Bulletin of Electrical Engineering and Informatics
Vol. 8, No. 4, December 2019, pp. 1496~1501
ISSN: 2302-9285, DOI: 10.11591/eei.v8i4.1621

Comparative analysis of classification algorithms for chronic kidney disease diagnosis

Zainuri Saringat[1], Aida Mustapha[2], R. D. Rohmat Saedudin[3], Noor Azah Samsudin[4]
[1,2,4]Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia
[3]School of Industrial Engineering, Telkom University, 40257 Bandung, West Java, Indonesia

Article Info

Article history:

Received Apr 30, 2019
Revised Jun 25, 2019
Accepted Jul 10, 2019

Keywords:

Chronic kidney disease
Classification
Data mining

ABSTRACT

Chronic Kidney Disease (CKD) is one of the leading cause of death contributed by other illnesses such as diabetes, hypertension, lupus, anemia or weak bones that lead to bone fractures. Early prediction of CKD is important in order to contain the disesase. However, instead of predicting the severity of CKD, the objective of this paper is to predict the diagnosis of CKD based on the symptoms or attributes observed in a particular case, whether the stage is acute or chronic. To achieve this, a classification model is proposed to label stage of severity for kidney diseases patients. The experiments then investigated the performance of the proposed classification model based on eight supervised classification algorithms, which are ZeroR, Rule Induction, Support Vector Machine, Naïve Bayes, Decision Tree, Decision Stump, k-Nearest Neighbour, and Classification via Regression. The performance of the all classifiers is evaluated based on accuracy, precision, and recall. The results showed that the regression classifier perform best in the kidney diagnostic procedure.

Corresponding Author:

Zainuri Saringat,
Faculty of Computer Science and Information Technology,
Universiti Tun Hussein Onn Malaysia,
86400 Parit Raja, Batu Pahat, Johor, Malaysia.
Email: zainuri@uthm.edu.my

1. INTRODUCTION

Human kidneys function as a filtration system for blood by removing toxins from body. The toxins are circulated from kidneys to the bladder where the toxins are then dispelled from the body system through urination. Failure to filter waste from the blood causes the body to be overloaded with toxins. Kidney failure can result in death because its damage leads to urine elimination problem and causes blood loss to the kidneys. Kidney problems can be either acute or chronic, where the former refers to the sudden loss of kidney function occurs when high level of waste product of the body metabolism accumulates in the blood. The latter, Chronic Kidney Disease (CKD) is a gradual development of permanent kidney disease.

CKD happens when the kidneys are damaged or are not functioning for a long duration of time. CKD also cause malfunction of body waste system. One can develop complication like high blood pressure, anaemia, weak bones, poor nutritional health and nerve damage. Kidney disease is also known to increase a person's risk of having heart and blood vessel diseases. Chronic kidney disease may be triggered by diabetes, high blood pressure and other disorders that developed over a long period of time. Kidney disease can lead to fatalities as it will eventually lead to kidney failure, which requires dialysis or a kidney transplant.

Major contributors to CKD include diabetes, hypertension, lupus and complications from other medications such as anemia or weak bones that lead to bone fractures. Blood and urine test, ultrasound and other tests can check the status of kidney function. Early detection and treatment are important to contain the CKD. The severity of CKD is classified into five stages [1, 2]. Stage 1 is the mildest stage and can be

identified with only few symptoms. Stage 5 on the other hand is considered severe and can potentially threatens life if untreated. Stage 5 of CKD is also known as end-stage kidney disease, renal disease, or kidney failure, and is synonymous with the now outdated terms chronic renal failure or chronic kidney failure.

In this modern era, healthcare industries are in need of tools to assist in identifying smarter treatment methodologies, hospital infection control and effective hospital resource management. Medical data mining has wide explored different types of data mining tasks such as classification, clustering, and association rule mining. Therefore, it is imperative for a healthcare industry to perform in-depth analysis of massive medical data in making operational or strategic decisions. Among examples of the medical or clinical data includes the breast cancer [3], heart and coronary diseases [4, 5], liver cancer [6], diabetes [7], Parkinson's disease [8, 9], and epilepsy [10]. Specific to the Chronic Kidney Disease (CKD) classification and prediction, other data mining algorithms that have been used are the multilayer perceptron, radial basis functions network and logistic regression [11, 12], decision forest [13], time-series analysis [14], naive Bayes and artificial neural networks [15]. Instead of predicting the severity of CKD, the objective of this paper is to predict the diagnosis of CKD as shown in Figure 1 based on the symptoms or attributes observed in a particular case. The remainder of this paper proceeds as follows. Section 2 presents the research methodology, Section 3 presents the results, and finally Section 4 concludes with some indication for future work.

