JANOLI International Journal of Artificial Intelligence and Applications
(A publication unit of Shanti Educational Research Foundation)

Comparative Study of Machine Learning Techniques for Prediction of Kidney Disease

Mainka Saharan and Pradeep Upadhyay
NIMS University, Jaipur
Email: pu982712@gmail.com

Abstract: As kidney chronic disease is nowadays widely increasing which either caused by kidney disease or reduce the function of the kidney, it also affects the cardiac problems- scientifically which can lead to sudden heart attacks at the end-stage. Early diagnosis and adequate therapies can only help in stopping this disease, where dialysis and kidney transplantation is the only way to save the life of the patient. Detecting kidney disease through machine learning and through data mining techniques which can reveal the hidden problem of the kidney. Therefore, the current article is based on the comparative study using various Machine Learning techniques to detect kidney disease. This survey supports to find the accuracy of algorithms which are more useful to find the kidney disease in early stage. The comparative study of all the algorithms and by implementing the models on different platforms, and it is analyzed that which is the best algorithm to predict CKD (Chronic Kidney Disease). The machine learning techniques are compared like Probabilistic Neural Network (PNN), Multilayer Perceptron Algorithm (MLP), Logistic Regression (LOGR), Regression Tree (RPART), Support Vector Machine (SVM) and Radial Basis Function (RBF).

Keywords: Classification, Machine Learning, Kidney disease Detection, Feature Extraction, Data Mining Technique

Introduction

Data mining strategies include association rule mining to find common patterns, forecasting, grouping, and clustering. The data mining techniques are more useful especially in the prediction of heart disease and kidney disease. Data mining was used to discover patterns from the vast quantity of collected information and then to create predictive models. There is a large amount of data to be interpreted in the medical field. The mining of medical data improves the quality of patient care and the prediction of medical trends.

It is a chronic condition associated with the condition of being diseased and mortality, it may have a high risk of heart disease and it may cost a huge amount for recovery. To stay alive, there are millions of people worldwide sufferings from this disease and only a small percentage of people get the treatment to live. The main objective is to predict the best technique and accuracy, sensitivity and specificity of the dataset.
The technique is been used to reveal and extract hidden problems from the patients so that it may be easier for a physician to maximize the accuracy in the identification of the disease. Among the 24 parameters, they only will have seven parameters they are: Hemoglobin, albumin, diabetes, blood glucose, creatinine, and pus cells.

They have predicted the CKD (Chronic kidney disease) in many different forms by using different algorithms. They have seen that mostly the UCI repository data set is used for the prediction. They compare all the algorithms and the different models used to predict CKD. They get to know that PNN, random subspace, boosted decision tree, ANN, Naive Bayes gives the best result from other algorithms. In other papers also prediction is done on the same dataset and the accuracy was taken out by using different algorithms. In an another paper, naive Bayes gives the best result among other algorithms.

Rest of the paper is organized as follows. Section 2 provides literature review work, Section 3 presents technical discussion of different machine learning algorithms, the comparison model is considered in the section 4, Experimental result is illustrated in section 5. Finally Section 6 presents conclusion and future work.

Literature Review

The aim of our survey is to find out the best technique based on the accuracy. The current research compares the mining technique like probabilistic neural network (PNN), multilayer perceptron algorithm (MLP), logistic

regression (LOGR), regression tree (RPART), Support vector machine (SVM) and Radial Basis Function (RBF) and many different algorithms also.
In the research , author have taken the probabilistic neural network are a kind of neural network of Radial Basis Function which is an algorithm for one-pass processing and very linear design. The strength of it is fast testing and simple adjusting. It consists of 4 layers as the input layer, pattern layer, summation layer, and output layer. MLP is one of the main neural network types, consists of an input layer; it may contain 1 or more hidden layers and output layer. It is been successfully done to solve complex or differentiated by the best algorithm that is the error backpropagation algorithm, to train them in a supervised manner.

SVM is the method for both relevant and irrelevant classification. It uses a nonlinear mapping to restore to a higher dimension the unique training data. It examines in a "decision boundary" the linear optimal hyperplane separating relevant data from another. The hyperplane is been used to separate the wanted and unwanted data [1]. Another tool i.e. RBF also consider for the same problem. RBF is a neural network algorithm that needs fewer networks learning computation time. It contains 3 layers: data layer, hidden layer, and output layer. These nodes are been connected within each network. Input variables are passed from the input layer without any weights directly to the hidden layer
There are 5 stages of the kidney but they take only 2 stages and measured by eGFR(estimated Glomerular Filtration Rate) formula. In this creatinine (It is the waste product which is removed from the body and filter the blood and release into urine) is taken, the age of the female is been taken. They have done the testing and validation of different algorithms at different stages.

