International Journal of Research in Engineering, Science and Management
Volume-2, Issue-3, March-2019
www.ijresm.com | ISSN (Online): 2581-5792

Diagnosis of Chronic Kidney Disease using Random Forest Algorithms

S. Venkata Lakshmi[1], M. K. Meena[2], N. S. Kiruthika[3]
[1]*Assistant Professor, Department of Computer Science, K L N College of Engineering, Sivagangai, India*
[2,3]*Student, Department of Computer Science, K L N College of Engineering, Sivagangai, India*

***Abstract*: Machine learning (ML) is a category of algorithm that allows software applications to become more accurate in predicting outcomes without being explicitly programmed. Chronic kidney disease (CKD) is a slow and progressive loss of kidney function over a period of several years. Medical tests for other purposes sometimes contain useful information about CKD disease. Attributes of different medical tests are investigated to identify what attributes contain useful information about CKD. A database with several attributes of CKD are analysed with different techniques. Common Spatial Pattern (CSP) filter and Linear Discriminant Analysis (LDA) are first used to identify the dominant attributes that could contribute in detecting CKD. Classification methods are used to identify the dominant attributes. These analyses suggest that haemoglobin, albumin, specific gravity, hypertension and diabetes mellitus together with serum creatinine are the most important attributes in the early detection of CKD. Further, it suggests that in the absence of the information of hypertension and diabetes mellitus, the attributes blood glucose random, and blood pressure may be used. The main objective of this project is to determine the kidney function failure by applying the Random Forest algorithm and to classify the chronic and non-chronic kidney diseases.**

***Keywords*: Machine Learning, Random Forest Classification, Chronic Kidney Disease.**

1. Introduction

A. Machine learning

Machine learning is an application of artificial intelligence (AI) that provides systems the ability to automatically learn and improve from experience without being explicitly programmed. Machine learning focuses on the development of computer programs that can access data and use it learn for themselves.

The process of learning begins with observations or data, such as examples, direct experience, or instruction, in order to look for patterns in data and make better decisions in the future based on the examples that we provide. The primary aim is to allow the computers learn automatically without human intervention or assistance and adjust actions accordingly.

Some machine learning methods:

Machine learning algorithms are often categorized as supervised or unsupervised.

Supervised machine learning algorithms can apply what has been learned in the past to new data using labelled examples to predict future events. Starting from the analysis of a known training dataset, the learning algorithm produces an inferred function to make predictions about the output values. The system is able to provide targets for any new input after sufficient training. The learning algorithm can also compare its output with the correct, intended output and find errors in order to modify the model accordingly.

In contrast, unsupervised machine learning algorithms are used when the information used to train is neither classified nor labelled. Unsupervised learning studies how systems can infer a function to describe a hidden structure from unlabelled data. The system doesn't figure out the right output, but it explores the data and can draw inferences from datasets to describe hidden structures from unlabelled data.

Semi-supervised machine learning algorithms fall somewhere in between supervised and unsupervised learning, since they use both labelled and unlabelled data for training – typically a small amount of labelled data and a large amount of unlabelled data. The systems that use this method are able to considerably improve learning accuracy. Usually, semi-supervised learning is chosen when the acquired labelled data requires skilled and relevant resources in order to train it / learn from it. Otherwise, acquiring unlabelled data generally doesn't require additional resources.

Reinforcement machine learning algorithms is a learning method that interacts with its environment by producing actions and discovers errors or rewards. Trial and error search and delayed reward are the most relevant characteristics of reinforcement learning. This method allows machines and software agents to automatically determine the ideal behaviour within a specific context in order to maximize its performance. Simple reward feedback is required for the agent to learn which action is best; this is known as the reinforcement signal.

The kidneys function as blood filters that drain waste products while retaining other valuable blood contents like proteins. If these filters are damaged, they initially may become "leaky," and substances like proteins can seep from blood into urine. At later stages, these filters slowly shut down and lose their ability to filter. When kidney impairment lasts for more than 3 months, it is called chronic kidney disease. This process ultimately results in decreased urine production and kidney failure, with build-up of waste products in the blood and body

tissues. One common reason for kidney failure in the United States is diabetes.

Sometimes chronic kidney disease is accompanied by high blood pressure, which not only can be caused by kidney damage but also further accelerates kidney injury and is a major reason for the negative effects of chronic kidney disease on other organs, including increased risk of heart disease and stroke, collection of excess body fluids, anaemia, weakening of bones, and impairment of the way the body eliminates medications.

In addition to diabetes and high blood pressure, age >60 years, female sex, African American ethnicity, obesity, high cholesterol, lack of physical exercise, smoking, and excessive salt intake are factors that increase risk of kidney disease. Other contributing circumstances include infections or inflammatory diseases that affect the kidneys; inappropriate use of medications like aspirin, ibuprofen, and other painkillers; and use of herbal supplements that are known to cause damage to kidneys. Also, imaging studies that use iodine contrast substances can have a negative effect on kidneys.

