementation of a tool, Rapid Miner compares these two classification methods. Naganna Chetty et al., (Chetty, N. et al. 2015) have proposed chronic kidney disease as generally considering as kidney damage and can be measured with GFR rate. Chronic kidney disease patient's classification at their initial stage is a benefit providing treatment either to cure or to prevent this disease. Classification algorithms used here, Wrapper subset attribute evaluator and best first search method to Predict and classify the CKD and non-CKD patients. This subset is a recent model approached to differentiate patients with and without chronic kidney disease. Classifiers perform better on reduce the data compared with the original data.
Ani R et al., (Ani R et al.2016) proposed Chronic renal failure as a significant disease in human life initially chronic renal failure starts with loss of renal function later affects the functioning of the kidney, precautions done are the replacement of kidney or dialysis process. Artificial intelligence is used to learn from complex data sets. Classification techniques like neural network based back propagation (BPN), probability based Naive Bayes, LDA classifier, lazy learner K-Nearest Neighbour (KNN), tree-based decision tree, and Random Subspace Classification algorithms are analyzed, from these the better result giving algorithms used for developing clinical decision support system.

Zeinab Sedighi et al., (Sedighi Z. et al. 2015) chronic kidney disease is a common disease prevented by early detection and cure. Practical guidance needs classification of kidney disease as a global improvement, data mining, and machine learning techniques supports to discover knowledge in identifying patterns for classification.

Feature selection in data mining only defines the useful set of information from raw data which produces quick classified models making it popular in data mining and also in machine learning techniques. A set of filters and wrapper methods in data mining including machine learning techniques are used to classify chronic kidney disease. Feature selection method uses minimum time and appropriately performs the classifications.

3. Motivation

In disease diagnosis, any one of data mining classifier is required for the prediction of disease with the assistance of existing dataset. Maintaining accuracy of the classifier is essential to obtain better performance in medical data diagnosis. The accuracy depends on the operational procedure of the applied classification technique.

So far many researchers have proposed their idea for classification, and some of the recent research is listed in the literature review.

There are various techniques used for the classification such as decision tree, artificial neural network, support vector machine, k-nearest neighbour, fuzzy, etc. Every algorithm has its benefits; however, the accuracy of prediction is not up to the mask. Thus a vacuum exists for the research to develop a better classification technique.

**Primary objective:** The major purpose of this work is to analyse the prediction of Chronic Kidney Disease (CKD) by using the proposed technique.

**Secondary objective:** In this case, the purpose of this work is to develop a dataset with environmental factors for the ease of enhancing the accuracy of the prediction of Chronic Kidney Disease (CKD) by using the proposed technique.
4. Optimal Fuzzy K-Nearest Neighbour (OF-KNN)

Contribution

Chronic Kidney Disease is a kind of global health issue, can cause by several factors including environmental and genetic. In most of the past works, the CKD has analysed and forecasted based on the common factors without much concentration on environmental factors. Hence this work planned to include the environmental factors along with the patient's activities. In medical data prediction, the accuracy must be in acceptable range so that the system can replace a human source. Hence the overall objective of this work is to develop a novel classification technique for the prediction of CKD with better accuracy.

Novel Method

The modelling of a novel CKD predictor is the motivated research topic in this work. An Optimal Fuzzy K-nearest neighbour (OF-KNN) technique is developed for the prediction of CKD. In the OF-KNN technique an optimization algorithm such as Bat is utilized to tune fuzzy then the OF is used to measure similarity in KNN. Thus the proposed technique can be capable of providing better accuracy and acceptable speed.

