Chronic Kidney Disease Prediction with Reduced Individual Classifiers

Merve Doğruyol Başar¹, Aydın Akan²

¹Department of Electrical and Electronics Engineering, Istanbul University Cerrahpaşa School of Engineering, Istanbul, Turkey
²Department of Biomedical Engineering Department, İzmir Katip Celebi University School of Engineering, İzmir, Turkey


ABSTRACT

Chronic kidney disease is a rising health problem and involves conditions that decrease the efficiency of renal functions and that damage the kidneys. Chronic kidney disease may be detected with several classification techniques, and these have been classified using various features and classifier combinations. In this study, we applied seven different classifiers (Naïve Bayes, HoeffdingTree, RandomTree, REPtree, Random Subspaces, Adaboost, and IBk) for the diagnosis of chronic kidney disease. The classification performances are evaluated with five different performance metrics, i.e., accuracy, kappa, mean absolute error (MAE), root mean square error (RMSE), and F measures. Considering the classification performance analyses of these methods, six reduced features provide a better and more rapid classification performance. Seven individual classifiers are applied to the six features and the best results are obtained using individual random tree and IBk classifiers.

Keywords: Chronic kidney disease, reduced individual classifiers, data mining, classification
cause of many related problems in the community, such as economic, social and medical problems. Ultrasound, CT scan, kidney biopsy, and a greater rate of drug consumption can cause economical and medical problems. In addition, the constant use of drugs and the ensuing complications can cause social problems as well [7].

In the last few decades, numerous studies have been conducted in order to understand and examine CKD. In the literature, many scholars have used different classification techniques for the perception of CKD. Data mining is the process of pulling out intended data from the major dataset. These techniques are used in various applications such as medical diagnosis, face recognition and data filtering.

Data mining techniques, such as clustering, classification and so on, play a big role in the extraction of unknown knowledge from the major databases. Classification is a supervised learning technique and predefines subgroups. Classification algorithm necessitates the classes to be identified based on the data attribute value. It defines the classes taking into account the characteristics of the data. The training algorithm uses these predefined specimens to determine the set of parameters required for appropriate segregation [8].

Jena and Kamila [9] predicted and analyzed kidney disease using Waikato Environment for Knowledge Analysis (Weka) tool and they used different algorithms such as Support Vector Machine (SVM), J48 classifier, Naïve Bayes classifier, Multilayer perceptron, conjunctive rule and decision table [10].

Chaudhary and Garg [11] developed a prediction system using A-priori and K-means algorithms for prediction and these algorithms were also used to predict patients with kidney failure. The authors analyzed 42 attributes of the data using machine learning tools and evaluated the data using Receiver Operating Characteristic (ROC) plots.

Baby and Vital [12] proposed a diagnosis and prediction system and analyzed the data using the Weka tool. The authors used AD Trees, J48 classifier, K-Star algorithm, Naïve Bayes classifier and Random Forest and evaluated ROC plots. Their study showed that K-Star algorithm and Random Forest classifier were the best methods for their dataset.

Sinha [13] proposed the performance evaluation of two data mining techniques. They used K-Nearest Neighbor (KNN) and SVM to compare accuracy and precision. The authors showed that the KNN classifier provided better results than SVM in terms of accuracy and precision values [13].

Vijayarani and Dhayanand [14] used SVM and Artificial Neural Network (ANN) for the diagnosis of kidney disease. The authors compared the performance of two classifiers in terms of accuracy and execution time and used Weka tool for execution. After the performance a comparison of the two data mining techniques was undertaken and the authors concluded that ANN is better than the other algorithm for the diagnosis of kidney stone.

Against this background, in this work, we use seven different classification techniques to evaluate the existence of CKD in humans. With this purpose, we focus on Naïve Bayes, Hoeffding Tree, Random Tree, REP Tree, Random Subspaces, Ada-boost and IBk classifiers and compare their accuracy, kappa, root mean square error (RMSE), mean absolute error (MAE) and F measure values.

Our analyses show that classifiers can be applied to all features and reduced features. After that, the best results are obtained using individual Random Tree and IBk classifiers using 6 best reduced features. The aim of this work is to reduce the number of classifiers and thus help early treatment of the CKD patients.

Methodology

In this section, we provide the considered methodology with the purpose of analyzing the data.

