Classification of Chronic Kidney Disease (CKD) Using Rule based Classifier and PCA

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

Abstract

Feature selection is a process which attempts to minimize the problems caused by high feature dimensionalities. This is normally achieved in feature extraction step in data mining. The main task of feature extraction is to select or combine the features that preserve most of the information and remove the redundant components in order to improve the efficiency of the subsequent classifiers without degrading their performances. For feature selection, Principal Component Analysis (PCA) is used in this paper. PCA is the one of the most popular methods for feature selection. The rule based approach is most useful in the classification problem. In this paper, rule based classification algorithms, namely PART, RIDOR and JRIP are used for classification of Chronic Kidney Disease (CKD) dataset. The classification results of normal rule based algorithm and feature selection based classification results are compared and analyzed.

Keywords: Feature selection, Rule based Classifier, PCA, PART, RIDOR, JRIP, CKD.

1. Introduction

A large number of algorithms and data mining tools have been developed and implemented for feature selection and Classification. Feature selection is a closely related to dimension reduction. The objective of feature selection is to identify features in the data-set as important, and discard any other irrelevant and redundant feature. Since feature selection reduces the dimensionality of the data, it holds out the possibility of more effective and rapid operation of data mining algorithm.

Classification is a supervised procedure that learns to classify new instances based on the knowledge learnt from a previously classified training set of instances. It takes a set of data already divided into predefined groups and searches for patterns in the data that differentiate those groups supervised learning, pattern recognition and prediction. Typical Classification Algorithms are Decision trees, rule-based induction, neural networks, genetic algorithms and Bayesian networks. Rule based classification algorithm also known as separate-and-conquer method is an iterative process consisting in first generating a rule that covers a subset of the training examples and then removing all examples covered by the rule from the training set. This process is repeated iteratively until there are no examples left to cover [1].

Data mining task is useful for several applications area. It is also important for health care to discover trends in patient data in order to make better their health [2]. Chronic Kidney Disease (CKD) has become a superior cause of deaths recent decades. The advantages of data mining have been creating a scope of research in health care informatics (HCI) [3]. In this study, three rule based algorithms and PCA based feature selection method is used for classification of chronic kidney disease. The popular data mining tool WEKA is used to compare those techniques. The rest of the paper is organized as follows. Section 2 review the related work and section 3 presents the overview of chronic kidney disease. We describe the feature selection algorithm PCA and rule based classifier in section 4 and 5. The empirical results are presented in section 5. Finally, we conclude this study in section 7.

2. Literature Review

There are many research works that proposed rule based classifier and feature selection technique for efficient classification. Lakshmi Devasena C., proposed the rule based classifier algorithm namely, RIDOR, ZeroR and PART Classifiers for credit risk prediction. Using the open source machine learning tool the test is completed. RIDOR Classifier performs best, followed by PART Classifier and then by ZeroR Classifier for credit risk prediction by taking various measures [4].

M. Thangaraj et al., presented rule based classifier across multiple database relations using tuple-id propagation technique. The overall position is done based on the number of relations, number of tuples, number of attributes, number of foreign keys, classification accuracy and runtime. Based on the results, PART Classifier appears to be superior to Decision tree, RIPPER and RIDOR [5]. The paper [7] gives an assessment of rule-based classifiers for Iris data set from UCI machine learning repository using an open source machine learning tool WEKA. The classification accuracy, mean absolute error and root mean squared error are calculated for each machine learning algorithm.

Włodzisław Duch [6] provided the introduction to the Rule based system. Rules descried in paper are the one used in classification (Classification, Machine Learning), regression (Regression, Statistics) and association tasks. Paper explained the various forms of rules that allow expression of different types of knowledge classical prepositional logic (C-rules), association rules (Arules), fuzzy logic (F-rules), M-of-N or threshold rules (T-rules) and prototype-based rules (P-rules) .All these types of rules are explained in detail with their advantages and disadvantages. S. Vijayarani and M. Muthulakshmi [8] provided the classification of the computer files based on their extension category using classification rule techniques.

Manmeet Kaur [9] proposed classification of patents by using the text mining approach based On principal component analysis and Logistics. Fungun Kuang et al. [10] presented the intruder detection system using the kernel principal component analysis, SVM and GA model. In the research article [11], authors introduced an ANFISAdaptive Neuro Fuzzy Inference System for to detect the CKD-Chronic Kinney Disease based on real medical data. Dr. S. Vijayaran [12], proposed the algorithms for kidney disease classification using naïve bayes and support vector machine and compare the algorithm performance. In [13] the author proposed the comparative study of chronic kidney disease classification using KNN and SVM. In recent year, there are many other approaches and algorithms are invented for feature selection and classification tasks.

