Indian Journal of Science and Technology, Vol 9(31), DOI: 10.17485/ijst/2016/v9i31/95634, August 2016
ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

An Improved Prediction of Kidney Disease using SMOTE

Sai Prasad Potharaju[1*] and M. Sreedevi[2]

[1]K L University, Guntur - 522502, Andhra Pradesh, India; psaiprasadcse@gmail.com
[2]Department of Computer Science and Engineering, K L University, Guntur - 522502, Andhra Pradesh, India; msreedevi_27@kluniversity.in

Abstract

Objectives: This article presents a framework to improve the accuracy of rule induction and decision tree models. **Analysis:** In this paper, we used a rebalancing algorithm called SMOTE to enhance the accuracy of different induction and decision tree models in order to predict kidney disease of patients. For this prediction, data collected from Apollo Hospitals, Tamil Nadu, India has been analysed. **Findings:** In this research, initial dataset is not balanced i.e. most of the instances belong to the same class. If dataset is imbalanced, the traditional models can't produce accurate results. Thus the proposed framework improves the accuracy of models by balancing the imbalanced dataset. For this, a technique for sampling the minority class called SMOTE is applied on existing dataset and percentage of variation between classes is minimized. The examined findings with various classifiers algorithms and with the use of over sampling algorithm, the produced findings proves an increasing accuracy and also those results are compared with balanced and imbalanced dataset. In particular, this method can attain the average accuracy of 98.73%. **Applications:** This method can be applied in other areas to improve the accuracy in case of imbalanced dataset. In case of Big Data also SMOTE can be applied using Hadoop framework and Mapreduce programming model with new algorithmic approach.

Keywords: Classification, Data Mining, Health Informatics, Kidney Failure, SMOTE

1. Introduction

Data mining is a process of hidden knowledge discovery technique to analyse data and reduce it into utilizable information from large dataset[1]. Data mining techniques such as association rule, clustering, classification regression plays an important functions in getting hidden knowledge through the huge databases. The present research destine to predict the chances of occurring kidney disease from the collected patient dataset. The primary objective of data mining is predictions and descriptions of data, in practice[2]. Prediction in data mining includes attributes or variables in the dataset to know the hidden or future state of data.

The main aim of anticipation in data mining of health care is to know the hidden patterns in patient data in order to make better their health[3]. Because of the transformations in regular life style of citizens, Chronic Kidney Disease (CKD) has become a leading cause of deaths. Large scale available data about patient healthcare provides a goldmine that can be used to understand the functioning of different parts of body. With the high availability of healthcare data it has been become a rapid research[4] area in healthcare informatics with the advantages of data minng. To examine or to know, how patient's body is reacting with different medical test reports, unveiled information is useful.

Health Care Informatics (HCI) is a collection of computer science and information science within the domain of healthcare. There are number of research areas presented within the domain of HCI, which includes Translational Bio Informatics (TBI), Bioinformatics, Image Informatics, Clinical Informatics and also Public Health Care Informatics (PHCI). Research in HCI can range from data gathering, accessing, storage and analytics by applying data mining strategies, and so on. Recent

*Author for correspondence

healthcare research work has focused on predicting many diseases. Example of these include, prediction of breast cancer[5], prediction of Dermatological disease, prediction of diabetes[6], prediction of heart disease[7], *prediction* of *hepatitis* C virus (*HCV*), prediction of liver cancer. However, the scope of current study targets to examine the use of various techniques of classification over the kidney disease data particular to Apollo Hospitals, Tamil Nadu, India.

The dataset gathered from Apollo Hospitals, Tamil Nadu, India is strongly equal to data available at other hospitals in India. Hence, the produced findings by this framework is generalized, and same can be strongly applied for such kind of problem, with the study of techniques of classification like rule-based induction & Decision tree. The obtained hidden knowledge not limited to predict the chronic kidney disease, but also used to discuss the rationale of this conclusion[8]. In addition, the current study of this article gives useful set of observations, from which high sophisticated and accurate classification modeling techniques can be built. Prediction from imbalanced datasets i.e. huge variance in instances of different classes may not give accurate results.

