COMPARATIVE STUDY OF CLASSIFICATION ALGORITHMS FOR DIABETES AND CHRONIC KIDNEY DISEASE DIAGNOSIS

Aung Nway Oo
Faculty of Computer Science
University of Information Technology
aungnwayoo78@gmail.com

Khin Thuzar Win
Department of Information Technology Supporting and Maintenance
University of Computer Studies Hinthada
khinthuzarwin87@gmail.com

Abstract: Now a day, data mining and machine learning methods are used to analyse the medical dataset. These techniques can reduce the number of tests to be taken by a patient, can save cost and can also save time for both, doctors and patients. Classification is a classic data mining technique based on machine learning. There are many classification algorithms that can be used for medical domain. In this paper, Naïve Bayes, Random Forest, KStar and PART classification algorithms are used to classify Diabetes dataset and Chronic Kidney Disease (CKD) dataset. The main objective of this paper is to compare the classification results of each classifier for Diabetes dataset and Chronic Kidney Disease (CKD) dataset.

Keywords: data mining, machine learning and classification,

1. INTRODUCTION

In recent decades, Diabetes and Chronic Kidney Disease (CKD) are challenges for health sector. Diabetes is due to either the pancreas not producing enough insulin, or the cells of the body not responding properly to the insulin produced [1]. There are three main types of diabetes mellitus:

1. Type 1 diabetes results from the pancreas's failure to produce enough insulin due to loss of beta cells.
2. Type 2 diabetes begins with insulin resistance, a condition in which cells fail to respond to insulin properly.
3. Gestational diabetes is the third main form, and occurs when pregnant women without a previous history of diabetes develop high blood sugar levels.

The main job of the kidneys is to remove waste from the blood and return the cleaned blood back to the body. Kidney failure means the kidneys are no longer able to remove waste and maintain the level of fluid and salts that the body needs. One cause of kidney failure is diabetes mellitus, a condition characterized by high blood glucose (sugar) levels. Over time, the high levels of sugar in the blood damage the millions of tiny filtering units within each kidney. This eventually leads to kidney failure. A person with diabetes is susceptible to nephropathy whether they use insulin or not. The risk is related to the length of time the person has diabetes [2].

In 2017, 425 million people had diabetes worldwide, up from an estimated 382 million people in 2013 and from 108 million in 1980. Chronic kidney disease affected 753 million people globally in 2016, including 417 million females and 336 million males. In 2015 it resulted in 1.2 million deaths, up from 409,000 in 1990. Therefore diagnosis of chronic kidney disease and diabetes is important.

In recent decade, the uses of data mining techniques in medical studies are growing gradually. Medical diagnosis is an important but complicated task that should be performed accurately and efficiently and its automation would be very useful.

Data mining is the process of discovering patterns in large data sets involving methods at the intersection of machine learning, statistics, and database systems. Data mining is an interdisciplinary subfield of computer science and statistics with an overall goal to extract information from a data set and transform the information into a comprehensible structure for further use. Data Mining has been used in a variety of applications such as marketing, customer relationship management, engineering, and medicine analysis, expert prediction, web mining, mobile computing and health care.

Medical data mining is the new area for exploring hidden data pattern from huge amount of data. Hospitals, clinics and medical analysis laboratories accumulate a large amount of patient data over the years. These data provide a basis for the analysis of risk factors for many diseases (various types of cancer, heart diseases, diabetes, chronic kidney disease, etc.).

Classification is one of the data mining methodologies used to predict and classify the predetermined data for the specific class. The various classification algorithms applied on different kinds of medical datasets for diagnosis.

The rest of the paper is organized as follows. Section 2 and section 3 provide introduction and the related work. General description of diabetes and chronic kidney disease datasets are presented in section 4. The experimental results are discussed in section 5. Finally, conclusion of this study was provided in section 6.

2. RELATED WORK

There are many research papers proposed for medical diagnosis. There are many research papers proposed for medical diagnosis.

In 2016, S.Yuvarani and R. Selvarani [3] proposed decision tree models for diabetes classification using LADtree, NBtree and a Genetic J48tree. AsmaB.M. Patil [4] performed different classification algorithms with varying accuracies and suggested improved prediction accuracy using weighted least squares SVM. Gaganjot Kaur predicted a modified J48 Classification Algorithm for the Prediction of Diabetes [5].