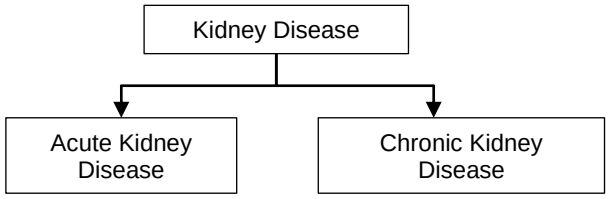

Figure 1. Types of kidney diseases

2. RESEARCH METHOD

Classification is one of data mining task that classify an instance mutually exhaustive and exclusive into one of the target target variable or class. The goal of classification is to predict the target class for each instance in the dataset. In medical data mining, a classification model can be used to label stage of severity for kidney diseases patients. Classification task is applied very frequently in medical decision making.

This paper is set to construct a classification model for Chronic Kidney Disease (CKD) based on a number of supervised learning algorithms. The methodology used is the data mining methodology that is defined as a process of revealing meaningful patterns from large database [16]. The data mining methodology is shown in Figure 2.

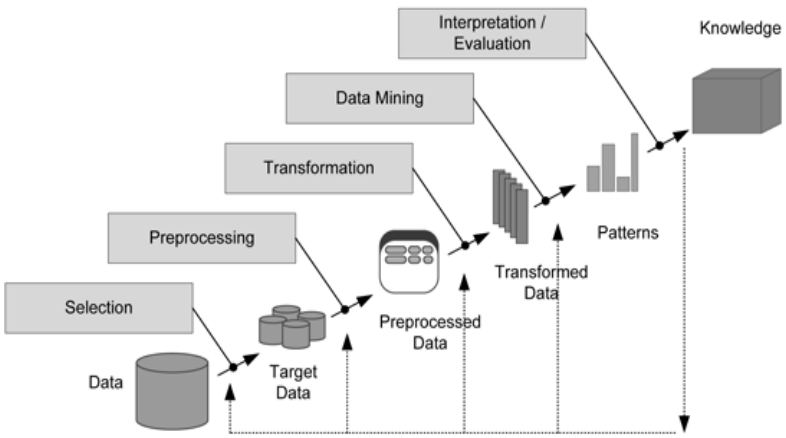

Figure 2. Data mining methodology

Eight supervised classification algorithms will be tested and compared against each other, which are ZeroR, Rule Induction, Support Vector Machine, Naïve Bayes, Decision Tree, Decision Stump, k-Nearest Neighbour, and Classification via Regression. The following subsections introduce the kidney dataset, all the classification algorithms, as well as the evaluation metrics used to evaluate the performance of the classifiers.

2.1. Dataset

The Chronic Kidney Disease Data Set is acquired from UCI Machine Learning Repository website [17]. The attributes in the dataset are both from the types nominal and numerical with two classes; CKD and NOTCKD. It contains 400 instances and 25 attributes. Data were collected nearly two months of period. During data cleaning, missing values are replaced with the average values. Next, nominal attributes are normalized and validated. Table 1 shows dataset of chronic kidney disease.

Table 1. Dataset of chronic kidney disease

Attribute	Name	Possible Values	Types
Age	Age	Age in years	Numerical
Bp	Blood Pressure	Bi in mm/Hg	Numerical
Sg	Specific Gravity	1.005,1.010,1.015,1.020,1.025	Nominal
Al	Albumin	0,1,2,3,4,5	Nominal
Su	Sugar	0,1,2,3,4,5	Nominal
Rbc	Red Blood Cells	normal, abnormal	Nominal
Pc	Pus cells	normal, abnormal	Nominal
Pcc	Pus Cell Clumps	present, not present	Nominal
Ba	Bacteria	present, not present	Nominal
Bgr	Blood Glucose Random	bgr in mgs/dl	Numerical
Bu	Blood Urea	bu in mgs/dl	Numerical
Sc	Serum Creatinine	sc in mgs/dl	Numerical
Sod	Sodium	sod in mEq/L	Numerical
Pot	Potassium	pot in mEq/L	Numerical
Hemo	Hemoglobin	hemo in gms	Numerical
Pcv	Packed Cell Volume	-	Numerical
Wc	White Blood Cell Count	wc in cells/cumm	Numerical
Rc	Red Blood Cell Count	rc in millions/cmm	Numerical
Htn	Hypertension	yes, no	Nominal
dm	Diabetes Mellitus	yes, no	Nominal
cad	Coronary Artery Disease	yes, no	Nominal
appet	Appetite	good, poor	Nominal
pe	Pedal Edema	yes, no	Nominal
ane	Anemia	yes, no	Nominal
class	Class	ckd, not ckd	Nominal