Chronic kidney disease develops slowly, with few symptoms. It is often not recognized until the disease is advanced. If it is detected early, treatment can slow down or avoid kidney function decline and diminish the negative effects on other body functions. A blood test measuring glomerular filtration rate assesses how well the kidneys clear the blood of a waste product called creatinine. A value of 60 to 90 may be an early sign of kidney disease; a value below 60 is usually considered abnormal. A test using a urine sample evaluates the presence of protein (albumin) in the urine; repeated results of 30 mg or more can indicate a problem. High blood pressure may also point to underlying chronic kidney disease.

Human body organs are interconnected with each other, so if one organ does not work properly then there will be symptoms due to this improperness. When the kidney is not working properly this would cause some changes in attributes such as serum creatinine, blood pressure, blood sugar and haemoglobin. Therefore, this correlation among the attributes can be used to identify CKD. Doctors inherently use these attributes and their inter-relationships from reports such as blood reports and urine reports to identify the diseases.

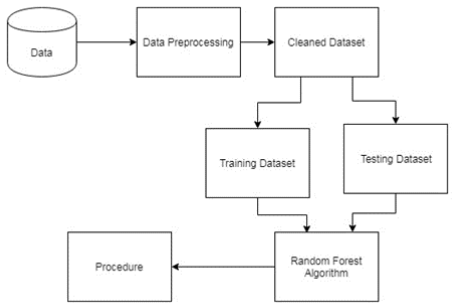

Fig. 1. System architecture

2. Dataset collection

The kidney disease dataset has been used for analysis of kidney disease. This dataset contains four hundred instances and twenty five attributes are used in this comparative analysis. Out of 25 attributes 11 attributes are numeric and 14 attributes are nominal type. The attributes in the dataset are Age, blood pressure, specific gravity, albumin, sugar, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose random, blood urea, serum creatinine, sodium, potassium, hemoglobin, Packed cell volume, White Blood Cell Count, Red Blood Cell Count, hypertension, Diabetes Mellitus, appetite, Pedal Edema, Anaemia, and Class. The class distribution is in two classes' chronic kidney dieses and not chronic kidney dieses; these are interpreted as "ckd" and "notckd" respectively.

3. Pre-Processing

Data pre-processing is a way to convert the noisy and huge data into relevant and clean data, as the data available is Real world data, so it contains inaccurate data, missing values and other Noisy data, for removing this inconsistent data from the Dataset, the proposed system have to clean the raw data. This is an important part to complete the prediction model. It reduces the dimensionality and helps the machine to achieve better results. This is one of the most time consuming stage in building a classification model.

Following data pre-processing steps are followed:

A. Looking up for proper format

As we have made our model using python, so we need a csv file (comma separated value) for our code. The data downloaded is in the form of RAR file, so we extract the data from the text file available and save it into a csv file so that our python code can read it. This is the first most important step, if the data is not available in requires format then we cannot design the classification model.

B. Finding Missing Values

When the data collected is real world data, and then it will contain missing values. This brings more change in the prediction accuracy. Sometimes these missing values can be simply deleted or ignored if they are not large in number. It is the simplest way to handle the missing data but it is not considered healthy for the model as the missing value can be an important attribute contributing to the disease. The missing values can also be replaced by zero this will not bring any change as whole, but this method cannot be much yielding. So an efficient way to handle missing values is to use mean, average of the observed attribute or value. This way we lead to more genuine data and better prediction results.

C. Data transformation:

In this step we transform the given real data into required format. The data downloaded consist of Nominal, Real and Decimal values. In this step we convert the Nominal data into numerical data of the form 0 and 1. The positive value is

assigned the value of 1 and the negative value is assigned the value of 0. Now the resultant csv file comprises of all the integer and decimal values for different CKD related attributes.

4. Feature selection

In this step we select subset of relevant attributes from the total give attributes. This stage helps in reducing the dimensionality and making the model simpler and easy to use, thus leading to short training time and high accuracy. To obtain highly dependent features for CKD prediction we have used Correlation and dependence method. The term correlation can be defined as mutual relationship between two. In this those attributes are chosen which highly influence the occurrence of Chronic Kidney Disease. By using the correlation it is found that 10 attributed were highly correlated to the occurrence of CKD from the total of 25 attributes.

The 10 attributes selected from a total of 25 attributes are:

- Specific gravity
- Albumin
- Sugar
- Blood Glucose Random
- Serum Creatinine
- Potassium
- Packed Cell Volume
- White Blood Cell
- Red Blood Cell
- Diabetes millitus

5. Dataset training

The dataset is divided into two sub datasets both containing 25 attributes.

- *Training data:* Training dataset is derived from main dataset and it contains 300 out of 400 records in main dataset of CKD.
- *Testing data:* Testing dataset is of 100 out of 400 records from main CKD dataset.

