IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 10, October 2017

Prediction of Kidney Disease using Hybrid System based on Regression, Decision Tree and K-Near Neighbour method

Mansooreh Fateh[1], **Seyyed Javad Mirabedini**[2]

Department Computer Engineering, Islamic Azad University, Damavand Branch, Damavand, Iran[1]

Department Computer Engineering, Islamic Azad University, Central Tehran Branch, Tehran, Iran[2]

Abstract: In the world, many people are suffering from kidney failure, and a large percentage of people are not aware until the advanced stage. Therefore, early detection and diagnosis of disease can be an effective step in the treatment and survival of the patient. In this study, the proposed model is a hybrid system consisting of decision tree, linear regression and nearest neighbour. The first step is created based on the training dataset of the decision trees models, linear regression and nearest neighbour. In the second step, the related class of the available samples in the test dataset is predicted based on the obtained models. In the last step, final output is determined based on the voting strategy. The obtained accuracy of this method is 89.47% in the proposed method, which had the best performance in comparison with the decision tree, linear regression and nearest neighbour.

Keywords: Decision Tree, Regression, K-NN, Hybrid Methods, Kidney Disease.

I. INTRODUCTION

With the more increasing advance of data collection and storage, the production of a large amount of information in various sciences is provided [1]. Today, large databases such as Wikipedia, Facebook, and Google contain a huge amount of data which analyse data with a variety of goals and theses analysis can be effective to make good managerial decisions in various domains. Hence, data mining can be resulted in success of making decision based on knowledge. Today's data mining-based methods have been interred to extract knowledge in various fields such as commerce, industry, government, medicine, the web, social media, and bioethics. Data mining in the medical field has been applied to predict, diagnose the disease, and also to design medical professional systems based on biomedical data. Biomedical data includes laboratory data, images (MRI, Radiology, CT Scan,...) and physician reports. Artificial intelligence techniques have been applied to manage knowledge in the medicine field since the 1970s. That time, the MYCIN program was developed for doctor's advice and making decision. In MYCIN, the obtained knowledge by experts is presented under a set of if- then rules that these types of systems later became known as expert systems [3]. Recent researches indicate the application of data mining to diagnose and predict the disease in the health domain. For example, the use of artificial intelligence techniques in identifying thalassemia types [2-5], predicting diabetes in thalassemia patients [6-7], identifying benign and malignant breast cancer [8], diagnosis of diabetes in women [9-13]. This paper identifies kidney disease based on the Chronic Kidney Disease dataset using a complex system including data mining methods, such as decision tree, nearest neighbour and regression. A timely diagnosis of renal failure can lead to an increase the patient life and to reduce the risks of the diseases. Identification at an early stage may not be easy due to lack of accurate data and adequate information. So, the study has tried to offer a proposed method to increase the accuracy in the diagnosis of healthy individuals from patients with renal problem.

II. DATASET

In this paper, the kidney disease dataset is used which was collected in 2015, This collection includes 24 features, which include: age, bp, sg, al, sugar, rbc, pus cell, ba, bgr, bu, sc, sod, pot, hemo, pvc, wbcc, rbcc, htn, dm, cad, appet, pe, ane.

The samples in dataset are in two groups, patients with kidney failure and healthy people. Among the 400 available records in the dataset, there are 250 records, including class 1 (patients with kidney failure) and 150 records, including grade 0 (non-renal).

The sample selection strategy is random selection for the training and testing dataset. Here, 105 random samples are selected from each class for the training data set, and the rest of the data is used for the test set. So, the number of records in the training set is 210, and for the test set are 190.

DOI10.17148/IJARCCE.2017.61002

III. DECISION TREE

The decision tree is one of the most powerful tools to categorize and predict that have a tree structure similar to a flowchart. This classification method is a natural and visual approach. The classification is based on the sequence of raised questions. The next question is based on the current question value. This approach is useful for nonmetric data because the answer is "yes / no" or "true / false." The questions sequence forms a simple tree.
The first question makes the first node of the tree, which is called the root and then, the next questions form the inner nodes, until it reaches the leaf or end of the tree. The amount of the leaf determines the class of the pattern [14]. The two main issues in creating the decision tree are the selection of the most suitable attributes for each node and the condition of the algorithm end. The Information Gain, Gini index, Gain ratio, Like Hood Ratio and DKM criteria are several criteria for the attributes selection. In this study, the decision tree is modelled according to the accuracy criterion based on the training dataset. Table 1 shows the results of the decision tree.