3. Chronic kidney Disease

Chronic kidney disease (CKD) is a type of kidney disease in which there is gradual loss of kidney function over a period of months or years. Early on there are typically no symptoms. Later, leg swelling, feeling tired, vomiting, loss of appetite, or confusion may develop. Complications may include heart disease, high blood pressure, bone disease, or anemia.

Causes of chronic kidney disease include diabetes, high blood pressure, glomerulonephritis, and polycystic kidney disease. Risk factors include a family history of the condition. Diagnosis is generally by blood tests to measure the glomerular filtration rate and urine tests to measure albumin. Further tests such as an ultrasound or kidney biopsy may be done to determine the underlying cause. A number of different classification systems exist [20].

Screening at-risk people is recommended. Initial treatments may include medications to manage blood pressure, blood sugar, and lower cholesterol. NSAIDs should be avoided. Other recommended measures include staying active and certain dietary changes. Severe disease may require hemodialysis, peritoneal dialysis, or a kidney transplant. Treatments for anemia and bone disease may also be required.

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. The causes that contribute to the greatest number of deaths are high blood pressure at 550,000, followed by diabetes at 418,000, and glomerulonephritis at 238,000.

4. PRINCIPAL COMPONENT ANALYSIS (PCA)

The main advantage of PCA is that once these patterns are found in the data, the data is then compressed without much loss of information. It is widely used in most of the pattern recognition applications like face recognition, image compression, and for finding patterns in high dimensional data. Principal component analysis (PCA) is an essential technique in data compression and feature extraction [14]. It is well known that PCA has been widely used in data compression and feature selection.

Overview of PCA is briefly described as follows.

Assume that $\{x_t\}$ where t =1, 2..., N are stochastic n dimensional input data records with mean (μ). It is defined by the following Equation:

$$\mu = \frac{1}{N}\sum_{t=1}^{N} x_t \qquad (1)$$

The covariance matrix of x_t is defined by

$$C = \frac{1}{N}\sum_{t=1}^{N}(x_t - \mu)(x_t - \mu)^T \qquad (2)$$

PCA solves the following eigenvalue problem of covariance matrix C:

$$C_{vi} = \lambda_i v_i \qquad (3)$$

where λ_i (i =1 ,2,...,n) are the eigenvalues and v_i(i = 1,2,...,n) are the corresponding eigenvectors. To represent data records with low dimensional vectors, we only need to compute the m eigenvectors (called principal directions) corresponding to those m largest eigenvalues (m<n). It is well known that the variance of the projections of the input data onto the principal direction is greater than that of any other directions.

Let

$$\varphi =[v_1,v_2,.....,v_m],\ \Lambda= diag[\lambda_1,\lambda_2,.....,\lambda_m] \qquad (4)$$

Then

$$C\Phi = \Phi\Lambda \qquad (5)$$

The parameter v denotes to the approximation precision of the m largest eigenvectors so that the following relation holds.

$$\frac{\sum_{i=1}^{m} \lambda i}{\sum_{i=1}^{n} \lambda i} \geq v \qquad (6)$$

Based on (5) and (6) the number of eigenvectors can be selected and given a precision parameter v, the low dimensional feature vector of a new input data x is determined by

$$x_f = \Phi^T x \qquad (7)$$

5. Rule based Classifiers

Basically, in rule based classification, classification model is represented in the form of IF-THEN rules. Such rules have two part, the IF part is known as the rule antecedent and the else part is known as consequent. This structure tells that the antecedent part covers the tuples which satisfies the rule and part of consequent covers the tuples which does not satisfy the rule. Coverage and accuracy are the measures used to assess the rules. The percentage of tuple covered under rule is the coverage of that rule and percentage of correctness is the accuracy of that rule. This IF-THEN classification rules can be extracted from decision tree as this rules are easily interpreted by humans.

PART Algorithm (Rule-based Classification algorithm): Full form of PART is Projective Adaptive Resonance Theory [15]. PART is refined method of rule generation [16]. After rule generation entire tree generated, the best tree is selected and its leaves are translated into rules. PART support all type of classes like Binary and Nominal class and supports all type of attributes. PART is a separate-and-conquer rule learner proposed by Eibe and Witten [19]. 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.