2.2. Evaluation metrics

In this paper, the performance of the classification algorithms is measured in terms of accuracy, precision, and recall.

- Accuracy–Accuracy is the proportion from the set of tuples correctly being classified by the classifier. The formula for accuracy is shown in (1).

$$\text{accuracy} = \frac{TP+TN}{\text{TP+FP +TN + FN}} \quad (1)$$

- Precision–Precision is the proportion of the true positives (TP) against all the positive results including the false positives. Precision is shown in (2).

$$\text{precision} = \frac{TP}{\text{TP+FP}} \quad (2)$$

- Recall–Recall or Sensitivity is also referred as true positive (TP) rate. This is essentially the proportion of positive tuples that are correctly identified. TP is the true positive rate, TN is true negative rate, FP is false positive rate and FN is false negative rate. Recall is shown in (3).

$$\text{recall} = \frac{TP}{\text{TP+FN}} \quad (3)$$

2.3. Classification algorithms

Six supervised classification algorithms are chosen for the purpose of benchmarking the results. The classifiers can be catergorized in rule-based, plane-based, Bayesian approach, tree-based, lazy learning, and regression-based.

2.3.1. Rule-based classifier

Two rule-based classifiers are used, the ZeroR classifier as the benchmark classifier and the Rule Induction. ZeroR is the simplest classification method that ignore all predictors and depends on the target class by means of majority class. The next rule-based algorithm used is the Rules induction, which is one of the most fundamental data mining algorithm. The dataset is represented with instances or cases in rows and attributes as columns. In this paper, the decision value is assigned by an expert to each case. Attributes are independent and the decision is a dependent variable [18].

2.3.2. Support Vector Machine

Support Vector Machine (SVM) is a plane-based classifier that create a discrete hyperplane in the descriptor space of the training data and compounds. The instances or cases are classified based on the side of the hyperplane. In a linearly separable dataset, SVM separates the hyperplane that passes through the middle of the two classes, separating the two. Once this linear function is determined, new data instance x_n can be classified by plotting the function $f = x_n$; x_n belongs to the positive class if $f = x_n > 0$ [18]. The advantage of SVM is that the distance between an instance across the hyperplane can be calculated in a nonlinear feature space, therefore does not require explicit transformation of the original descriptors.

2.3.3. Naïve Bayes

Naïve Bayes is a simple probabilistic classifier based on Bayes Theorem known for its independent assumption. Naive Bayes assumes that the presence (or absence) of an attribute is unrelated to the presence (or absence) of any other attributes. This independant assumption is beneficial because the algorithm only requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. Because independent variables are assumed, only the variances of the variables for each class need to be determined and not the entire covariance matrix. Naive Bayes learn rapidly in various supervised classification problems [19].

2.3.4. Tree-based

In this paper, two tree-based classification are applied, which are decision tree and deicion stump. Decision tree is a tree-based algorithm that nodal structures including root, branches and leaves. Each internal nodes denotes a test on an attribute, each branch denotes the outcome of a test and each node holds a class label. Decision tree are commonly used in decision analysis to help identify a strategy most likely to reach a goal.

Meanwhile, decision stump is a classification algorithm that only has one-level of decision tree. It only has one internal node or the root that is connected to the terminal nodes or leaves. Decision stump classify and instance only based on the value of just a single input feature. This model also perform unexpectedly well on some normally benchmark datasets from the UCI repository which shows that learners with a high bias and low variance may perform well because they are less prone to overfitting [20].

2.3.5. k-Nearest neighbor

Nearest Neighbor classification is a type of lazy learning and is mainly used when all attribute values are continuous, although it can be modified to deal with categorical attributes [21]. *k*-Nearest Neighbor (*k*-NN) classification predict the class membership of the given instances. An object is assigned to its *k* nearest neighbors, where *k* is typically a small integer number such as 1 or 2. The assignment is based on majority vote of its *k* neighbors.