In a paper , authors have investigated the kidney chronic disease with the help of some methods of machine learning. In this study, they have predicted the methodology in which they compare the data parameters and the attributes. In machine learning, they will test the classifiers by 4 methods by which we will get the accuracy, sensitivity, and specificity.

It contains the MI methods which are regression tree (RPART), Support Vector Machine (SVM), MLP and Logistic Regression (LOGR). They compare them based on sensitivity, accuracy, precision, specificity, and error. By the help of dataset, the information which they have to analyze have following things, which shows the relationship between input parameters which will reduce the number of parameters needed for prediction of CKD and will remove unwanted parameters, it is also analyzed with the help of urine test and blood test for diagnoses, and also used to analyze the best method to overcome from CKD early. The dataset has some missing values so it is represented in the form of the matrix, it contains parameters and participants. In this, the MI uses regression analysis for filling the missing values. The parameters derived from a urine test and blood test states that in blood test the plasma volume gets down and the red blood cells increases. They started with the 24 parameters and will become over with only 7 subset parameters. In this, the highest sensitivity is achieved by the MLP and LOGR and the other methods are followed. This method is more stable and shows the best result. As per the F1 score, it is identified that the MLP and LOGR are the best and RPART overcomes the SVM model.

In the research of , authors have seen that kidney diseases are also affecting the cardiac diseases which are known as cardiorenal syndrome. When a patient goes through the ECG that CKD can also undergo which can be a major issue for the patient. The database is been collected from the PTB and fantasia. It's between the old age from 50 to 70 years. In this the first step will be doing is that extracting digitalized ECG data from the database. To find the best features we have to use the QT(it is used in ECG and the wave start from Q and ends with T wave) and RR (it is the time elapsed between two R waves) intervals. The author have used supervised learning because data set contains some pre label data and they have used SVM. As the SVM is best for classification tool and gives the best accuracy at less time.

In this research paper they have come to know about other data mining technique and they are as: ANN, SVM, KNN, RBF, and random subspace were used in the dataset consist of 400 samples and 24 attributes and with particle swarm optimization. The result is been compared with the test results. The highest accuracy was of random subspace mining. In this paper, they have used many different algorithms with their authors and then they will predict which is best. Random subspace is used because in this the smaller parts are trained and it uses the subspace for actual data size. It is used in large datasets attributes. These algorithms reduce the time, over learning and easy to understand, simple .All these data mining algorithms are applied with the help of PSO (Particle Swarm Optimization), random subspace has the best accuracy.

The paper , presented the study using WEKA tool. They have identified that which is the best among J48 and random tree by using WEKA Tools. It a powerful tool that is used for the classification model. It is the collection of algorithms for data mining tasks. It has 4 modes to do work: Simple CLI, Explorer, Experimenter, Knowledge flow. In this paper, they have used the decision tree also for classification of relationships and to identify subgroup differences .J48: By using the information gain and examines the same for the result of an attribute for splitting the data from the training data set to make a decision tree. After this, it requires smaller subsets. If all instances contained in the same class then the procedure stops. J48 is also known as C4.Random tree: It is the method for constructing a tree that needs k random features at each node. Weka generates full classification for each node. Powerful and accurate, good performance on . This paper is classified on the bases of blood groups in different regions of Gujarat.

In this paper author have explored the different association rule algorithm i.e. Apriori association rule mining, filtered, Tertius, predictive. Among all Apriori association rule mining is best. They have found the pattern of system usage by teacher and student for developing the learning system model is used for data. By this, they can prevent different types of cancers. By comparing all 4 we predict apriori is best and it is best because the output it presents is all yes.

The research is based on the comparison of the accuracy on various algorithms like Apriori, Tertius. All the algorithms are compared on the different measures like support, confidence and predictive accuracy. Support measure frequency; trust measure strength and predictive precision is used to measure the concern of generated rule. They can take the mean, median and mode from this algorithm.