RIDOR Algorithm (Rule-based Classification algorithm): The algorithm was introduced by Compton and Jansen, ripple-down rule technique as a methodology for the acquisition and maintenance of large rule-based systems [19]. Ridor algorithm is the implementation of a RippleDown Rule learner proposed by Gaines and Compton. RIDOR learns rules with exceptions by generating the default rule, using incremental reduced error pruning to find exceptions with the smallest error rate, finding the best exceptions for each exception, and iterating. The exceptions are a set of rules that predict classes other than the default. IREP is used to generate the exceptions.

JRIP Algorithm (Rule-based Classification algorithm): JRIP sometimes called as RIPPER is one of popular classifier algorithm [17][18]. In JRIP instances of the dataset are evaluated in increasing order, for given dataset of threat a set of rules are generated. JRIP (RIPPER) algorithm treats each dataset of given database and generates a set of rules including all the attributes of the class. Then next class will get evaluated and does the same process as previous class, this process continues until all the classes have been covered.

6. Results and Discussions

The above three algorithms are compared using dataset namely Chronic Kidney Disease. These dataset are collected from UCI Repository in the website www.ucirepository.com. The dataset contains 400 instances and 25 attributes. The WEKA application is used for the evaluation. For each classifier 2/3 of the dataset is used for training and 1/3 of datasets is used for. The accuracy parameters and error rate of test results are described in the following table 1 and table 2.

Table 1. Accuracy Parameters for CKD Evaluation

Algorithm	TP Rate	FP Rate	Precision	Recall
PART	0.989	0.000	1.000	0.989
RIDOR	0.977	0.000	1.000	0.977
JRIP	0.955	0.042	0.977	0.955
PCA+ PART	1.000	0.000	1.000	1.000
PCA+ RIDOR	0.989	0.000	1.000	0.989
PCA+ JRIP	1.000	0.021	0.989	1.000

The error measurement four parameters are considered as:
(i) RMSE Root mean squared error
(ii) MAE Mean absolute error
(iii) RRSE Root relative squared error
(iv) RAE Relative absolute error

Table 2. Error rate for CKD Evaluation

Algorithm	RMSE	MAE	RRSE	RAE
PART	0.1225	0.0392	25.5737	8.3948
RIDOR	0.1213	0.0147	25.3109	3.1502
JRIP	0.2051	0.0553	42.8033	11.8563
PCA+ PART	0	0	0	0
PCA+ RIDOR	0.0857	0.0074	17.8975	1.5751
PCA+ JRIP	0.0861	0.018	17.9755	3.8637

Table 3. Accuracy results of CKD Evaluation

PART	RIDOR	JRIP	PCA+ PART	PCA+ RIDOR	PCA+ JRIP
99.2647 %	98.5294 %	95.5882 %	100 %	99.2647 %	99.2647 %

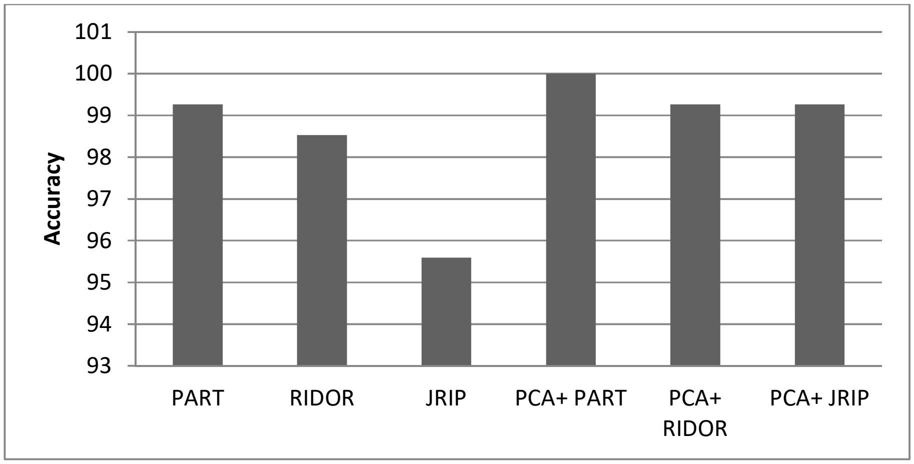

Figure 1. Visualize results for accuracy of different classifiers

7. Conclusion

The comparative analysis of the three rule based algorithms for Chronic Kidney Disease (CKD) classification is presented in this paper. The results of experiment are presented in tabular and graphical form. From this study it is found that combination of PCA and PART is best algorithm for CKD classification.

