Feature Selection System for the Classification of Chronic Kidney Disease Based on Information Gain and Pearson Correlation

Yi Mon Aung
Faculty of Information Science
University of Computer Studies (Magway)
yimonaung@ucsmgy.edu.mm

Abstract

Recently, large amounts of data are widely accessible in information structures, and students have attracted great attention of academics to turn such data into useful knowledge. Researchers also need related data from vast records of patients using the feature selection method. The selection of features is the process of identifying the most important attributes and removing redundant and irrelevant attributes. The system implements obtaining information in the original data set without taking into account class labels to create the best functional subset using information gain and Pearson correlation. This system was tested on a chronic kidney disease dataset obtained from the UCI machine learning repository. The best data on chronic kidney disease with all the functions using data mining algorithms such as Naive Bayes, Multilayer Perceptron (MLP), J48, K Nearest Neighbor (KNN) to test the effectiveness of the system.

1. Introduction

Modern clinical information systems supply numerous data in medical databases. This facilitates the mining of useful knowledge from medical databases and arrange for valuable information to support medical decision making. The field of data mining, known as medical data mining, is currently measured one of the most popular research topics in the data mining community.

Chronic kidney disease (CKD) is a major public wellbeing problem worldwide, particularly in low- and middle-income nations. CKD means that your kidneys do not work as you expect and cannot properly filter your blood. About 10% of the world's population feel pain from CKD, killing millions of people each year due to a lack of affordable treatment, and this number is increasing among older people. In 2010, the International Association of Nephrologists identifies CKD as an important cause of death worldwide, with an increase of 82.3% over the past 20 years [4].

The main purpose of this paper is to help professionals get the best set of features from the original data set and provide accurate and timely analytic information for decision making. Because data mining is an experimental learning, making accurate predictions can be difficult. It is important to choose the appropriate method for selecting feature, because we need to maximize the accuracy of each classifier. Feature selection is very important for predictive analytics and should not be missed. This reduces lead time and provides more accurate and consistent outcomes.

The rest of this paper is organized in the following order: Section 2 presents the related works. The data mining algorithm is discussed in detail in Section 3. Section 4 contains the results and analysis. Finally, section 5 describes the paper with the conclusion.

2. Related Work

Kehinde [3] used three classifiers to predict a set of chronic kidney disease data: a multilayer perceptron, naive bayes, and a J48 decision tree. The purpose of this study is to assess the effectiveness of the classifiers used based on the following indicators for accuracy, specificity, sensitivity, error rate and precision. Based on the above performance indicators, the results show that the J48 decision tree gave the best results, but the naive bayes was the fastest classifier, because it had the shortest execution time.

S. D. Harish, K. Vinay Kumar, K. Taraka Ram and G. Pradeepini [6] conduct the analysis of various open source Python modules and output the results predicted by machine learning algorithms to determine accuracy of chronic kidney disease. It prepared for predictive modeling of kidney disease data. It compare various algorithms such as KNN and Logistic Regression, which are mainly used to classify data. This algorithm predicts a data set collected from a patient’s medical record. If someone has a chronic kidney disease, it does not primarily depend on the level of potassium in the blood.

K. B. Anusha, T. Pandu Ranga Vital, K. Sangeeta [2] performed statistical analysis, machine learning (ML) and neural network applications for the clinical Uddanam CKD dataset for the prevention and quick recognition of CKD. According to statistical analysis, CKD can be prevented in the Uddanam. According to ML analysis, if a process model is built in 0.06 seconds and a forecast accuracy of 99.9%, a naive Bayesian model is the best choice. In ANN analysis, an artificial neural network (ANN) with a 9-neural hidden layer (HL) provides 100% accuracy for predicting CKD and is much better than all other models that take 0.02 seconds of process time.

3. Methodology

Four methods of data mining: Naive Bayes, MLP, J48 and KNN were used in this study to understand which system gives better classification results in terms of accuracy.

3.1. Naïve Bayes

The Naive Bayes classifier offers high classification accuracy with performance comparable to the best decision trees and neural networks. Based on the probabilities determined from the data, it is determined that the new objects belong to classes with different degrees of probability. The Naive Bayes classification avoids this problem by assuming that the attributes are independent and that you do not have to consider combinations of attributes. In other words, the effect of attribute values on a particular class is independent of the values of other variables. The starting point for the naive Bayes classification is the Bayes theorem [1]:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \qquad (1)$$

where P(h) is the prior probability of hypothesis h, P(D) is the prior probability of training data D, P(h|D) is the probability of h given D, and P(D|h) is the probability of D given h.