In order to get accuracy in results, imbalanced dataset should be balanced by applying different techniques. In this research, rebalance algorithm i.e. SMOTE is applied on imbalanced dataset to make it balanced dataset.

The remaining part of this article is structured in different sections. Related work is introduced in Section 2 which will provide the road map for the remaining stages. Section 3 includes the data mining stages to predict kidney disease, it includes specification of dataset, dataset extraction, cleaning of dataset and transformation of dataset. To handle the problem of imbalanced data i.e. huge difference in the size of class, the suggested analytical framework has included artificial datasets into the process of learning. As a result, a traditional classifier can be employed to generate the needed predictive model. Section 4 gives an examination setup and the respective findings with different classification techniques like Decision tree & rule based induction. Conclusion and chances of future work is given in Section 5.

2. Related Work

With the observations of[9], the authors compared efficiency and accuracy using BPA-Back Propagation Algorithm, RBF-Radial Basis Function, SVM-Support Vector Machine .The important aim of authors research work was to suggest the advisable tool for detecting stone in kidney, to minimize the examine time and to increase the efficiency and the performance accuracy.

In research article[10], Performance Analysis of different Data Mining strategies in the detection of Heart Disease was introduced. The authors of[11] discussed about different clustering methods for Big Data Mining. In the research paper[12], the authors presented the framework to reduce the cost and exploit of choosing patients for clinical investigations. In this study, authors applied different data mining algorithms for getting hidden data in the form of decision rules. Patient can be selected on the basis of prediction output and also the process of discovering most weighted parameters is discussed in this paper. In the research article[13], authors proposed an ANFIS-Adaptive Neuro Fuzzy Inference System to predict renal failure time frame of CKD based on real clinical data. In the research article[14], authors compared SVM and Naive Bayes classifier algorithms for kidney disease prediction. With the examined findings it is concluded that the accuracy of the SVM is improved than the Naive Bayes (NB) classifier technique. But analysis of algorithm is carried out on imbalanced dataset. In addition, the author of[15] used decision tree algorithm called ID3 in order to predict the Information Technology students performance status. The obtained results intimates that those students in Computer Science have better performance than other students in different departments. Authors of research article[16] compared NBTree (NBT), C4.5, Bayesian Network(BN) and Decision Forest classification models for the purpose of planning the registration of course. That study suggests to teach the probability of course dropout. This study is examined using the popular tool called WEKA and it is observed that NBTree(NBT) generated the good level of accuracy. In research paper[17], the authors verified the findings of C4.5, CART and ID3 algorithms for detecting the progress report of 1st year candidates. In the same study, it has been concluded that ID3 is performed better than the other models tested. The techniques applied in[15-17] are considered in this research to predict kidney disease according to dataset availability. The research paper[18] introduced sampling technique for minority class called SMOTE, which is able to resolve the problem of imbalanced data. It is also suggested to read other learning techniques for better understanding of this study.

3. Methodology

In this section, designed and implemented data mining framework for current study is explained. **Figure. 1** shows various data mining stages including data gathering, data preparation, generating model and interpreting the data. Blood Pressure, Specific Gravity, Sugar, Bacteria, Blood Glucose Random, Blood Urea, Sodium, Potassium, Haemoglobin, Serum Creatinine, Diabetes Mellitus, etc are the various factors which leads to improper functioning of kidney.

