National Conference on Engineering Innovations and Solutions (NCEIS – 2018)
International Journal of Scientific Research in Computer Science, Engineering and Information Technology
© 2018 IJSRCSEIT | Volume 4 | Issue 6 | ISSN : 2456-3307

Data Mining Techniques Used To Predict Chronic Kidney Disease

Chithra A G, Chandana B, Darshan R, Harshitha H S, Nasreen Fathima
Department of Computer Science & Engineering, ATME College of Engineering, Mysuru, Karnataka, India

ABSTRACT

Chronic kidney disease is a global health issue and area of concern, associated with an increased risk of cardio vascular diseases and chronic renal failure[1]. It is a symptom where kidney fails to filter toxic wastes from the body, which results in decomposition of wastes in human body and leads to dangerous results. The two main causes of this disease are diabetes and high blood pressure, which are responsible for up to two-third of cause[5]. The healthcare sector has huge medical data but the main difficulty is how to cultivate the existing information into useful practices[3]. To unfold this hurdle the concept of data mining is best suited. The main objective of this paper is to use data mining technique such as random forest, RBF, K-means clustering and Naïve Bayes for the prediction of chronic kidney disease and to summarize the efficiency of Naïve Bayes method by generating suitable results.

Keywords: Data mining, Classification, Chronic Kidney disease, Random forest, RBF, K-means clustering.

I. INTRODUCTION

Data mining is a practice of examining large pre-existing databases in order to generate new information[10]. Data mining is gaining popularity in disparate research fields due to its application and approaches to mine the data in an appropriate manner which improves prediction and reduces cost[9]. Hence we are using this technique to predict chronic kidney disease.

Chronic kidney disease which is also called as renal failure is slow continuous loss of kidneys functionality over a time of several years[1]. CKD has become a major public health problem[2]. The disease comprises circumstances that harm kidney and reduce its ability to keep us healthy[5].

The National Kidney foundation determines the different stages of chronic kidney disease based on the presence of kidney damage and glomerular filtration rate (GFR), which measures a level of kidney function prediction beginning with the identification of symptoms in patients and then identifying patients who are suffering from CKD among huge patient's record[5]. Thus, the prime objective of this paper is to organize the data from CKD dataset using classification techniques to predict class accurately in each case.

II. METHODOLOGY

Data Mining is one of the most significant stages of the Knowledge Data Discovery process[15]. The process involves data collection from various sources with preprocessing of the chosen data. The data is then transformed into suitable format for further processing. Various data Mining technique are applied on the data to extract valuable information and evaluation is done at the end[15]. Some of the techniques are discussed below:

1.Random forest Algorithm

The random forests algorithm for prediction or classification task can be explained as follows:

i. Using original samples data draw n tree bootstrap.
ii. For each of the bootstrap sample, produce an unpruned classification tree, by following modification:
    At each node, instead of choosing the best split among all predictors, arbitrarily sample m try of the predictors and select the best split among those variables.
iii. Predict new data by aggregating the predictions of the ntree trees using majority votes for classification.

An estimation of the error rate can be found, based on the training data, by the following steps:

i. At every bootstrap iteration, predict the data not in the bootstrap sample (what Breiman calls "out-of bag", or OOB, data) by considering the tree developed with the bootstrap sample.
ii. Cumulate the OOB predictions. (On the average, every data point would be out-of-bag around 36% of the times, so cumulate these predictions.) Calculate the error rate, and call it the OOB estimate of error rate

2. K-Means Algorithm

In K-Means choose k cluster centers in the input space. Mark each training point as "captured" by the cluster to which it is closest. Move each cluster center to the mean of the points it captured. Repeat until convergence[18].

The k-Means clustering algorithm picks up the number of k centres randomly assigning the data points {x^p} to k subsets. It then uses a simple re-estimation procedure to end up with a partition of the data points into k disjoint sub-sets or clusters S_j containing N_j data points that minimizes the sum squared clustering function[18].

$$J=\sum_{J=1}^{k} \sum_{p€sj} ||x^p-\mu j||^2$$

Where, μj is the mean/centroid of the data points in set S_j given by

$$\mu j=1/Nj \sum_{p€s} x^p$$

It does that by iteratively finding the nearest mean μ_j to each data point x^P reassigning the data points to the associated clusters S_j, and then recomputing the cluster means μ_j.

The clustering process terminates when no more data point switch from one cluster to another. Multiple runs can be carried out to find the local minimum with lowest *J*.