The author suggests that as CKD affect the body and the treatment is at high cost. The study uses the UCI datasets by using the Bayesian classification algorithm, KNN results are obtained by classifying algorithms and SVM. The classification is done on basis of raw data, gain ratio, relief. The data pre-processing is done in which feature selection algorithm is done on basis of relief and gain ratio. For the classification the algorithm used are SVM, KNN, naïve bayes it is done on the 10 floor cross validation method. The best result is given by KNN.

In this author contains the same dataset from UCI machine learning repository in this they have to predict the CKD by a boosted decision tree, deep support vector machine. And by this, they predict the best accuracy by boosted decision tree it is robust and not over fit in training data. Their model can work with any kind of data set. Their model is based on cloud platform because of the fast evaluation of data. It is implemented with the help of the Azure platform in machine learning .

This paper tells that they have unwanted data which is collected from the healthcare industry to decide for the diagnosis of diseases, in this they will use the algorithms of KNN and Naive Bayes produces more result than KNN. Classification is supervised learning assign objects into classes. In Naive Bayes the dimensionality is high and in these variables is independent on each other. The Bayes theorem provides the favorable outcomes done upon the total no of outcomes. The algorithm is run on the rapid miner tool .

This paper also tells about the prediction of kidney disease through the implementation of data mining algorithms. Algorithms are as follows back propagation, neural network, naive Bayes, decision table, decision tree, KNN and one rule classifier in this also naive Bayes is best. In this, no classifiers are used to predict the CKD. ANN is a nonlinear approach that uses spread back to learn with one or more hidden layers. Naive Bays is also been used and J48 is used which helps in making binary tree acc to divide and conquer concept to divide data. In this decision, a table is used for visualizing and an inducer. There are three components action row, condition row, and rules. It was done on the weka tool on this it is classified based on the instance, sensitivity, specificity, time cost, mean absolute error and ROC area. The six algorithms is used in its naive Bayes, multilayer perception. J48, KNN, one rule, and decision table. Naive Bayes is best of all.

Comparison Model

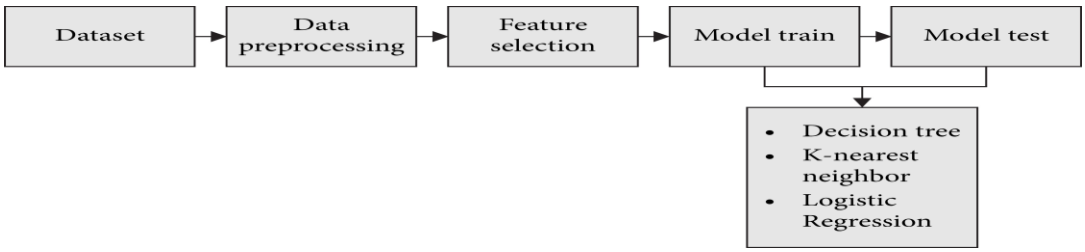

Fig.1. The Machine Learning Model.

PNN: It is a neural feed forward that is used in problem of classification and pattern recognition problem. It is faster and more accurate then multilayer perceptron networks. It approaches Baye's optimal classification and them relatively insensitive to outliers. The network contains the four layers. Input layers, pattern layer, summation layer, output layer. PNN is slower then MPN for classifying of new cases
It requires more space to store the model.

There are data patterns yet which has it predefined classes h=1…H, the probability of y belonging to class h equals rg, the cost is y CG , the probability density function is x1(y),x2(y),x3(y)….x g(y).[12]

MLP: It is a neural network connecting multiple levels in a directed graph which states a single path will go only to one path. It is used in supervised learning. It is a finite acyclic graph. It contains nodes which are neurons with logistic activation. By it, we can calculate the complex functions by combing many neurons. There are three layers in it: an input layer, hidden layer, output layer.

Logistic Regression: It is going to a method for binary classification problems, it uses a logistic function to model the binary dependent variable and it contains the two values 0 or 1, true or false. It is used to predict the risk of developing the disease. In regression analysis, logistic regression is estimating the parameters of the logistic model.

Regression Tree: It is the iterative process that splits data into branches or partition. It has target values. Deciding on regression is much easier than another method. We can continue to splits each branch into smaller groups.

Support Vector Machine: It is a supervised learning model related to learning algorithms that are used for classification and regression. It constructs the hyperplane in high dimensional space. It separates the linear and nonlinear points. In this, the straight-line equation is been used y = mx+c, and for mid hyper planes, mx+c will be equal to 0, when mx+c will be 1 then all positive values will be there, and when mx+c equals to -1 then negative values.