TABLE1 RESULTS FROM ACCURACY CRITERION

Accuracy rate	Validation rate	Precision rate	Class Name
84.74%	100%	83.33%	**patient**
	35.56%	100%	**healthy**

As seen in Table 1, the resulted accuracy rate of the decision tree is 84.74 based on the precision evaluation criterion. But, the diagnosis rate of healthy individuals is 35.56%, which indicates poor performance of this model in classification.

IV.K-NEAR NEIGHBOUR

The nearest neighbour approach is one of the weak learner methods in the Rapid Miner. This method is based on computing. The K-NN operator is used to model the nearest neighbour in the Rapid Miner. In this paper, the number of neighbours is 3. Table 2 shows the results of the neighbourhood number.

TABLE 2 RESULTS FROM THE NEAREST NEIGHBOUR

Accuracy rate	Validation rate	Precision rate	Class Name
86.32%	89.66%	92.20%	**Patient**
	75.56%	69.39%	**Healthy**

As seen in Table 2, the accuracy rate of the nearest neighbour model is 86.32%. The performance of this model is better than the decision tree. Because 75.56% of healthy samples were precisely diagnosed.

V. REGRESSION

Regression is a powerful statistical technique that can be used at all stages of the data mining process. In linear regression, two different data types are modelled as a straight line. Regression line is a tool to predict the value of a variable based on its dependent variable. In fact, in order to model the values of two characteristic attributes, we find a line that is close to all the pairs of values of these two specific attributes. If the couple is not on a straight line, there is an error for each pair and regression line. This line will be selected in some way to minimize the error. Minimizing the sum of errors squares is the conventional method that is used in most cases. For the two variables X and Y, the following linear equation is considered [14].

$$Y=\alpha+\beta X \qquad (2\text{-}2)$$

The values of α and β are calculated based on the following formulas, which are called regression correlation coefficients.

$$\alpha=\bar{Y}-\beta\bar{X} \qquad (3\text{-}2)$$

β=(∑_(i=1)^n▒〖(x_i-(X))̅(y_i-Y̅)〗)/(∑_(i=1)^n▒〖(x_i-(X))̅〗^2)   (4-2)

Where X̅ and Y̅ are the ordered pairs average, respectively. And y_i and x_i are also sample values that their number reach n. Table 3 shows the results of linear regression in RapidMiner

TABLE 3 RESULT OF REGRESSION

Class name	Precision rate	Validation rate	Accuracy rate
Patient	88.61%	96.55%	87.89%
Healthy	84.38%	60%	

DOI10.17148/IJARCCE.2017.61002

As shown in Table 3, the accuracy rate of linear regression is 87.89%. In this method, the important issue is the classification precision of patient samples, 96.55%, but it has a low accuracy compared to healthy ones. The performance of this model is better than the decision tree, but is weaker than the nearest neighbour.

VI. PROPOSED METHOD

In this research, a hybrid classification system has been used to increase the accuracy of the healthy and patient samples diagnosis. This hybrid system consists of three approaches, the nearest neighbour, regression, and decision tree. Figure 1 describes how to model these three methods.

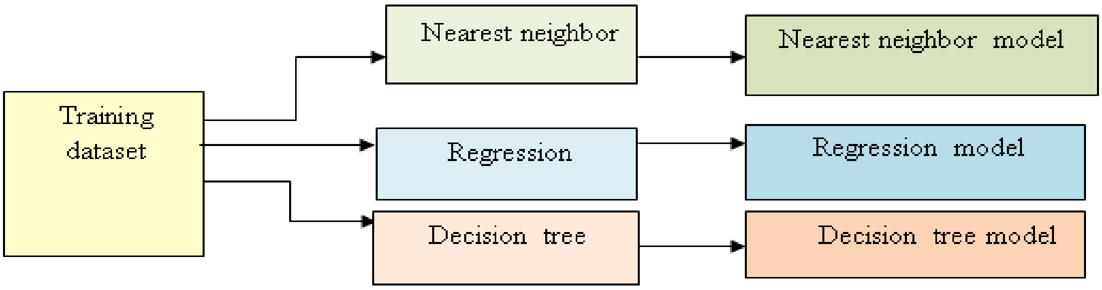

Figure 1 - The first step in the proposed method (creating regression models, decision trees, and closest neighbours)