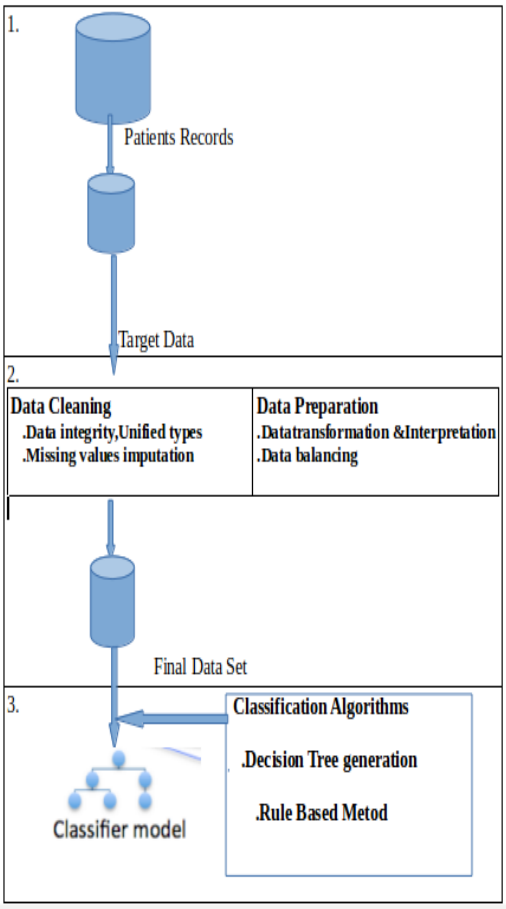

Figure 1. The data mining approach to prediction of improper functioning of Kidney.

3.1 Data Gathering and Preparation of Dataset

To examine and analyse the proposed approach of data mining technique of this current problem, particularly the patient's data of Apollo Hospitals as a case is collected. This collected data is over the period of two months in the year 2015. **Table 1** describes the list of 25 features involved in this present work.

Table 1. Description of investigated attributes, where class attribute values are used as the target classes

Attribute	Type
Age	numerical
Blood Pressure	numerical
Albumin	nominal
Sugar	nominal
Red Blood Cells	nominal
Pus Cell	nominal
Pus Cell clumps	nominal
Bacteria	nominal
Blood Glucose	numerical
Blood Ure	numerical
Serum	numerical
Sodium	numerical
Potassium	numerical
Hemoglobin	numerical
Packed Cell Volume	numerical
White Blood Cell Count	numerical
Red Blood Cell Count	numerical
Hypertension	nominal
Diabetes Mellitus	nominal
Coronary Artery Disease	nominal
Appetite	nominal
Pedal Edema	nominal
Anemia	nominal
Class (cdk or not cdk)	nominal

Before integrating the gathered instances with any classification technique, it is essential to prepare the complete set of data, with unique representation.

At first, aggregated data may contains noise problem, i.e. unidentified or wrong representation of attribute values. In particular to this work, any instance with such noise problem is discarded from the final data records. After solving the noise problem, the final dataset having total 400 instances. For attributes with numerical values, mean of the attributes value is considered to minimize and remove the bias in the process of learning.

3.2 Model Generation

Various Decision tree & rule-based induction techniques are examined in order to generate a classifier model and same can be analysed later. However, the objective of

both these models are same i.e. classification of data set, but the working of these models are different. The algorithms listed in **Table.2** are analysed in this work.

Table 2. Algorithms analyzed

Category	Algorithms
Rule induction	Jrip, OneR, Ridor
Decision tree algorithms	J48, SimpleCart, ADTree, RandomTree and REPTree

The evaluation technique of K-fold cross validation is applied to predict the performance and details of the resulting knowledge. For the prediction of this, dataset which consists of 400 instances are divided randomly into training datasets & test datasets. The partitioned datasets included 66% of training datasets and 34% of test datasets. This is iterated for 10 (ten) times i.e. K = 10(K-fold cross validation). For each of these ten iterations, the training dataset is given to the process of learning with a specific technique of classification. After completion of training phase, the respective test dataset is applied to examine the performance of classifier. The average of these ten iterated classification techniques result is considered as final accuracy.

But, the issue of imbalanced data that exists in the final dataset has not been considered. Specifically, the number of CDK instances (Improper Functioning of Kidney) and that of not cdk (Proper Functioning of Kidney) are not equal. In the final dataset 150 number of instances are belongs to class of cdk category and 250 are not cdk. Because of imbalanced data, the algorithm tends to bias class of majority category So, the result obtained may not be accurate.

Such type of imbalanced dataset difficulties are solved by applying a technique of over sampling called SMOTE[18]. In this particular study, the same technique is applied to balance the dataset. As a result of this technique, it increases the instances of the class of minority category, with artificial records generated using the principle of k-nearest neighbours. As for the findings given in the later section, the value of k is set to 5. After sampling the imbalanced dataset total 1170 instances were generated, in which 600 instances are belongs to not cdk (Proper Functioning of Kidney) class and 570 instances belong to cdk (Improper Functioning of Kidney) class after 3 samplings.