2.3.6. Regression

ClassificationViaRegression is a classification algorithm that is based on linear regression to perform classification [22]. In this algorithm, each class is binarized and the algorithm will build a regression model for each class.

3. RESULTS AND ANALYSIS

Table 2 presents results of the experiment across eight classifiers in terms of accuracy, precision and recall. All of the classifiers achieve more than 90% accuracy except ZeroR. Although Rule Induction and

Support Vector Machine classifiers have achieved more than 90% accuracy, the results show less than 50% in terms of precision and recall. The Naïve Bayes, Decision Tree, and k-nearest neighbour classifiers achieve less than 70% in terms of precision and recall. For the tree-based algorithms, both classifiers achieve more than 90% accuracy, however, the decision stump has higher percentage than the decision tree in terms of precision and recall. The regression algorithm is the only classifier that has achieved more than 97% in terms of accuracy, precision, and recall. Overall, the results suggest that the regression classifier perform best in the kidney diagnostic procedure. Figure 3 shows the excerpt of prediction results for chronic kidney disease cases based on decision tree only.

Table 2. Comparison of results

Category	Classification Algorithms	Accuracy (%)	Precision (%)	Recall (%)
Rule-based	ZeroR	62.50	62.50	100.00
	Rule Induction	92.50	46.27	46.11
Plane-based	Support Vector Machine	90.25	44.79	46.37
Bayesian	Naive Bayes	98.50	65.40	66.00
Tree-based	Decision Tree	95.50	63.41	63.78
	Decision Stump	92.00	97.40	89.60
Lazy learning	*k*-Nearest Neighbor	94.75	62.73	63.91
Regression	ClassificationViaRegression	98.25	99.60	97.60

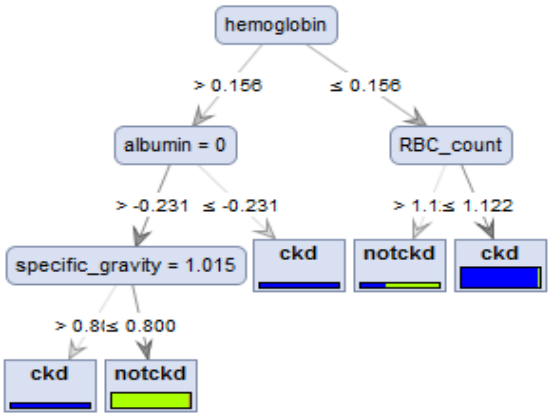

Figure 3. Excerpt of decision model for CKD

4. CONCLUSION

This paper compared eight classification algorithms: ZeroR, Rule Induction, Support Vector Machine, Naïve Bayes, Decision Tree, Decision Stump, k-Nearest Neighbour, and Classification via Regression in predicting the Chronic Kidney Disease (CKD). The obtained result of this experiment shows that the conventional Naive Bayes performed the highest despite its independent assumptions on the variables due to nature of the dataset. In the future, further investigations will be made in feature context, with possibility of employing association rule mining, case-based reasoning or ontology to select best correlated and impactful CKD features.

ACKNOWLEDGEMENTS

This paper is funded by International Grant Scheme vot W004 at Universiti Tun Hussein Onn Malaysia.

REFERENCES

[1] Levey AS, Coresh J, Balk E, Kausz AT, Levin A, Steffes MW, Hogg RJ, Perrone RD, Lau J, Eknoyan G. "National Kidney Foundation practice guidelines for chronic kidney disease: evaluation, classification, and stratification". *Annals of internal medicine*. 2003 Jul 15;139(2):137-47.