Material

Chronic Kidney Disease dataset was created for the analysis of kidney disease and obtained from “UC Irvine Machine Learning Repository” database. This dataset contains twenty-four features and four hundred instances are used in this comparative analysis. The features in the CKD dataset are age, blood pressure, specific gravity, albumin, sugar levels, serum creatinine, hemoglobin, hypertension, red blood cells, pus cell, pus cell clumps, bacteria, blood glucose, blood urea, serum creatinine, sodium, potassium, hemoglobin, packed cell volume, white blood cell count, red blood cell count, hypertension, diabetes mellitus, coronary artery disease, appetite, pedal edema and anemia.

Our features contain 11 numeric and 13 nominal parts. Table 1 shows the description of the features in CKD. These factors are closely associated with kidney disease [15].

Kidney Disease Factors

Chronic Kidney Disease occurs slowly and gradually over many years. In deficient kidney function, the conditions cannot operate as in their former stages. Five stages of kidney disease can be given. As shown in Table 2, the GFR number is a reference for the kidney function activation. If the GFR number decreases, kidney disease becomes worse [16].

Classification Techniques

In this work, we use seven different classification techniques for the CKD data [17].

Naïve Bayes: The Naïve Bayes algorithm is based on Bayes theorem and is a probabilistic classifier. It calculates a set of proba-
basar and akan prediction of kidney disease by using reduced classifiers

Electrica 2018; 18(2): 249-255

Basaר and Akan. Prediction of Kidney Disease by Using Reduced Classifiers

bilities by calculating the frequency and combinations of values in the data. It is exemplified by a vector \( X = (x_1, x_2, \ldots, x_n) \) representing \( n \) features (independent variables) and appoints to this sample probabilities \( p(C_k | x_1, \ldots, x_n) \) for each of \( k \) possible outcomes or classes.

Using Bayes’ theorem, the conditional probability can be decomposed as follows:

\[
p(C_k | X) = \frac{p(C_k)p(X | C_k)}{p(X)} \tag{1}
\]

The Naive Bayes algorithm performs well and learns rapidly in several supervised classification processes.

Hoeffding Tree

A Hoeffding tree is an algorithm that is competent at learning from a big dataset. Hoeffding trees take advantage of small samples to choose a proper splitting feature. This idea is assisted by the Hoeffding bound, which quantifies the number of observations needed to estimate some statistics within a detected precision. The bound states probability is \( 1 - d \), the mean of a random variable of range \( R \) will not differ from the estimated mean after \( n \) independent observations by more than,

\[
e = \sqrt{\frac{R^2 \ln(1 / d)}{2n}} \tag{2}
\]

where \( R \) is the base of the number of possible classes, \( n \) is the number of observations and \( e \) is the bound.

Random Tree (RT)

Random Tree is a supervised classifier and an ensemble learning algorithm. RT can generate many individual learners and cope with both classification and regression problems.