4. Results

In the initial examination, all specified techniques listed in Table 2 are applied to produce classifier models over the original training sets. The corresponding observations are listed in **Table 3**, where Decision tree techniques like ADTree and J48 techniques found to be better accurate than rule-based induction models. The least accuracy has been found with OneR model.

Table 3. Results with original data set, where the presented accuracy is the average across10-fold cross validation

Algorithm	Accuracy
OneR	92
JRIP	96.5
Ridor	97
ADTree	99
J48	98
RandomTree	97.5
REPTree	98
SimpleCart	97.75

In the second experiment, over sampling technique is applied on each training dataset for balancing the dataset. As a result of this, training dataset contained 1170 instances after rebalancing (570 CDK and 600 not cdk). It need not to be fully balanced, 5% of variation may be allowed for better accuracy. In this study, variation is 1.05%. **Table.4** shows the findings of this examination, in which all classification algorithms listed in Table 2 produced good results in comparison with imbalanced data. **Figure. 2** supports this conclusion. These observations suggests that the imbalanced problem has an improper impression on decision tree category than the others, analyse **Figure 3.**

Table 4. Results with balanced data set (applying SMOTE)

Algorithm	Accuracy
OneR	94.6
JRIP	99.65
Ridor	98.8
ADTree	99.65
J48	99.57

RandomTree	99.05
REPTree	99.23
SimpleCart	99.31

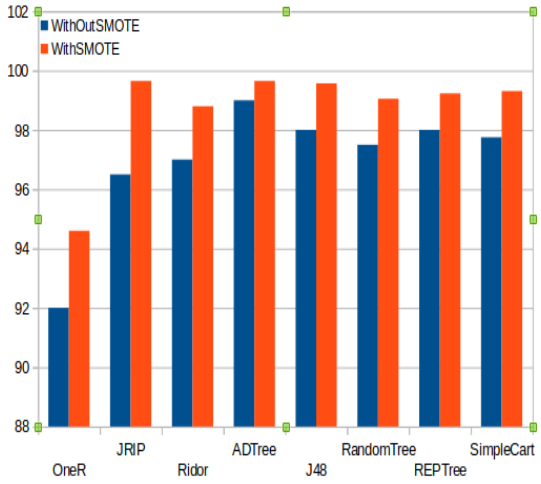

Figure 2. The algorithm-specific comparison of the results with two different data settings: original data and balanced data using SMOTE.

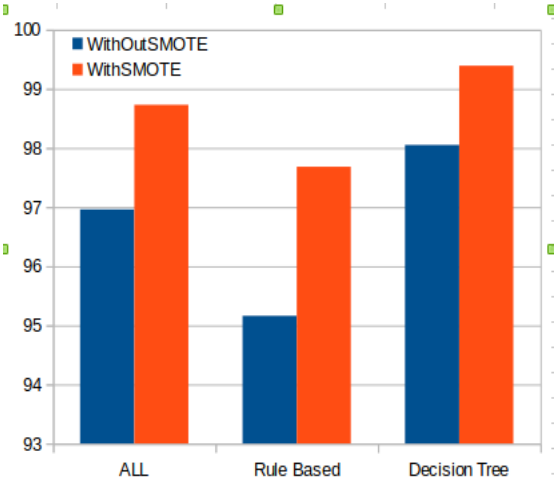

Figure 3. The improvement over original data using SMOTE algorithm, as the averages acrossall, and two categories of classification models.