3.2. Multilayer Perceptron (MLP)

A typical multilayer perceptron network (MLP) consists of a set of source nodes which form an input layer, one or more hidden layers of compute nodes and a node output layer. The input signal propagates through the network layer by layer. The signal flow of such a network with each layer hidden. Train the network and test the network performance using a direct acting multilayer propagation algorithm. MLP networks are commonly used in supervised learning problems. There is a training set of input and output pairs, and the network must learn to model the dependencies between the pairs. MLP are popular learning algorithms in the sense that neural networks recognize the desired output and adjust the weighting factors so that the calculated outputs are as close as possible [7].

3.3. Decision Tree (J48)

Decision trees can handle multidimensional data. The learning and classification steps to guide the decision tree are simple and quick. In general, decision tree classifiers are very precise. However, success depends on the available data. The decision tree induction algorithm is used for classification in many fields of application, such as medicine, manufacturing and manufacturing, financial analysis, astronomy and molecular biology. The successor to ID3, J48 (C4.5), uses an information gain extension called gain ratio which attempts to overcome this bias. Apply some sort of normalization to information gain using the defined value of "information sharing" as in Info (D) [8].

3.4. K-Nearest Neighbor (KNN)

The nearest neighbor classifier is based on similarity learning. The training samples are described by n-dimensional numerical attributes. Each sample represents a point in n-dimensional space. With this way, all training samples are stored in the n-dimensional pattern space. Given an unknown sample, the nearest neighbor classifier k searches the pattern space k for the closest learning sample to the unknown sample. "Proximity" is defined by Euclidean distance. Euclidean distance is the distance between two points. The closest k algorithm is the simplest of all machine learning algorithms. The objects are classified by majority vote in the district. The object is assigned to the most general class of k neighbors [5].

3.5. Information Gain (IG)

Acquiring information is one of the ways to select features. The selection of features is heuristic, allowing you to choose the partition criteria that best isolate a particular data partition. This measurement is based on the work of Claude Shannon on information theory. The information gain is used for the induction of decision trees using the following formula [8]:

$$Info(D) = -\sum_{i=0}^{m} p_i log_2(p_i) \qquad (2)$$

Info(D) is the average amount of information needed to identify the class label of the tuple of D. Acquisition of information is defined as the difference between the original information requirement and the new requirement. The attribute with the highest information gain (gain (A)) is selected as the split attribute. This means partitioning by the attribute "best classification".

3.6. Pearson Correlation

One of the simplest ways to understand the relationship between characteristics and response variables is the Pearson correlation coefficient [10]. It measures the linear correlation between two variables. Calculations are quick and easy and are often done first on the data. The correlation between a composite and an external variable is a function of the number of component variables in the composite, the extent of the cross-correlation between them and the extent of the correlation between the component and the external variable. The higher the correlation between the component and the external variable, the higher the correlation between the composite and the external variable. As the number of components in the composite increases (assuming the additional components are the same as the original component in terms of mean cross-correlation with other components and external variables), the correlation between the composite and external is higher.

4. Results and Discussion

4.1. Dataset

The data set was collected from the UCI machine-learning standard [9]. The data are blood tests and other measures for patients with and without CKD. These are patients who were seen in a hospital in Tamil Nadu, India, for about two months at some point before July 2015. There are 400 instances, 24 attributes and 1 class attribute. The dataset contains 400 instances (250 CKD instead of CKD) and 24 attribute counts + 24 = class = 11 numbers, 14 nominal. In this study, WEKA (version 3.8.4) was used for the selection and classification of the characteristics.