Koushik Chandra Howlader et al. [6] research to predict the severity of diabetes and find out significant features and gathered diabetes patients records from Noakhali Diabetes Association, Noakhali, Bangladesh. And then, preprocessed the raw dataset and used CDT, J48, NBTree and REPtree decision tree based classification techniques to analyze the dataset. In paper [7] proposed the rule based classification for diabetic patients using cascaded K-Means and Decision Tree C4.5. This paper presented the development of a hybrid model for classifying Pima Indian diabetic database (PIDD).

In the research work of [8], Comparative analysis of different classification algorithms have been done using various criteria's like accuracy, execution time and how much instances are correctly classified or not classified correctly for chronic kidney disease dataset.

In [9], KNN, SVM, RBF and Random subspace data mining methods were applied on the data set consisting of 400 samples and 24 attributes taken from UCI for classification of chronic kidney disease with particle swarm optimization (PSO) based feature selection method. M. Praveena and N. Bhavana [10] proposed prediction of chronic kidney disease by using C4.5 algorithm and predicted whether the person is normal or suffering from kidney problem. Paper [11] aimed at using 3 classifiers: multilayer perceptron, naive bayes and J48 decision tree in the prediction of chronic kidney disease dataset. The aim of this paper is to evaluate the performance of the classifiers used based accuracy, specificity, sensitivity, error rate and precision. Pallavi Sharma et al. [12] reviewed the different data mining techniques for chronic kidney disease prediction.

In this paper, comparative study of classification algorithms for diabetes and chronic kidney disease diagnosis was proposed.

3. CLASSIFICATION ALGORITHMS

Classification is a data mining function that assigns items in a collection to target categories or classes. The main goal of classification techniques is to predict accurately. The aim is to predict each class accurately for each and every class of data. This paper compared the classification results of following classification algorithms.

Naive Bayes: the Naive Bayes algorithm is based on the Bayesian theorem and operates on conditional probability. Despite its simplicity, it is a powerful algorithm for predictive modeling. Additionally, the Naive Bayes classifier works quite well concerning real-world situations. An example is spam filtering, which is a well-known problem for which the Naive Bayes classifier is suitable. As with the BayesNet algorithm, there should be no missing data in this algorithm and the variables must be discrete. Since there are no missing data in this dataset, the Naive Bayes algorithm can be applied after discretization of the continuous variables [13].

Random forest: Random forest is an ensemble machine learning algorithm that is used for classification and regression problems. It is a supervised machine learning algorithm which creates forest and makes it random. The forest is an ensemble of decision trees, mostly trained with training method [14]. In simple word, Random

forest algorithm generates multiple decision trees and merges them together to get the maximum accuracy and immutable prediction.

K-Star: K-star is a classifier based on instances, which is a test instance class. It has as foundation, the class of those training instances similar to it, as it is driven by some similarity function. It uses distance operation based on entropy; this makes it stand apart from distant learners based on instances [15].

PART: It is a separate-and-conquer rule learner proposed by Eibe and Witten [16]. The algorithm producing sets of rules called decision lists which are ordered set of rules. A new data is compared to each rule in the list in turn, and the item is assigned the category of the first matching rule (a default is applied if no rule successfully matches). PART builds a partial C4.5 decision tree in its each iteration and makes the best leaf into a rule. The algorithm is a combination of C4.5 and RIPPER rule learning.

4. DATASETS INFORMATION

The dataset used in this paper was obtained from UCI machine learning repository. The diabetes dataset contained 768 instances and 8 attributes. There are two types of class values tested positive (patient with diabetes) and tested negative (patient without diabetes). The attributes of diabetes dataset was described in following table.

Table 1. Attribute list of diabetes dataset from UCI Machine Learnig Repository [17]

Attribute Name	Attribute Description
Preg	Number of times pregnant
Plas	Plasma glucose concentration a 2 hours in an oral glucose tolerance test
Pres	Diastolic blood pressure (mm Hg)
Skin	Triceps skin fold thickness (mm)
Insu	2-Hour serum insulin (mu U/ml)
Mass	Body mass index (weight in kg/(height in m)^2)
Pedi	Diabetes pedigree function
Age	Age

The following table shows the description of chronic kidney disease dataset. There are 400 patient records and 24 attributes. The class values are ckd (patient who suffers chronic kidney disease) and not ckd (patient who does not suffer chronic kidney disease).