The same may be improved by applying other strategy called as SVM . However, the producing model might not be clear to the common user. The applied rule-based induction methods has diverted to be the benefit of this present work. The rules obtained from ADTree technique and J48 technique are given in **Figure 4 and Figure 5** respectively.

```
| (1)sc < 1.213: -1.045
| | (8)bgr < 138.768: -0.635
| | (8)bgr >= 138.768: 0.815
| (1)sc >= 1.213: 3.063
| (2)sg < 1.02: 2.59
| (2)sg >= 1.02: -0.838
| | (4)al < 0.124: -0.956
| | | (6)hemo < 12.85: 0.959
| | | (6)hemo >= 12.85: -1.249
| | (4)al >= 0.124: 1.855
| (3)hemo < 12.983: 2.167
| (3)hemo >= 12.983: -0.893
| (5)dm = no: -0.13
| (5)dm = yes: 1.703
| (7)sg < 1.02: 0.931
| (7)sg >= 1.02: -0.32
| (9)al < 0.008: -0.093
| (9)al >= 0.008: 0.828
| (10)pcv < 42.012: 0.495
| (10)pcv >= 42.012: -0.581
Legend: -ve = notckd, +ve = ckd
Tree size (total number of nodes): 31
Leaves (number of predictor nodes): 21
```

Figure 4. The set of rules obtained by ADTree algorithm.

```
htn = no
|  sg <= 1.019125: ckd (103.12/2.84)
|  sg > 1.019125
|  |  appet = good
|  |  |  dm = no
|  |  |  |  pe = no
|  |  |  |  |  pcv <= 40.992378
|  |  |  |  |  |  pcv <= 39: ckd (7.94/0.21)
|  |  |  |  |  |  pcv > 39: notckd (20.56/0.03)
|  |  |  |  |  pcv > 40.992378: notckd (576.66/0.95)
|  |  |  |  pe = yes: ckd (3.0/0.0)
|  |  |  dm = yes: ckd (6.16/0.01)
|  |  appet = poor: ckd (7.87/0.01)
htn = yes: ckd (374.68/0.68)
```

Figure 5. The set of rules obtained by J48 algorithm.

Precision Without Smote	**Precision With Smote**	Recall Without Smote	**Recall With Smote**	Fmeasuer Without Smote	**Fmeasuer With Smote**	ROC Area	**ROC Area With Smote**	Class	Algorithm
0.954	**0.926**	0.916	**0.973**	0.935	**0.949**	0.921	**0.945**	ckd	
0.869	**0.926**	0.927	**0.973**	0.897	**0.949**	0.921	**0.945**	notckd	OneR
0.972	**1**	0.972	**0.993**	0.972	**0.996**	0.961	**0.996**	ckd	
0.953	**0.993**	0.953	**1**	0.953	**0.997**	0.961	**0.996**	notckd	Jrip
0.988	**0.998**	0.964	**0.977**	0.976	**0.988**	0.972	**0.988**	ckd	
0.942	**0.979**	0.98	**0.998**	0.961	**0.988**	0.972	**0.988**	notckd	Ridor
0.992	**1**	0.976	**0.991**	0.984	**0.996**	0.995	**0.998**	ckd	
0.961	**0.992**	0.987	**1**	0.974	**0.996**	0.995	**0.998**	notckd	J48
0.992	**0.998**	0.968	**0.982**	0.98	**0.99**	0.999	**0.999**	ckd	
0.949	**0.984**	0.987	**0.998**	0.967	**0.991**	0.999	**0.999**	notckd	RandomTr
0.976	**1**	0.992	**0.984**	0.984	**0.992**	0.997	**0.997**	ckd	
0.986	**0.985**	0.96	**1**	0.973	**0.993**	0.997	**0.997**	notckd	RePTree
0.969	**1**	0.996	**0.986**	0.982	**0.993**	0.996	**0.998**	ckd	
0.993	**0.987**	0.947	**1**	0.969	**0.993**	0.996	**0.998**	notckd	SimpleCart

Figure 6. Comparision of different Parameters with and without SMOTE.

After completion of analysis using different algorithms, various parameters like Precision,Recall,F-Measure,ROC Area also compared as shown in **Figure 6.**

5. Conclusion and Future Work

This research article presented on the improvement of analytical framework for the prediction of kidney functioning. The approach used in this work follows the basic stages in data mining as shown in Figure 1. Initially, dataset has been gathered and integrated to form the target dataset. In the data preparation phase, traditional techniques for removing missing values are employed and normalization technique is applied to remove bias and bugs identified in the data. At the end, various rule-based induction models and Decision tree models are applied for the classification, such that the generating hidden knowledge is interpretable. The accuracy obtained by Decision tree models are performed better than rule based models. The accuracy recorded is better with the application of SMOTE in case of imbalanced data. Further, SMOTE can be applied for Big Data analysis using Hadoop framework with the help of mapreduce programming model with new algorithmic approach, which is our future work.