In the first step, the three classification methods, regression, the nearest neighbour and the decision tree are modelled based on the training dataset. In the second step, the models obtained from each method classify the available samples in the test dataset. The final decision is based on the voting strategy according to the results from each model. For example, if the output of the two models is class 1 and the final output is class 1. In the last step, the final output is evaluated based on the confusion matrix. In Figure 2, testing stage of the models and the evaluation of the proposed.

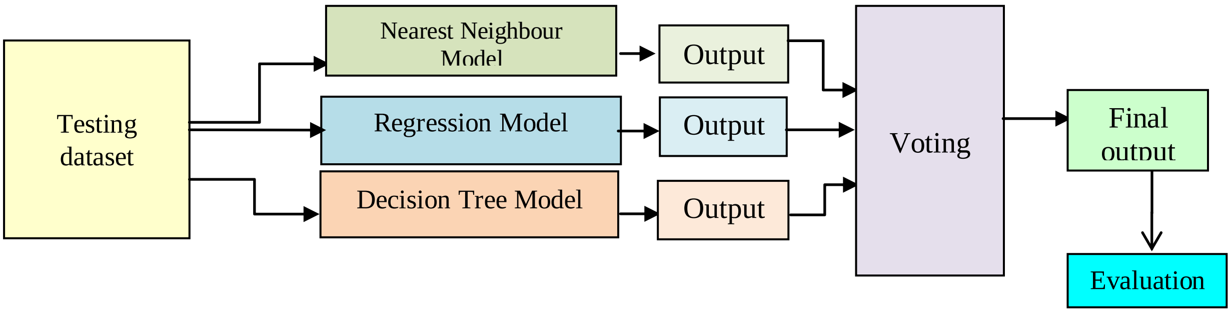

Figure 2. Testing stage in the proposed method (test of regression model, decision tree model and nearest neighbour model)

According to Fig. 2, this system uses the output of regression models, decision tree model, and nearest neighbour model based on the voting strategy. The class which has the most votes, is considered as output. Table 4 shows part of the voting. "Ckd" represents the patient people and "notckd" represents the healthy one.

TABLE 4 VOTING IN HYBRID SYSTEM

Number Row	Actual Class	K-NN	Regression	Decision Tree	Hybrid System
1	ckd	ckd	ckd	ckd	ckd
2	ckd	ckd	ckd	ckd	ckd
3	ckd	ckd	ckd	ckd	ckd
4	ckd	ckd	ckd	ckd	ckd
5	ckd	ckd	ckd	ckd	ckd
6	notckd	notckd	notckd	notckd	notckd
7	notckd	notckd	notckd	notckd	notckd
8	ckd	ckd	ckd	ckd	ckd
9	ckd	ckd	ckd	ckd	ckd
10	ckd	ckd	ckd	ckd	ckd
11	notckd	notckd	notckd	notckd	notckd
12	notckd	notckd	ckd	notckd	notckd

DOI10.17148/IJARCCE.2017.61002

Table 5 presents the results of a hybrid system consisting of decision tree, regression, and nearest neighbour. The system uses the outputs of the regression, the decision tree, and the nearest neighbour models based on the voting strategy, which were provided in the previous sections.

TABLE 5 RESULT FROM PROPOSED METHOD

Class name	Precision rate	Validation rate	Accuracy rate
Patient	87.87	100	89.47
Healthy	100	55.55	

Regarding the presented values in Table 5, accuracy rate of the proposed method is 89.47%, which has the highest accuracy rate compared to other methods.

VII. EXPERIMENTAL RESULTS

In this section, the comparison and evaluation are conducted based on the obtained results from the previous sections. Table 6 shows comparison among the decision tree, regression, nearest neighbour results and suggested method.