Data set

In the implementation validation, the CKD dataset is used and serene from the UCI Machine Laboratory. It is composed of 400 instances of 24 +5 attributes. Among the 24 attributes, 11 possess numerical values, and 13 have nominal values. The remaining five attributes are the environmental factors smoking habit, drinking habit, body mass, hereditary kidney diseases and congenital kidney diseases. In this data set, 250 instances are identified as chronic kidney diseases, and the remaining are 150 are not identified. The diagrammatical representation of proposed method is portrayed in Fig 1.
Initially, the test data is fed to the input port of KNN. The KNN acts as the predictor or classifier in our work. The KNN predicts the output class based on the distance between the input data and the training set. In the conventional system, Euclidean distance is applied to find the nearest data in KNN. To attain accurate prediction, the Euclidean distance is changed into the Optimal Fuzzy. The Optimal Fuzzy system can help to find the optimal data which is similar to the test data. This OF-KNN system can provide improved accuracy than the conventional system, especially for the chronic kidney disease.

The steps involved in the proposed Optimal Fuzzy K-nearest neighbour (OF-KNN) technique are described below.

**Steps in Optimal Fuzzy K-Nearest Neighbour (OF-KNN)**

The process flow of the planned methodology is shown in Fig. 2.

**Data Pre-Processing**

In this step, the suitable and usable format of the data is prepared, and then the knowledge extraction is applied (Malik JS et al. 2010). The data pre-processing is utilized to solve the real-world problems associated with the original dataset.
Examples of these problems are: (i) missing, and corrupted data elements (ii) data with noise (iii) different granularity measure of data (iv) dependent and large kind of data with irrelevant information and (v) data form multiple sources.

**Major Tasks in Data Pre-Processing**

**i) Data Cleaning**

The unpredictable, incomplete and noisy data is significantly a part of the dataset. Therefore in data cleaning, the missing values in the dataset are cleared, noise smoothening is carried out, and irregularities of the data is corrected.

**ii) Data Integration**

The Data integration part maintains the coherent nature of the data. The data from different sources are added and stored in a perfect storage. These sources may include multiple data cubes, databases or flat files. While integrating the data, multiple issues are taken into consideration.

The correlation analysis by chi-square test is analysed utilizing the below equation.

\[ \delta^2 = \sum \frac{(Observed \ value - Expected \ value)}{Expected \ value} \]  

The cells that contribute the most to the \( \delta^2 \) value are those whose actual count is very different from the expected count.

**iii) Data Transformation**

The original data is converted and merged into the form which is appropriate for further processing. The smoothening and aggregation are the two major steps involved in data transformation.

a) **Smoothing:** The techniques like clustering the data, regression and binning are worked out to remove the noise from the data applied.

b) **Aggregation:** In aggregation, the instantaneous or accumulated operations are applied to the data.

**Data Reduction**

In order to reduce the volume of the data without affecting the integrity of the original data the data reduction technique is applied. The strategies utilized as data reduction techniques are Aggregation, Sampling, Dimensionality Reduction, Feature subset selection, Feature creation, Discretization and Binarization and Attribute Transformation.

**5. Modelling of Fuzzy Technique based on Bat Optimization**

There are no real strategies for evaluating the predefined shape and extent of membership functions in fuzzy technique. The parameters of the membership functions can be decided based on the requirement of the application. The actual function utilised to in the engineering problems is the triangular membership
functions. The performance upgradation of fuzzy is obtained from the tuning of membership functions. Some of the computational methods were described in the earlier literature. In this paper, the BAT algorithm is utilized to optimize the membership functions (Setnes M. and Roubos H, 2000; Herrera F et al. 1995).

**Problem Definition**

![Triangular Membership Function](image)

Figure 3: Triangular Membership Function

The triangular membership function depicted in the above Fig.3 was chosen as the problem for optimization utilizing BAT. The duration of the membership function to use the BAT algorithm to optimize MF, the problem was set up for a triangular MF (The same argument can be extended to other MF types) as follows: Define the span of the MF using \( a_i, b_i \) and \( c_i \) where, \( a_i \geq x_{minimum} \) and \( b_i \geq x_{maximum} \). The \( x \) value is lies between \( x_{max} \) and \( x_{min} \) and \( c_i \) is denoted as the point of maximum support of the fuzzy set \( i \). If \( c_i \) is the addition of \( a_i \) and \( b_i \) which is multiplied with 0.5 then \( c_i \) is considered as the midpoint.