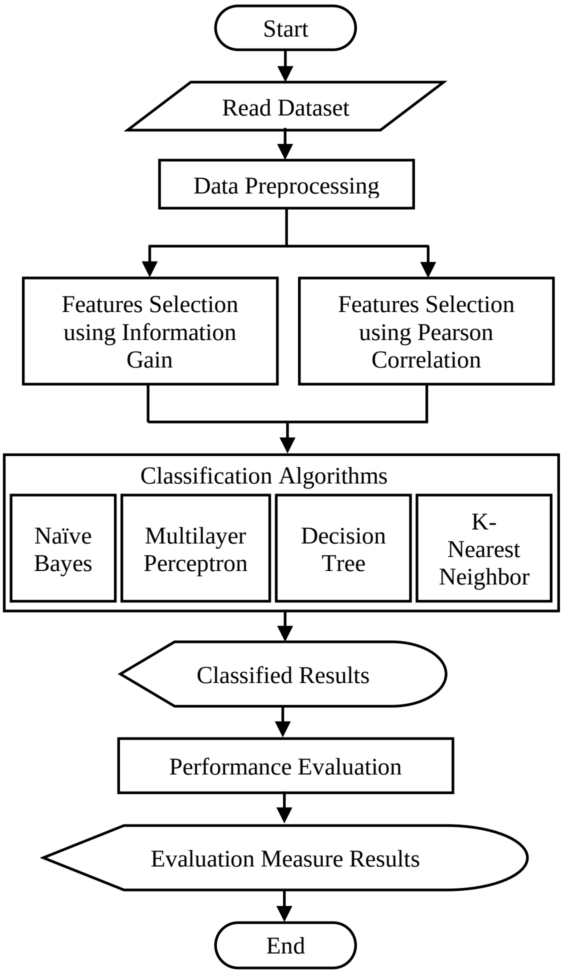

Figure 1. System Flow

Figure 1 shows the system flow diagram of classification of chronic kidney disease using features selection methods.

4.2. Accuracy

Accuracy is the most intuitive measure of performance and is simply the relationship between correctly classified or predicted perception and overall perception.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \qquad (3)$$

where TP is the number of true-positive classifications, FN is the number of false-negative classifications, TN is the number of true-negative classifications, and FP is the number of false positive classifications.

Precision is also known as positive predictive value. It is defined as the average probability of relevant searches.

$$Precision = \frac{TP}{TP + FP} \qquad (4)$$

Recall is defined as the average probability of complete retrieval.

$$Recall = \frac{TP}{TP + FN} \qquad (5)$$

4.3. Experimental Results

The table shows the experimental results for each classifier with and without the feature selection method. After all attributes have calculated the information gain, the attribute with the highest information gain is selected. The methods of gain search and classification of information and the last selected attributes classified by these classifications are displayed as follows:

Table 1. Rank of attributes based on information gain

No	Attributes	Values	Rank
1	hemo (hemoglobin)	0.5175	1
2	sc (serum creatinine)	0.4979	2
3	sg (specific gravity)	0.4486	3
4	pcv (packed cell volume)	0.4461	4
5	al (albumin)	0.3907	5
6	htn (hypertension)	0.3224	6
7	dm (diabetes mellitus)	0.2921	7
8	rbcc (red blood cell count)	0.2711	8
9	bu (blood urea)	0.2617	9
10	bgr (blood glucose random)	0.2285	10
11	sod (sodium)	0.1902	11
12	bp (blood pressure)	0.1739	12
13	appet (appetite)	0.1566	13
14	pc (pus cell)	0.1471	14

15	pe (pedal edema)	0.1434	15
16	pot (potassium)	0.1308	16
17	rbc (red blood cells)	0.1188	17
18	su (sugar)	0.1170	18
19	ane (anemia)	0.1096	19
20	age (age)	0.1046	20
21	wbcc (white blood cell count)	0.0850	21
22	pcc (pus cell clumps)	0.0692	22
23	cad (coronary artery disease)	0.0577	23
24	ba (bacteria)	0.0349	24

Once all attributes have calculated the Pearson correlation, the attribute with the highest information gain is selected. The last selected attributes and their rankings classified by Pearson's correlation search and classification methods are displayed as follows:

Table 2. Rank of attributes based on pearson correlation

No	Attributes	Values	Rank
1	hemo (hemoglobin)	0.7296	1
2	pcv (packed cell volume)	0.6901	2
3	rbcc (red blood cell count)	0.5909	3
4	htn (hypertension)	0.5904	4
5	dm (diabetes mellitus)	0.5591	5
6	al (albumin)	0.4770	6
7	bgr (blood glucose random)	0.4014	7
8	appet (appetite)	0.3933	8
9	pc (pus cell)	0.3752	9
10	pe (pedal edema)	0.3752	10
11	bu (blood urea)	0.372	11
12	sg (specific gravity)	0.3505	12
13	sod (sodium)	0..3423	13
14	ane (anemia)	0.3254	14
15	su (sugar)	0.3009	15
16	sc (serum creatinine)	0.2941	16
17	bp (blood pressure)	0.2906	17
18	rbc (red blood cells)	0.2826	18
19	pcc (pus cell clumps)	0.2653	19
20	cad (coronary artery disease)	0.2361	20
21	age (age)	0.2254	21
22	wbcc (white blood cell count)	0.2053	22
23	ba (bacteria)	0.1869	23
24	pot (potassium)	0.0769	24

Table 3 shows the information about the accuracy of all algorithms with precision and recall.