6. References

1. Mircea GH, Robert J. Howlett, Jain LG. Knowledge-based Intelligent Information and Engineering Systems. 8th International Conference, KES 2004, Wellington, New Zealand, Sep 20-25. Proceedings, e-book Springer publication. 2004.
2. Rao R. Survey on prediction of heart morbidity using datamining techniques. International Journal of Data Mining and Knowledge Management Process (IJDKP). 1(3):14–34.
3. Sen AK, Patel SB, Shukla DP. A Data Mining Technique for Prediction of Coronary Heart Disease Using Neuro-Fuzzy Integrated Approach Two Level. International Journal of Engineering and Computer Science. 2(9):1663–71.
4. Herland M, Khoshgoftaar TM, Wald R. A review of data mining using big data in health informatics. Journal of Big Data. 2014; 1:2**.** Doi: 10.1186/2196-1115-1-2.
5. Safavi AA, Parandesh NM, Salehi M. Predicting breast cancer survivability using data mining techniques, School of Electrical and Computer Engineering, Shiraz University, Iran. Doi: 10.1109/ICSTE.2010.5608818.
6. Meng X-H, Huang Y-X, Rao D-P, Zhang Q. Comparison of three data mining models for predicting diabetes or prediabetes by risk factors. Elsvier.
7. Doi: http://dx.doi.org/10.1016/j.kjms.2012.08.016,
8. Masethe HD, Masethe MA. Prediction of Heart Disease using Classification Algorithms. Proceedings of the World Congress on Engineering and Computer Sciences, WCECS, San Francisco, USA. 2014; 2014 October 22-24; 2.
9. Pandey M, Sharma VK. A decision tree algorithm pertaining to the student performance analysis and prediction. International Journal of Computer Applications. 61(13).
10. Bharadwaj A, Thakur M, Dolly Gupta. International Journal of Computer Science and Information Technologies, 3(3):3900–4.

11. Lohita K, Sree AA, Poojitha D, Renuga Devi T. A. Umamakeswari Performance Analysis of Various Data Mining Techniques in the Prediction of Heart Disease. Indian Journal of Science and Technology. 2015 Dec; 8(35), Doi: 10.17485/ijst/2015/v8i35/87458.
12. Sajana T, Sheela Rani CM, Narayana KV. A Survey on Clustering Techniques for Big Data Mining. Indian Journal of Science and Technology. 2016 Jan; 9(3). Doi: 10.17485/ijst/2016/v9i3/75971.
13. Kusiak A, Dixonb B, Shaha S. Predicting survival time for kidney dialysis patients: a data mining approach, Elsevier Publication. Computers in Biology and Medicine 35:311–27.
14. Norouzi A, Yadollahpour A, Mirbagheri SA, Mazdeh MM, Hosseini SA. Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System. Hindawi Publishing Corporation, Computational and Mathematical Methods in Medicine. Article ID6080814.
15. Vijayarani S, Dhayanand S. Data mining classification algorithms for kidney disease prediction. International Journal on Cybernetics and Informatics (IJCI). 2015 Aug; 4(4).
16. Kumar V, Singh S. Classification of students data using data mining techniques for training and placement department in technical education. Proceedings of International Conference on Computer Science and Networks. 121–6.
17. Pumpuang P, Srivihok A, Polgrang P. Comparisons of classifier algorithms: bayesian network, C4.5, decision forest and NBTree for course registration planning model of undergraduate students. Int Proceedings of IEEE International Conference on Systems, Man and Cybernetics. 3647–365.
18. Bunkar K, Singh UK, Pandya B, Bunkar R. Data mining: prediction for performance improvement of graduate students using classification. Int Proceedings of Ninth International Conference on Wireless and Optical Communications Networks. 1–5.
19. Chawla NV, Bowyer KW, Hall LO, Kegelmeyer WP. SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research. 16:321–57.