[2] Levey AS, Eckardt KU, Tsukamoto Y, Levin A, Coresh J, Rossert J, Zeeuw DD, Hostetter TH, Lameire N, Eknoyan G. "Definition and classification of chronic kidney disease: a position statement from Kidney Disease: Improving Global Outcomes (KDIGO)". Kidney international. 2005 Jun 1;67(6):2089-100. Tsai JH. Data Mining for DNA Viruses with Breast Cancer and its Limitation. INTECH Open Access Publisher. 2008.
[3] H. G. Lee, K. Y. Noh and K. H. Ryu, "A Data Mining Approach for Coronary Heart Disease Prediction using HRV Features and Carotid Arterial Wall Thickness," *2008 International Conference on BioMedical Engineering and Informatics*, Sanya, 2008, pp. 200-206.
[4] D. Pal, C. Chakraborty and K. M. Mandana, "Data mining approach for coronary artery disease screening," *2011 International Conference on Image Information Processing*, Shimla, 2011, pp. 1-6.
[5] S. Bahramirad, A. Mustapha and M. Eshraghi, "Classification of liver disease diagnosis: A comparative study," *2013 Second International Conference on Informatics & Applications (ICIA)*, Lodz, 2013, pp. 42-46.
[6] Ibrahim NH, Mustapha A, Rosli R, Helmee NH. "A Hybrid Model of Hierarchical Clustering and Decision Tree for Rule-based Classification of Diabetic Patients". *International Journal of Engineering and Technology*. 2013, Vol. 5.
[7] G. Yadav, Y. Kumar and G. Sahoo, "Predication of Parkinson's disease using data mining methods: A comparative analysis of tree, statistical and support vector machine classifiers," *2012 National Conference On Computing And Communication Systems*, Durgapur, 2012, pp. 1-8.
[8] Ramani RG, Sivagami G. "Parkinson disease classification using data mining algorithms". *International journal of computer applications*. 2011 Oct; 32(9):17-22.
[9] Ghannad-Rezaie M, Soltanian-Zadeh H. "Interactive knowledge discovery for temporal lobe epilepsy". INTECH Open Access Publisher. 2008.
[10] Rubini LJ, Eswaran P. "Generating comparative analysis of early stage prediction of Chronic Kidney Disease". *International Journal of Modern Engineering Research.* 2015, 5(7), pp. 49-55.
[11] Kunwar V, Chandel K, Sabitha, AI, Bansal A. "Chronic Kidney Disease Analysis Using Data Mining Classification". INTECH Open Access Publisher. 2016.
[12] W. H. S. D. Gunarathne, K. D. M. Perera and K. A. D. C. P. Kahandawaarachchi, "Performance Evaluation on Machine Learning Classification Techniques for Disease Classification and Forecasting through Data Analytics for Chronic Kidney Disease (CKD)," *2017 IEEE 17th International Conference on Bioinformatics and Bioengineering (BIBE)*, Washington, DC, 2017, pp. 291-296.
[13] Perotte A, Ranganath R, Hirsch JS, Blei D, Elhadad N. "Risk prediction for chronic kidney disease progression using heterogeneous electronic health record data and time series analysis". *Journal of the American Medical Informatics Association*. 2015, 22(4), pp. 872-880.
[14] V. Kunwar, K. Chandel, A. S. Sabitha and A. Bansal, "Chronic Kidney Disease analysis using data mining classification techniques," *2016 6th International Conference-Cloud System and Big Data Engineering (Confluence)*, Noida, 2016, pp. 300-305.
[15] Han J, Kamber M. "Data Mining Concepts and Techniques". San Francisco, CA: Morgan Kaufmann. 2006.
[16] Padmanaban KA, Parthiban G. "Applying Machine Learning Techniques for Predicting the Risk of Chronic Kidney Disease". *Indian Journal of Science and Technology*. 2016, 9(29).
[17] Dua D, Graff, C. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml]. Irvine, CA: University of California, School of Information and Computer Science, 2019.
[18] Wu X, Kumar V, Quinlann J, Ghosh J, Yang Q, Motoda H. "Top 10 Algorithm in Data Mining" *IEEE International Conference on Data Mining (ICDM)*. London: Springer-Verlag. pp. 1-37.
[19] Dimitoglou G, Adams JA, Jim CM. "Comparison of the C4. 5 and a Naïve Bayes classifier for the prediction of lung cancer survivability". arXiv preprint arXiv:1206.1121. 2012.
[20] Sammut C, Webb GI. (Eds.). Encyclopedia of machine learning. Springer Science & Business Media. 2011.
[21] Bala S, Kumar K. "A literature review on kidney disease prediction using data mining classification technique". *International Journal of Computer Science and Mobile Computing*. 2014 Jul; 3(7):960-7.
[22] Jena L, Kamila NK. "Distributed data mining classification algorithms for prediction of chronic-kidney-disease". *Int. J. Emerg. Res. Manag. &Technology*. 2015 Nov; 9359(11): 110-8.