Table 3. Precision, recall and accuracy of classifiers

Classifiers	Precision	Recall	Accuracy(%)
Naive Bayes	0.956	0.950	95%
Multilayer Perceptron	0.998	0.998	99.75%
Decision tree	0.990	0.990	99%
K-nearest neighbor	0.962	0.958	95.75%

Table 3 presents a comparison of the four classifiers performed on the CKD data. The multilayer perceptron has the highest accuracy in all measurements 99.75%. The information on the four classifiers and the precision tables 4 and 5 compare the selection of the characteristics based on the information gain and the Pearson correlation.

Table 4. Precision, recall and accuracy of classifiers with information gain based features selection

Classifiers	Precision	Recall	Accuracy(%)
Naive Bayes	0.962	0.958	95.75%
Multilayer Perceptron	0.995	0.995	99.5%
Decision tree	0.990	0.990	99%
K-nearest neighbor	0.970	0.968	96.75%

Table 5. Precision, recall and accuracy of classifiers with Pearson correlation based features selection

Classifiers	Precision	Recall	Accuracy (%)
Naive Bayes	0.962	0.958	95.75%
Multilayer Perceptron	0.993	0.993	99.25%
Decision tree	0.990	0.990	99%
K-nearest neighbor	0.972	0.970	97%

From all of these comparisons, we can see that the multilayer perceptron is the most powerful classifier for this dataset. However, KNN is the most powerful classifier for predicting CKD using feature selection based on the Pearson correlation. The decision tree algorithm has the same precision value in each prediction. For naive Bayes classifiers, the selection of features based on information gain and Pearson correlation has the greatest precision.

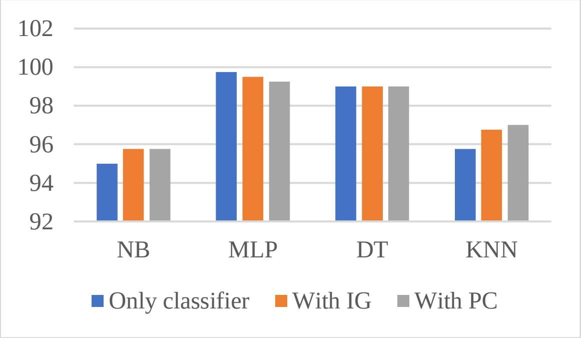

Figure 2. Accuracy evaluation

5. Conclusion

This article presents an empirical analysis of four different approaches to Naive Bayes (NB), Multilayer Perceptron (MLP), Decision Tree (J48) and K-Nearest Neighbor (KNN) with a selection of characteristics based on information gain and the Pearson correlation. Precision, recall, and accuracy are used as evaluation measures to evaluate each method. The overall results of this study recommend that the multilayer perceptron (MLP) work well in order to improve the accuracy of CKD predictions.

References

[1] J. Darzentas, G. Vouros, S. Vosinakis, A. Arnellos, “Artificial Intelligence: Theories, Models and Applications”, 5th Hellenic Conference on AI, SETN 2008, Syros, Greece, October 2-4, 2008, Proceedings (Lecture Notes in Computer Science (5138)).

[2] K. B. Anusha, T. P. R. Vital, K. Sangeeta, “Machine Learning Models and Neural Network Techniques for Predicting Uddanam CKD”, *International Journal of Recent Technology and Engineering*, ISSN: 2277-3878, Volume-8 Issue-2, July 2019.

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

[4] M. Almasoud, T. E. Ward, “Detection of Chronic Kidney Disease using Machine Learning Algorithms with Least Number of Predictors”, *International Journal of Advanced Computer Science and Applications*, Vol. 10, No. 8, 2019.

[5] R. E. Maleki, A. Rezaei, and B. M. Bidgoli, “Comparison of Classification Methods Based on the Type of Attributes and Sample Size”, *Journal of Convergence Information Technology*, DOI: 10.4156/jcit.vol4.issue3.14.

[6] S. D. Harish, K. V. Kumar, K. T. Ram, G. Pradeepini, “Chronic Kidney Disease Prediction based on Blood Potassium Levels using Machine Learning”, *International Journal of Innovative Technology and Exploring Engineering*, ISSN: 2278-3075, Volume-9 Issue-2, December 2019.

[7] https://ujjwalkarn.me/2016/08/09/quick-intro-neural-networks.

[8] https://towardsdatascience.com/decision-tree-algorithm-explained-83beb6e78ef4.

[9] https://archive.ics.uci.edu/ml/datasets/chronic-kidney disease.

[10] https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php.