**BAT Optimization Algorithm**

It comes under the global optimization criteria, and it is a metaheuristic algorithm. This was developed in 2010 by Xin-She Yang based on the below three idealized rules (Yang, 2013).

1. It can differentiate sustenance and obstacles by the magical sensing capability for locating the distance.
2. Inspired by the echolocation behaviour Bat algorithm was developed. The Bat can able to adjust the pulse rate and loudness automatically.
3. The variation of loudness is varied in many ways. Bu the most basic assumption is the value of loudness is varied from large to a small value.

The bat is assumed with a velocity \( \gamma \) and location \( A \) at the defined iteration \( T \). Based on the three idealized rules the frequency is defined as,

\[
F_n = F_{minimum} + (F_{maximum} - F_{minimum})\beta 
\]  

(2)

The \( \beta \) is the random vector from the uniform distribution which lies between 0 and 1. Each bat is associated with the current best solution that is location. The
velocity of each bat is updated by,

$$\gamma^T_n = \gamma^{t-1}_n + \left( A^{T-1}_n - A_{best} \right) F_n$$  \hspace{1cm} (3)

In the above equation $A_{best}$ is the current best solution.

$$A^T_n = A^{T-1}_n + \gamma^T_n$$  \hspace{1cm} (4)

The objective function to evaluate the fuzzy model performance is mean square error. The BAT algorithm is implemented with the below objective function (5).

$$\text{mean square error} = \frac{\sum_{s=1}^{p} \left( h_k - \hat{h}_k \right)^2}{\sum_{s=1}^{q} h_k^2}$$  \hspace{1cm} (5)

In (5) $h_k$ denotes the actual value; $\hat{h}_k$ denotes the expected value; $q$ denotes the number of data points. The actual and expected values are measured corresponding to the data point $k$.

The membership functions in the input and output are optimized so that the performance of the fuzzy model has enhanced with the optimized MFs.

**Optimization of Input and Output Membership Functions**

The same procedure is repeated for both input and output membership function.

Case 1: The membership functions with the expert initial MFs are equated to the number of BATs in the optimization algorithm.

Case 2: In the case of without expert initial MFs, the random generation of all the membership functions are defined.

**Modelling of Proposed OF-KNN Technique for the CKD Prediction**

In the proposed OF-KNN, the BAT optimization algorithm is utilized to tune fuzzy technique then the OF (optimal fuzzy) is used to measure similarity in KNN.

In general, there is no pre-processing is required for the sample set in the data in nearest neighbour classifier. In the original nearest neighbour algorithm the sample from the unknown class is equalised to the class of the nearest neighbour. The fuzzy KNN just deviates from the initial set KNN regarding the result. The membership function is assigned to the sample vector instead of assigning the vector to the particular class. If the vector is confused with two membership values, then the vector is inspected more than one time for the identification its class.
The idea behind the Optimal Fuzzy-KNN is to assign a membership function to the K-nearest neighbours vector distance and those neighbours in possible classes. Let \( S = \{Z_1, Z_2\} \) be the set of the function represents two classes from the CKD dataset. Let the membership function associated with the vector or sample as \( \mu(x) \).

\[
\mu_i(x) = \frac{1}{\|x - Z_i\|^{2/(m-1)}} \frac{1}{\|x - Z_m\|^{2/(m-1)}}
\]

(6)

The above equation holds the variable \( m \) which defines the weight attributed to a distance of each neighbor to the value of membership. As per the equation (6), the membership function of a particular sample is subjective to the neighbor’s distance’s inverse and its membership class. The inverse of the distance is considered as the weight it denotes the relation between the classified sample and the sample to be classified (Rosa JLA and Ebecken NFF, 2002).

6. Result and Discussion

This paper implemented a novel methodology for predicting the chronic kidney disease. This novel method is developed utilizing optimal fuzzy-K nearest neighbour algorithms. Along these techniques, the BAT algorithm is introduced to tune the membership functions of the fuzzy technique. Then this optimal fuzzy is applied for evaluating the similarity measure in KNN. The proposed methodology has implemented in the tool Matlab version of 2016a. The parameters used for the validation of the proposed method are accuracy, precision, recall, specificity, F-Measure and Mean square error. The simulation parameters of BAT optimization algorithm is given in Table 1.