3. RBF Algorithm

The Radial basis function network is an artificial neural network that uses radial basis function as activation functions[6]. Radial basis function network have many uses including system control classification.In the following we will assume that the choice of the radial basis function e(z) has already been made[13]. In order to have already been made in order to find the minimum of the cost function a learning algorithms must accomplish the following steps:

i. Select a search space (i.e. a subset of the parameter space);
ii. Select a starting point in the space (initialization);
iii. Search for the minimum(refining).

An RBFN is completely specified by choosing the following parameters:

i. The number n of radial basis functions;
ii. The centres ci and the distances k : ki, i.e. the matrixes Qi (i=1…n);
iii. The weights wi.

The number n of radial functions is a critical choice and depending on the approach can be made prior or determined incrementally. In fact, both the dimensions of the parameter space and consequently, the size of the family of approximations depend on the value of n.

4. Naive Bayes Algorithm

While looking for a way to classify short texts into several categories a simple but probably efficient method seems to be "Naive Bayes". An advantage of naive bayes is that it only requires a small number of training data to estimate the parameters necessary for classification[5]. This classifier is based on the Bayes rule of conditional probability. It makes use of the data, and analyses them individually as they are independent of all the attributes contained. This section introduces some of the basic facts about learning process:

A. Data Set

Total 400 instances of the dataset is used for the training to prediction algorithms, out of which 250 has label chronic kidney disease (CKD) and 150 has label non chronic kidney disease (NCKD). The clinical data of 400 records considered for analysis has been taken from UCI Machine Learning Repository. It has 25 attributes, 11 numeric and 14 nominal.

The below are the steps involved in this algorithm

Step 1: Scan the dataset

Step 2: Calculate the probability

Step 3: Apply the formulae

$$P=(n\_c + mp)/(n+m)$$

Where:

- n = the number of training examples v = vj
- n_c = number of examples for which v = vj and a = ai
- p = a prior estimate for P.
- m = the equivalent sample size

Step 4: Multiply the probabilities by p.

Step 5: Compare the values and classify the attribute values to one of the predefined set of class.

The above steps describes the working of the naïve bayes used to predict chronic kidney disease.

Although it's relatively simple idea, Navie Bayes can often outperform other more sophisticated algorithms and is extremely useful in common applications like spam detection and document classification.

III. RESULTS AND ANALYSIS

The experimental comparison of Naive Bayes and RF are done based on the performance vectors. It is statistical performance evaluation of classification tasks and contains list of performance criteria values.

Kappa statistic measures interrater eliability .Interrater reliability or precision happens when your data raters give the same score to the same data item.

The Kappa statistic differ from 0 to 1,where

- 0=agreement equivalent to chance.
- 0.1-0.20=slight agreement.
- 0.21-0.40=fair agreement.
- 0.41-0.60=moderate agreement.
- 0.61-0.80=substantial agreement.
- 0.81-0.99=near perfect agreement.
- 1=perfect agreement.

A. Performance Analysis (Naive Bayes vs RF)

Performance Vector:
Accuracy: 100.00%
Classification_error: 0.00%
Kappa: 1.000
Confusion Matrix:
Weighted_mean_recall: 100.00%, weights: 1, 1
Spearman_rho: 1.000
Kendall_tau: 1.000
Absolute_error: 0.000 +/- 0.000
Relative_error: 0.00% +/- 0.00%
Relative_error_lenient: 0.00% +/- 0.00%
Relative_error_strict: 0.00% +/- 0.00%
Normalized_absolute_error: 0.000
Root_mean_squared_error: 0.000 +/- 0.000
Root_relative_squared_error: 0.000
Squared_error: 0.000 +/- 0.000

Figure 1. Performance Vector for Naive Bayes

Figure 1. shows performance vector containing list of performance criteria values. Accuracy refers to number of correct predictions or how precise the dataset is being classified. Kappa takes into account the correct predictions occurring by chance. It gives a quantitative measure of the magnitude of agreement between observers. It lies in the range -1 to 1, where 1 is perfect agreement, 0 is chance agreement, and negative values indicate agreement less than chance i.e
disagreement between observers. The accuracy of Naive Bayes obtained is 100% and kappa value is 1 which indicates perfect agreement.

Performance Vector:
Accuracy: 87.3%
Kappa: 0.746
Spearman_rho: 0.542
Kendall_tau: 0.542
Absolute_error: 0.246 +/- 0.388
Relative_error: 24.63% +/- 38.75%
Relative_error_lenient: 24.63% +/- 38.75%
Relative_error_strict: 646.39% +/- 1,799.03%
Normalized_absolute_error: 0.493
Root_mean_squared_error: 0.459 +/- 0.000
Root_relative_squared_error: 0.918
Squared_error: 0.211 +/- 0.363

Figure 2. Performance Vector for RF

Figure 2 shows performance of RF with accuracy obtained as 87.3% and kappa value as 0.746 showing substantial agreement range.