Table 2. Attribute list of chronic kidney disease dataset from UCI Machine Learning Repository [17]

Attribute Name	Attribute Description
Age	Age
Bp	Blood pressure
Sg	Specific gravity
Al	Albumin
Su	Sugar
Rbc	Red blood cells
Pc	Pus cell
Pcc	Pus cell clumps
Ba	Bacteria
Bgr	Blood glucose random
Bu	Blood urea
Sc	Serum creatinin
Sod	Sodium
Pot	Potassium
Hemo	Haemoglobin
Pcv	Packed cell volume
Wc	White blood cell count
Rc	Red blood cell count
Htn	Hypertension
Dm	Diabetes mellitus
Cad	Coronary artery disease
Appet	Appetite
Pe	Pedal edema
Ane	Anemia

5. EXPERIMENTAL RESULTS

The experimental results of classifiers are discussed in this section. For each classifier 66% of the dataset is used for training and remaining of datasets is used for Testing. The classification results of diabetes dataset and chronic kidney disease dataset are described in Table 3 and Table 4. The visualize results of accuracy for each classifiers are illustrated in Figure 1.

Table 3. Diagnosis results of diabetes dataset

Classes	Statistics	Classification Algorithms			
		Naïve Bayes	Random Forest	K-Star	PART
tested_negative	Precision	0.824	0.824	0.766	0.849
	Recall	0.843	0.871	0.826	0.758
	F-Measure	0.833	0.847	0.795	0.801
	ROC	0.854	0.838	0.736	0.792
tested_positive	Precision	0.646	0.685	0.551	0.578
	Recall	0.614	0.602	0.458	0.711
	F-Measure	0.630	0.641	0.500	0.638
	ROC	0.854	0.838	0.736	0.792
Correctly Classified Instances		201	205	185	194
Incorrectly Classified Instances		60	56	76	67
Model building time		0.03 sec	0.03 sec	0.01 sec	0.05 sec
Accuracy		77 %	**78.5 %**	70.9 %	74.3 %

Table 2 Diagnosis results of chronic kidney disease dataset

Classes	Statistics	Classification Algorithms			
		Naïve Bayes	Random Forest	KStar	PART
Ckd	Precision	1.000	1.000	0.908	1.000
	Recall	0.920	1.000	0.898	0.989
	F-Measure	0.959	1.000	0.903	0.994
	ROC	1.000	1.000	0.901	0.996
not_ckd	Precision	0.873	1.000	0.816	0.980
	Recall	1.000	1.000	0.833	1.000
	F-Measure	0.932	1.000	0.825	0.990
	ROC	1.000	1.000	0.898	0.996
Correctly Classified Instances		129	136	119	135
Incorrectly Classified Instances		7	0	17	1
Model building time		0.02 sec	0.21 sec	0.01	0.07 sec
Accuracy		94.9 %	**100 %**	87.5 %	99 %