6. Random Forest Classification

The random forest is an ensemble approach that can also be thought of as a form of nearest neighbour predictor. Ensembles are a divide-and-conquer approach used to improve performance. The main principle behind ensemble methods is that a group of 'weak learners' can come together to form a 'strong learner'. The random forest starts with a standard machine learning technique called a 'decision tree' which, in ensemble terms, corresponds to our weak learner. The decision tree algorithm repeatedly splits the data set according to a criterion that maximizes the separation of the data, resulting in a tree-like structure. In this algorithm an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. The random forest takes this notion to the next level by combining trees with the notion of an ensemble. Thus, in ensemble terms, the trees are weak learners and the random forest is a strong learner. The advantages of a random forest classifier are that its' runtimes are quite fast, and that it is able to deal with unbalanced and missing data. Weaknesses of this algorithm are that when used for regression it cannot predict beyond the range in the training data, and it may over-fit data sets that are particularly noisy.

7. Diagnosis and Prediction

Kidney disease can be diagnosed with important attributes like serum creatinine, albumin, potassium, specific gravity, sugar, blood glucose random, white blood cell, red blood cell, packed cell value and diabetes millitus. Each attributes have some range, if the patients having the rate below the range may have kidney disease. The random forest classifier is build using multiple decision trees. To perform prediction using the trained random forest algorithm uses the below pseudo code.

- Takes the test features and use the rules of each randomly created decision tree to predict the outcome and stores the predicted outcome (target)
- Calculate the votes for each predicted target.
- Consider the high voted predicted target as the final prediction from the random forest algorithm.

To perform the prediction using the trained random forest algorithm we need to pass the test features through the rules of each randomly created trees. Each random forest will predict different target (outcome) for the same test feature. Then by considering each predicted target votes will be calculated. Based on that voting, the occurrence of the disease will be predicted.

8. Conclusion

By knowing the important attributes in detecting CKD, may provide a better opportunity in diagnosing individuals who do not exhibit explicit symptoms of the disease. These attributes can often be determined from medical tests taken for other purposes and may lead to CKD-specific tests. A weighting vector based on CSP filter and LDA analysis, and then classification analysis using LDA and KNN classifiers, were used to identify the dominant attributes. It was found that hemoglobin, albumin, specific gravity, serum creatinine, hypertension, and diabetes mellitus are the most important attributes in detecting CKD. Further, these analyses suggest that, when hypertension and diabetes mellitus are not available, random blood glucose checks and blood pressure tests can be used. This work is focused on predicting CKD status of a patient with high accuracy. We have analyzed 25 different attributes related to CKD patients and predicted accuracy for Random Forest algorithm. In conclusion, we predict that the patients may have chronic disease or not.

9. Future enhancement

The random forest is an ensemble approach that can also be thought of as a form of nearest neighbour predictor. Ensembles

are a divide-and-conquer approach used to improve performance. The main principle behind ensemble methods is that a group of 'weak learners' can come together to form a 'strong learner'. The random forest starts with a standard machine learning technique called a 'decision tree' which, in ensemble terms, corresponds to our weak learner. The decision tree algorithm repeatedly splits the data set according to a criterion that maximizes the separation of the data, resulting in a tree-like structure. In this algorithm an input is entered at the top and as it traverses down the tree the data gets bucketed into smaller and smaller sets. The random forest takes this notion to the next level by combining trees with the notion of an ensemble. Thus, in ensemble terms, the trees are weak learners and the random forest is a strong learner. The advantages of a random forest classifier are that its' runtimes are quite fast, and that it is able to deal with unbalanced and missing data. Weaknesses of this algorithm are that when used for regression it cannot predict beyond the range in the training data, and it may over-fit data sets that are particularly noisy.

References

[1] Anandanadarajah Nishanth, and Tharmarajah Thiruvaran, "Identifying important attributes for early detection of Chronic Kidney Disease," IEEE Reviews in Biomedical Engineering, Volume 11, 2017.
[2] Jayalakshmi, V. Lipsa Nayak and Dharmarajan, K, "A Survey on Chronic Kidney Disease Detection Using Novel Methods," Volume 119, pp.125-130, 2018.
[3] Pallavi Sharma, Gurmanik Kaur, "Review on Data Mining Techniques for Prediction of Chronic Kidney Disease," vol. 63, no. 1, September 2018.
[4] Pushpa M. Patil, "Review on Prediction of Chronic Kidney Disease using Data Mining Techniques," IJCSMC, Vol. 5, Issue. 5, pp.135 – 141, May 2016.
[5] Ramya, S. Radha, N, "Diagnosis of Chronic Kidney Disease Using Machine Learning Algorithms," Vol. 4, Issue 1, January 2016.
[6] Ravindra, B. V. Sriraam, N. Geetha, M., Classification of non-chronic and chronic kidney disease using SVM Neural networks," in International Journal of Engineering & Technology, vol. 7, pp.191-194, 2018.
[7] Sirage Zeynu, Shruti Patil, "Prediction of Chronic Kidney Disease Using Feature Selection and Ensemble Method," Volume 118, No. 24, 2018.
[8] Suman Bala, Krishan Kumar, "A Literature Review on Kidney Disease Prediction using Data Mining Classification Technique," IJCSMC, Vol. 3, Issue. 7, pp. 960 – 967, July 2014.