REFERENCES

[1] Phyu, Thair Nu. "Survey of classification techniques in data mining." Proceedings of the International Multi Conference of Engineers and Computer Scientists. Vol. 1. 2009.

[2] A. K. Sen, S. B. Patel, and D. D. Shukla, “A data mining technique for prediction of coronary heart disease using neuro-fuzzy integrated approach two level,” International Journal Of Engineering And Computer Science ISSN, 2013, pp. 2319–7242.

[3] M. Herland, T. M. Khoshgoftaar, and R. Wald, “A review of data mining using big data in health informatics,” Journal of Big Data, vol. 1, no. 1, p. 1, 2014.

[4] C. Lakshmi Devasena, T. Sumathi, V.V. Gomathi and M. Hemalatha," Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set", Bonfring International Journal of Man Machine Interface, Vol. 1, Special Issue, December 2011

[5] M. Thangaraj, C.R.Vijayalakshmi. Performance Study on Rule based Classification Techniques across Multiple Database Relations. International Journal of Applied Information Systems (IJAIS), Foundation of Computer Science FCS, New York, USA Volume 5–No.4, March 2013.

[6] Duch,Włodzisław " Rule discovery" Encyclopedia of Systems Biology 2013, pp 1879-1883.

[7] Lakshmi Devasena, J. Effectiveness Evaluation of Rule Based Classifiers for the Classification of Iris Data Set, Bonfring International Journal of Man Machine Interface, Vol. 1, Special Issue, December, 2011.

[8] Vijayarani1, S, M. Muthulakshmi ."Evaluating The Efficiency O f Rule Techniques for File Classification". International Journal of Research in Engineering and Technology eISSN: 2319-1163 |pISSN: 2321 -7308.

[9] Manmeet Kaur, Richa Sapra" International Journal of Engineering and Advanced Technology (IJEAT)", Volume-2, Issue-4, April 20, ISSN: 2249 – 8958, 711-714

[10] Fungun Kuang, Weihong Xu, Siyang Zhang,"Elsevier Journal, Applied Soft Computing, 18, 2014, 178-84

[11] J. Norouzi, A. Yadollahpour, S. A. Mirbagheri, M. M. Mazdeh, and S. A. Hosseini, "Predicting Renal Failure Progression in Chronic Kidney Disease Using Integrated Intelligent Fuzzy Expert System," Computational and Mathematical Methods in Medicine, vol. 2016, pp. 1 –9.

[12] V. S and D. S, "Data Mining Classification Algorithms for Kidney Disease Prediction," International Journal on Cybernetics & Informatics, vol. 4, no. 4, Aug. 2015, pp. 13–25.

[13] Comparative Study of Chronic Kidney Disease Classification Using KNN and SVM. Internation Journal of Engineering Research & Technology, Vol-4, Dec. 2015, 608-612

[14] E. Oja, "Principal components, minor components, and linear neural networks" Neural Networks, vol. 5, , 1992, pp. 927-935.

[15] Pinto C M A, Machado J A T. Fractional Dynamics of Computer Virus Propagation. Mathematical Problems in Engineering, 2014: 259-305.

[16] Himadri Chauhan, Vipin Kumar, Sumit Pundir and Emmanuel S. Pilli 2013 International Symposium on Computational and Business Intelligence "A Comparative Study of Classification Techniques for Intrusion Detection", department of Computer Science and Engineering Graphic Era University Dehradun India DOI: 10.1109/ISCBI.2013.

[17] Ripple Down Rule learner (RIDOR) Classifier for IRIS Dataset V. Veeralakshmi et al., International Journal of Computer Science Engineering (IJCSE) ISSN : 2319-7323 Vol. 4 No.03 May 2015

[18] Neeraj Bhargava, Sonia Dayma, Abhishek Kumar, Pramod Singh IEEE Sponsored 3rd International Conference on Electronics and Communication Systems (ICECS 2016) An Approach for Classification using Simple CART Algorithm in Weka, MDS University Ajmer India, 2016.

[19] http://archive.ics.uci.edu/ml

[20] https://en.wikipedia.org/wiki/Chronic_kidney_disease