TABLE 6 COMPARISON OF OBTAINED RESULTS FROM THE DECISION TREE, REGRESSION, NEAREST NEIGHBOUR AND PROPOSED METHOD

Method	Precision Rate of Patient Class	Precision Rate of Healthy Class	Validation Rate of Patient Class	Validation Rate of Healthy Class	Accuracy Rate
Decision Tree	%83.32	%100	%100	%35.56	%84.74
Nearest Neighbour	%92.2	%69.39	%89.66	%75.56	%86.32
Regression	%88.61	%84.38	%96.55	%60	%87.89
Proposed Method	**%87.87**	**%100**	**%100**	**%55.55**	**%89.47**

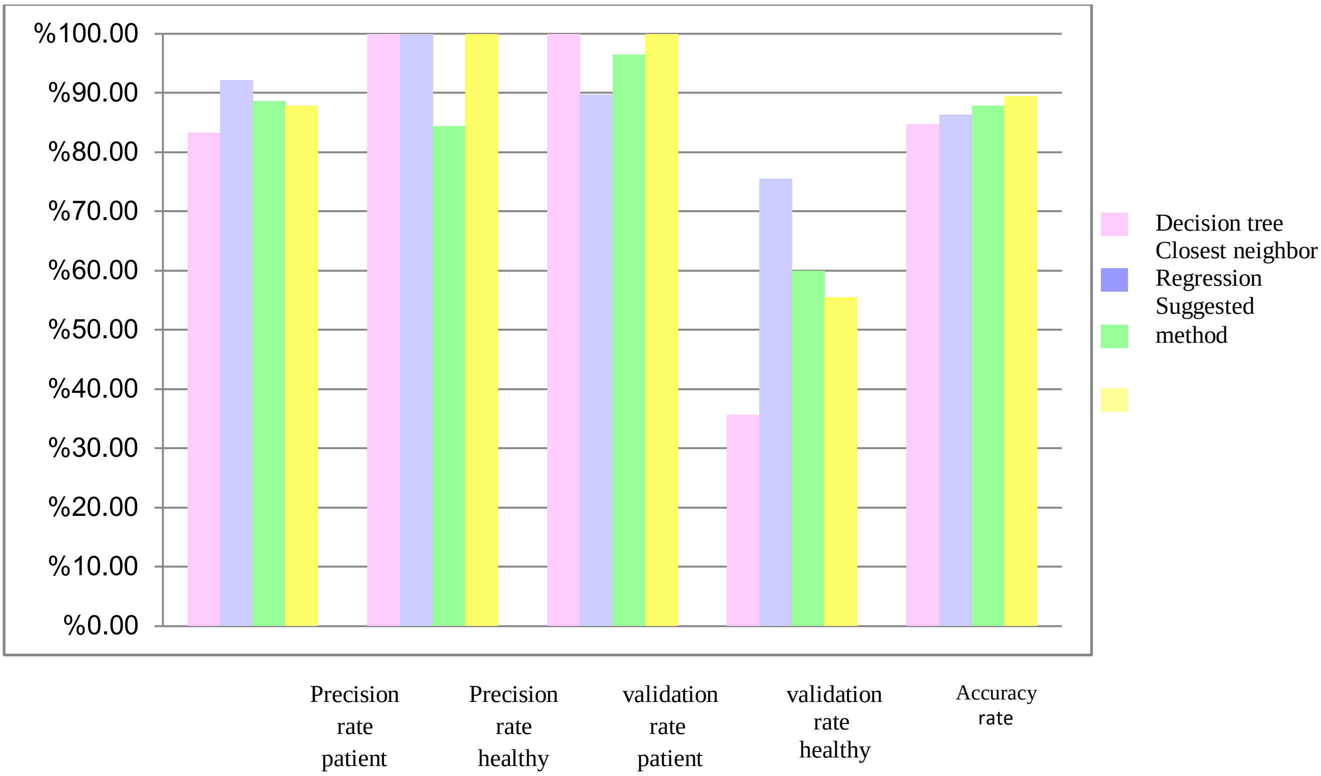

Figure 3. Chart of results comparison

DOI10.17148/IJARCCE.2017.61002

According to the presented results in Table 6, the accuracy rate of the proposed method has been improved, which indicates the good performance of this method in the classification of healthy and patient samples based on the dataset. Figure 3 shows the comparison chart of the results of decision tree, regression, closest neighbor, and suggested method.