Radial Basis Function: A radial basis function is a real-valued function which is defined as $\phi:[0,\infty)\rightarrow R$. It is a simple single-layer type of artificial neural network. It gives the approximate value of the given functions

Experimental Result

DATASETS: UCI machine repository contains a data set . The dataset of Apollo hospital India which has the 24 parameters and classified as non-CKD and CKD. Another dataset from online from PTB and fantasia.

Using the DTREG Predictive Modelling Program, the following analysis was done. The comparative analysis of the algorithms used was made on the basis of classification reliability and execution time performance measurements. A K-fold cross validation technique was used for model testing and validation. During the study, incomplete predictor parameter values have been replaced by press.

ALGORITHM	ACCURACY	RESULT

SVM	84.7%	Morphological operators are used along with canny edge detection method. After that the classification is done on the basis of SVM.
PNN	96.6%	For fast testing and simple adjusting PNN is been used
MLP	51.5 - 98%	Solve complex problems differentiate the problems
RBF	87%	It requires the less leaning computation time
RPART	95%	It helps in the splitting of data into branches
LOGR	98%	It signifies the features

Table 1: Comparison of accuracy of Various Machine Learning Algorithm

Conclusion

Finally, the PNN provides the highest percentage of overall detection accuracy. In this MLP, they require a minimum time interval. This algorithm is classified on reliability which is defined on the patient level. The PNN defines the best accuracy and efficiency prediction. In these days, and is a globally leading which causing the high death rate. It is because of the last stages of CKD may lead to the CVD. About CRS, many people are suffering from the cardiac disease may suffer from, CKD suffered patient may suffer from cvd whose treatment is limited. So that why the models are been classified to detect the disease from their digitized ECG in the early stages using algorithms.

References

- Rady, El-Houssainy A., and Ayman S. Anwar. "Prediction of kidney disease stages using data mining algorithms." Informatics in Medicine Unlocked (2019): 100178.
- Aljaaf, Ahmed J., et al. "Early Prediction of Chronic Kidney Disease Using Machine Learning Supported by Predictive Analytics." 2018 IEEE Congress on Evolutionary Computation (CEC). IEEE, 2018.
- Rahman, Tahsin M., et al. "Early Detection of Kidney Disease Using ECG Signals Through Machine Learning Based Modelling." 2019 International Conference on Robotics, Electrical and Signal Processing Techniques (ICREST). IEEE, 2019.
- Kemal, A. D. E. M. "Diagnosis of Chronic Kidney Disease using Random Subspace Method with Particle Swarm Optimization." UluslararasıMühendislikAraştırmaveGeliştirmeDergisi 10.3: 1-5.
- Solanki, Ashok kumar Vijaysinh. "Data mining techniques using WEKA classification for sickle cell disease." International Journal of Computer Science and Information Technologies 5.4 (2014): 5857-5860.
- Aher, Sunita B., and L. M. R. J. Lobo. "A comparative study of association rule algorithms for course recommender system in e-learning." International Journal of Computer Applications 39.1 (2012): 48-52.
- Mazid, Mohammed M., ABM Shawkat Ali, and Kevin S. Tickle. "Finding a unique association rule mining algorithm based on data characteristics." 2008 International Conference on Electrical and Computer Engineering. IEEE, 2008.
- Kayaalp, Fatih, Muhammet Sinan Basarslan, and Kemal Polat. "A Hybrid classification example in describing chronic kidney disease." 2018 Electric Electronics, Computer Science, Biomedical Engineerings' Meeting (EBBT). IEEE, 2018.
- Ranjan, Sourav, et al. "CHRONIC KIDNEY DISEASE RISK PREDICTION BASED ON MACHINE LEARNING TECHNIQUE USING CLOUD PLATFORM."
- Sunil, D., and B. P. Sowmya. "Chronic Kidney Disease Analysis using Data Mining." (2017).
- Alasker, Haya, et al. "Detection of kidney disease using various intelligent classifiers." 2017 3rd International Conference on Science in Information Technology (ICSITech). IEEE, 2017.
- Kusy, Maciej, and Roman Zajdel. "Probabilistic neural network training procedure based on Q (0)-learning algorithm in medical data classification." Applied Intelligence 41.3 (2014): 837-854.
- www.physionet.com, last accessed 2020/15/02