<table>
<thead>
<tr>
<th>Parameters</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of generations</td>
<td>1000</td>
</tr>
<tr>
<td>Loudness</td>
<td>0.5</td>
</tr>
<tr>
<td>Pulse rate</td>
<td>0.5</td>
</tr>
<tr>
<td>Minimum frequency</td>
<td>0</td>
</tr>
<tr>
<td>Maximum frequency</td>
<td>2</td>
</tr>
<tr>
<td>Number of dimensions</td>
<td>20</td>
</tr>
</tbody>
</table>

Accuracy

The prediction ability of classifier is measured from the performance parameter accuracy (Baldi, P et al. 2000). In machine learning, the accuracy of the classifiers is estimated using (7).

\[
\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN}
\]

(7)
From the accuracy comparison shown in Fig. 4, the proposed method with other methods the OF-KNN is 1.4% greater than SVM, 8% greater than KNN and 27% greater than the artificial neural network.

**Precision**

Precision is the measure defined by the ratio of true positive to the addition of true positive and false positive. The precision comparison is given in Fig. 5. It can be estimated as,

$$Precision = \frac{TP}{TP + FP} \quad (8)$$
The proposed classification method provides the precision of 4. The ANN classifier provides the accuracy of 75% with a precision of 0.3765. The precision for the SVM is 1 for the accuracy of 95.6%. The KNN method holds a precision value of 3 with accuracy 89%.

Recall

The recall is another measure for the performance evaluation of classifiers, and it is otherwise known as sensitivity. It is evaluated as by below equation (9) and shown in Fig. 6.

\[
\text{Recall} / \text{Sensitivity} = \frac{TP}{TP + FN}
\]  

(9)

Where, TP is the indication of the positive sample which is identified properly by the classification technique used. The FN is evaluated by positive sample mistakenly identified as the negative sample by the classifier.

Specificity

The specificity can be evaluated using the equation (10). TN is the identification of negative sample equally by the classification technique used. The FP is evaluated by negative sample mistakenly identified as the positive sample.

\[
\text{Specificity} = \frac{TN}{TN + FP}
\]  

(10)
From Fig. 7 the specificity value achieved by classifier support vector machine is 1, ANN is 0.6 and the other classifier KNN provides the value of 3. The specificity of proposed optimal fuzzy-KNN classifier is 4.

**F-Measure**

The F-measure is a parameter related to precision and sensitivity. The F-measure can be defined as,

$$F\text{-measure} = 2 \times \frac{\text{precision} \times \text{sensitivity}}{\text{precision} + \text{sensitivity}} \quad (11)$$

In the F-Measure of classifiers is nearest to zero it indicates the stumpy performance. The F-measure of the optimal fuzzy-KNN is 1.6, and the lowest value is indicated by artificial neural network classifier is 0.416 from the evaluation represented in Fig. 8.
Thus the validation of the proposed method is done utilizing the above parameters. From the analysis, the proposed method provides better classification accuracy for the two class of kidney diseases dataset. Additionally, the environmental factors are added to the dataset to enhance the classification accuracy of the proposed method, and it is the novelty of the proposed method.

7. Conclusion

This paper proposes a different approach for prediction of chronic kidney diseases with a modified dataset with five environmental factors. The methodology introduced in the paper for the prediction is OF-KNN (Optimal Fuzzy-K nearest Neighbour). The planned work has proved as the better classifier with two different classes of diseases with perfect classification rate. The comparative analysis has done with artificial neural network, support vector machines and K nearest neighbours. The performance metrics utilised for the analysis are accuracy, precision, recall, specificity and F-Measure. From the analysis, it has concluded that the proposed method is well organised to obtain the perfect classification. Thus for diagnosis of CKD, the proposed OF-KNN machine learning tool is resulting in high classification accuracy rate.

References