Given to Figure 3, the proposed method has the best performance at the precision rate. And then, the regression with 87.79% has the highest the accuracy rate to diagnose healthy and patient samples. In the meantime, the nearest neighbour method has the highest rate of validation in the healthy people class and the accuracy in the patient class.

VIII. CONCLUSION

Kidney disease is a hidden disease in the medical field, which the patient may aware of kidney failure after a few years. Therefore, timely and early diagnosis of the disease can play an important role in the health of individuals. In this study, the Chronic Kidney Disease Dataset was used to diagnose kidney failure. For this purpose, a hybrid classification system consisting of decision tree, nearest neighbour and linear regression has been used. Among the available 400 samples in the dataset, 210 random samples were considered for training and creating decision tree, linear regression, and nearer neighbors models and 190 remaining samples were used to evaluate the hybrid system model. Based on the obtained results from the test samples, the accuracy of this method is 89.47%.

REFERENCES

[1] J. Han and M. Kamber, “Data Mining Concepts and Technique”, Murgan Kaufman, Elsevier, 2006.

[2] S. R. Amondelia, G. Cossu, M. L. Ganadu, B. Golosio, G. L. Masala and G.M. Mura, "A Comparative Study of K-Nearest Neighbour, Support Vector Machine and Multi-Layer Perceptron for Thalassemia Screening", Chemometrics and Intelligent Laboratory Systems, Vol. 69, 2003.

[3] M. Payandeh, M. Aeinfar, V. Aeinfar and M. Hayati, "A New Method for Diagnosis and Predicting Blood Disorder and Cancer Using Artificial Intelligence", IJHOSCR, Vol. 3, 2009.

[4] E. H. Elshami and A. M. Alhalees, "Automated Diagnosis of Thalassemia Based on Data Mining Classifiers", the International Conference on Informatics and Applications, the Society of Digital Information and Wireless Communication, 2012.

[5] F. Yousefian, T. Bairostam, A. Azarkeivan, "Prediction Thalassemia Based on Artificial Intelligence Techniques: A Survey", International Journal of Advanced Research in Computer and Communication Engineering, Vol. 6, No. 6, pp. 281-287, 2017.

[6] F. Yousefian, T. Bairostam, A. Azarkeivan, "Predicting the Risk of Diabetes in Iranian Patients with β-Thalassemia Major / Intermedia based on Artificial Neural Network MLP", IEEE Transaction on Knowledge and Data Engineering, submitted for publication.

[7] F. Yousefian , T. Bairostam, A. Azarkeivan , "Prediction of Mellitus Diabetes in Patients with Beta- thalassemia using Radial Basis Network, and k-Nearest Neighbor based on Zafar Thalassemia Datasets", 5th International Conference on Science, Engineering and Technology Innovation, 2017.

[8] Bashir S, Qamar U, Hassan Khan F, WebMAC: A Web based Clinical Expert System, InfSyst Front: Springer, pp. 1-17, 2016.

[9] Ganji MF, Abadeh MS, "A Fuzzy Classification System based on Ant Colony Optimization for Diabetes Disease Diagnosis," Expert Systems with Applications, 38(12): 14650–14659, 2011.

[10] A. Iyar, S. Jeyalatha and R. sumbaly , “Diagnosis of Diabetes Using Classification Mining Techniques,” International Journal of Data Mining & Knowledge Management Process(IJDKP), Vol. 5, No. 1, pp.1-14, 2015.

[11] AlJarullah, A. "Decision Tree Discovery for the Diagnosis of Type II Diabetes", International Conference on Innovations in Information Technology. 303-307, 2011.

[12] Pradhan M, Sahu RK (2011). Predict the Onset of Diabetes Disease using Artificial Neural Network (ANN). International Journal of Computer Science & Emerging Technologies, 2(2): 303-311.

[13] Gajendran G, Venugopal T, Mathematical Approach to Design a Batch Mode Multilayer Feed forward Neural Network and Its Model on Type 2 Diabetes. Global Journal of Pure and Applied Mathematics (GJPAM), 2016.

[14] Duda, R; Hart, P. E and Stork, D. G, Pattern Classification. John Wiley &Sons, 2012.

DOI10.17148/IJARCCE.2017.61002