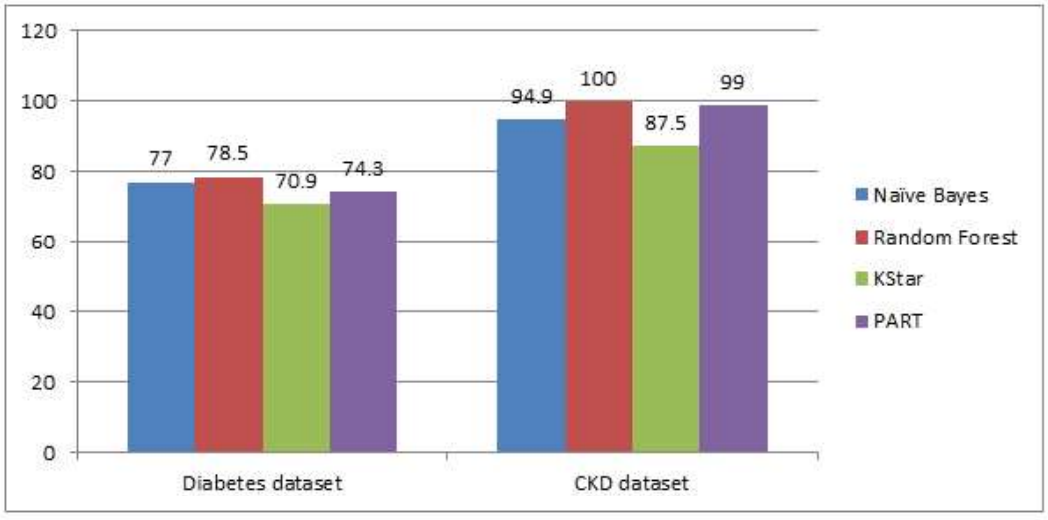

Figure 1. Accuracy results for diabetes and CKD dataset

6. CONCLUSIONS

In medical field accuracy of diagnosis result is important for both patients and doctors. Therefore to find the classification algorithm with high accuracy is important for diagnosis of medical dataset. In this paper, comparative study has been performed on the classification algorithms such as Naïve Bayes, Random forest, Kstar and PART for the UCI repository diabetes and chronic kidney disease diagnosis. In this study, the accuracy of Random Forest algorithm is high which achieve 78.5 % for diabetes dataset and 100% for chronic kidney disease.

7. ACKNOWLEDGMENTS

We would like to express our gratitude to all people for their suggestion and supports throughout this research work. We would like to express my thanks to Center for Machine Learning and Intelligent Systems at the University of California, Irvine for supporting free dataset for testing of my proposed method. Finally, we also thank all of my colleague for their participation and contribution for this study.

8. REFERENCES

[1] https://en.wikipedia.org/wiki/Diabetes

[2] https://www.betterhealth.vic.gov.au

[3] S.Yuvarani, R.Selvarani, "An Analysis of Decision Models for Diabetes", International Research Journal of Engineering and Technology(IRJET), 2016

[4] B.M Patil, R.C.Joshi, Durga Toshniwal, "Hybrid prediction model for type II diabetic patients", Expert Systems with applications, science direct, pp 8102-8108, 2012

[5] Gaganjot Kaur "Diabetes Research" Department of Computer Science and Diabetes Federation, Allagappa University, Karaikudi, India, 2005

[6] Koushik Chandra Howlader et al., "Mining Significant Features of Diabetes MellitusApplying Decision Trees: A Case Study In Bangladesh", bioRxiv, 2018

[7] Asha Gowda Karegowda et al.," Rule based Classification for Diabetic Patients using Cascaded K-Means and Decision Tree C4.5", International Journal of Computer Applications, May 2012

[8] Sakshi Saini et al., "Comparative Analysis of Classification Algorithms Using Weka", OSR Journal of Engineering (IOSRJEN), October 2018

[9] Kemal Adem, "Diagnosis of Chronic Kidney Disease using Random Subspace Method with Particle Swarm Optimization", International Journal of Engineering Research and Development, December 2018

[10] M. Praveena, N. Bhavana, "Prediction of Chronic Kidney Disease Using C4.5 Algorithm", International Journal of Recent Technology and Engineering (IJRTE), March 2019

[11] Kehinde A. Otunaiya, Garba Muhammad "Performance of Datamining Techniques in the Prediction of Chronic Kidney Disease", Computer Science and Information Technology 7(2): 48-53, 2019

[12] Pallavi Sharma, Gurmanik Kaur "Review on Data Mining Techniques for Prediction of Chronic Kidney Disease" International Journal of Engineering Trends and Technology (IJETT) – Volume 63 Number 1 – September 2018

[13] Begum Cigsar and Deniz Unal, "Comparison of Data Mining Classification Algorithms Determining the Default Risk",

[14] Theresa PrincyR and J. Thomas, "Human Heart Disease Prediction System using Data Mining Techniques" 2016 International Conference on Circuit, Power and Computing Technologies, March 2016

[15] Pratibha Devishri S et al., "Comparative Study of Classification Algorithms in Chronic Kidney Disease", International Journal of Recent Technology and Engineering (IJRTE) Volume-8, Issue-1,May 2019

[16] B.R. Gaines and P. Compton. Induction of ripple-down rules applied to modeling large databases

[17] https://archive.ics.uci.edu/