IV.CONCLUSION

The experimental results of our proposed method have demonstrated that Naive Bayes has produced superior prediction performance in terms of classification accuracy for our considered dataset. As enhancement to the work done, further analyses can be carried to predict the current state of CKD using algorithms such as C4.5.

V. REFERENCES

1. Vijayarani, S., & Dhayanand, S., 2015. Data mining classification algorithms for kidney disease prediction. International Journal on Cybernetics and Informatics (IJCI).
2. Kumar, K., & Abhishek, B., 2012. Artificial neural networks for diagnosis of kidney stones disease.
3. Abhishek, G. S. M. T., & Gupta, D., 2012. Proposing Efficient Neural Network Training Model for Kidney Stone Diagnosis. International Journal of Computer Science and Information Technologies, 3(3), pp. 3900-3904.
4. Durairaj M, Ranjani V, Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.
5. A. Sai Sabitha, Abhay Bansal, Khushboo Chandel, VeenitaKunwar, Chronic Kidney Disease Analysis Using DataMining Classification Techniques, 6th InternationalConference on Cloud System and Big Data Engineering,2016
6. David S. K., Saeb A. T., Al Rubeaan K., Comparative Analysis of Data Mining Tools and Classification Techniques using WEKA in Medical Bioinformatics, Computer Engineering and Intelligent Systems, 4(13):28-38,2013.
7. Durairaj M, Ranjani V, Data mining applications in healthcare sector a study. Int. J. Sci. Technol. Res. IJSTR, 2(10), 2013.
8. Yasodha P, Kannan M, Analysis of Population of Diabetic Patient Database in WEKA Tool, International Journal of Science and Engineering Research, 2 (5), May 2011.
9. Duraiaj M, Ranjani V, Data Mining applications in healthcare sector a study. Int . J. Sci. Technol. Res. IJSTR, 2(10), 2013.
10. Hall M, Reutemann P, WEKA Knowledge Flow Tutorial for version 3-5-8, July 2008.
11. Scuse D, Reutemann P, WEKA Experimenter Tutorial for version 3-5-5, January 2007.
12. Shomona Gracia Jacob, R.Geetha Ramani, Discovery of Knowledge Patterns in Clinical

Data through Data Mining Algorithms: Multiclass Categorization of Breast Tissue Data, International Journal of Computer Applications (0975– 8887)Volume 32– No.7, October 2011.
13. Mihaila C, Ananiadou S, Recognising Discourse Causality Triggers in the Biomedical Domain, Journal of Bioinformatics and Computational Biology, 11(06), October 2013.
14. Thitiprayoonwongse D., Suriyaphol P., Soonthornphisaj N., Data mining of dengue infection using decision tree, Entropy, 2: 2, 2012.
15. Alam M., Shakil K.A., Cloud Database Management System Architecture, UACEE International Journal of Computer Science and its Applications, 3(1):27-31, 2013.
16. Alam M., Shakil K.A., A decision matrix and monitoring based framework for infrastructure performance enhancement in a cloud based environment, International Conference on Recent Trends in Communication and Computer Networks, Elsevier, pp. 174-180, November 2013.
17. Alam M., Shakil K.A., An NBDMMM Algorithm Based Framework for Allocation of Resources in Cloud, arXiv preprint arXiv: 1412.8028, 2014.
18. Shakil K.A. and Alam M., Data Management in Cloud Based Environment using k-Median Clustering Technique, IJCA Proceedings on 4th International IT Summit Confluence 2013 - The Next Generation Information Technology Summit Confluence 2013, pp. 8-13, January 2014.
19. Alam M., Shakil K.A., and Sethi S. ,Analysis and Clustering of Workload in Google Cluster Trace based on Resource Usage, arXiv preprint arXiv: arXiv: 1501.01426, 2014.
20. Shakil K.A., Sethi S., Alam M., An Effective Framework for Managing University Data using a Cloud based Environment, arXiv preprint arXiv:1501.07056, 2015
21. Zareen F.J. and Jabin S., A Comparative Study of recent trends in biometric signature verification,IC3, IEE, 354-358, 2013.
22. Witten H, Ian H. 2011. Data mining: practical machine learning tools and techniques. Morgan Kaufmann Series in Data Management Systems.
23. Tom Fawcett, (2003). ROC graphs: Notes and practical considerations for data mining researchers. Technical report, HP Laboratories.