Table 1. Feature properties of CKD data

<table>
<thead>
<tr>
<th>Features</th>
<th>Type</th>
<th>Abbr.</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>num.</td>
<td>age</td>
<td>Avg. 51.5</td>
</tr>
<tr>
<td>Blood pressure (mm/Hg)</td>
<td>num.</td>
<td>bp</td>
<td>76.5</td>
</tr>
<tr>
<td>Blood glucose (mgs/dl)</td>
<td>num.</td>
<td>bgr</td>
<td>148.04</td>
</tr>
<tr>
<td>Blood urea (mgs/dl)</td>
<td>num.</td>
<td>bu</td>
<td>57.43</td>
</tr>
<tr>
<td>Serum Creatinine (mgs/dl)</td>
<td>num.</td>
<td>sc</td>
<td>3.07</td>
</tr>
<tr>
<td>Sodium (mEq/L)</td>
<td>num.</td>
<td>sod</td>
<td>Avg. 137.53</td>
</tr>
<tr>
<td>Potassium (mEq/L)</td>
<td>num.</td>
<td>pot</td>
<td>Avg. 4.63</td>
</tr>
<tr>
<td>Hemoglobin (gms)</td>
<td>num.</td>
<td>hemo</td>
<td>Avg. 12.53</td>
</tr>
<tr>
<td>Packed cell volume</td>
<td>num.</td>
<td>pcv</td>
<td>Avg. 38.88</td>
</tr>
<tr>
<td>White blood cell count</td>
<td>num.</td>
<td>wbc</td>
<td>Avg. 8406.12</td>
</tr>
<tr>
<td>Red blood cell count</td>
<td>num.</td>
<td>rbc</td>
<td>Avg. 4.71</td>
</tr>
<tr>
<td>Specific gravity</td>
<td>nom.</td>
<td>sg</td>
<td>1.005 7</td>
</tr>
<tr>
<td>(1.005, 1.010, 1.015, 1.020, 1.025)</td>
<td>nom.</td>
<td>su</td>
<td>0 1 2 3 4 5 199</td>
</tr>
<tr>
<td>Albumin</td>
<td>nom.</td>
<td>al</td>
<td>44 43 43 43 24 1</td>
</tr>
<tr>
<td>Sugar Degree</td>
<td>nom.</td>
<td>su</td>
<td>0 1 2 3 4 5 290</td>
</tr>
<tr>
<td>Red Blood Cells</td>
<td>nom.</td>
<td>rbc</td>
<td>47 abnormal</td>
</tr>
<tr>
<td>(normal, abnormal)</td>
<td>nom.</td>
<td>pc</td>
<td>76 abnormal</td>
</tr>
<tr>
<td>Pus cell (normal, abnormal)</td>
<td>nom.</td>
<td>pcc</td>
<td>42 present</td>
</tr>
<tr>
<td>Pus cell clumps</td>
<td>nom.</td>
<td>pcc</td>
<td>42 present</td>
</tr>
<tr>
<td>Bacteria (present/ notpresent)</td>
<td>nom.</td>
<td>ba</td>
<td>22 present</td>
</tr>
<tr>
<td>Hypertension (yes/no)</td>
<td>nom.</td>
<td>htn</td>
<td>147 yes</td>
</tr>
<tr>
<td>Diabetes Mellitus (yes/no)</td>
<td>nom.</td>
<td>dm</td>
<td>137 yes</td>
</tr>
<tr>
<td>Coronary artery disease</td>
<td>nom.</td>
<td>cad</td>
<td>34 yes</td>
</tr>
<tr>
<td>(yes/no)</td>
<td>nom.</td>
<td>appet</td>
<td>82 poor</td>
</tr>
<tr>
<td>Appetite (good/poor)</td>
<td>nom.</td>
<td>pe</td>
<td>76 yes</td>
</tr>
<tr>
<td>Pedal edema (yes/no)</td>
<td>nom.</td>
<td>pe</td>
<td>76 yes</td>
</tr>
<tr>
<td>Anemia (yes/no)</td>
<td>nom.</td>
<td>ane</td>
<td>60 yes</td>
</tr>
</tbody>
</table>

Table 2. GFR and CKD relation

<table>
<thead>
<tr>
<th>Stages of CKD</th>
<th>GFR (min/1.73m²)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stage 1</td>
<td>Normal kidney function</td>
</tr>
<tr>
<td>Stage 2</td>
<td>Digestible loss of kidney function</td>
</tr>
<tr>
<td>Stage 3a</td>
<td>Digestible to slight loss of kidney function</td>
</tr>
<tr>
<td>Stage 3b</td>
<td>Slight to severe loss of kidney function</td>
</tr>
<tr>
<td>Stage 4</td>
<td>Severe loss of kidney function</td>
</tr>
<tr>
<td>Stage 5</td>
<td>Kidney failure or dialysis</td>
</tr>
</tbody>
</table>

Stage 1 Normal kidney function 90 +
Stage 2 Digestible loss of kidney function 89 - 60
Stage 3a Digestible to slight loss of kidney function 59 - 45
Stage 3b Slight to severe loss of kidney function 44 - 30
Stage 4 Severe loss of kidney function 29 - 15
Stage 5 Kidney failure or dialysis - 15
dom trees are an ensemble of tree estimators called forest. RT takes the input features and classifies all the trees in the forest. The classifier reply is the average of the replies over all the trees in the forest.

**Reduce Error Pruning Tree (REPTree)**

REPTree makes a regression tree by reducing the data gain as well as pruning the regression tree. It only sorts the data by considering numeric values and copes with deficient values by splitting the data into smaller pieces, as J48 does.

**Random SubSpace**

Random Subspace classifier is based on the ensemble learning algorithm. It constitutes a decision tree-based classifier that maintains the highest accuracy on training data. The classifier involves multi-trees and constructs them systematically with the feature vector and trees structured in chosen subspaces randomly.

**Adaboost Algorithm**

This is an ensemble learning algorithm to determine data classification and degradation. At first, it starts with an equal data distribution, then it finds the best classifier which has a weight below the threshold. Then, the algorithm updates the weights and focuses on erroneously classified samples. Therefore, after a certain number of iterations, the most powerful classifiers are combined to improve the overall performance of the classification.

**IBk**

IBk is a KNN classifier that uses a number of nearest neighbors. Furthermore, it can be specified using leave-one-out cross-validation focalize to an upper limit given by the determined value. Different kinds of search algorithms can be used to accelerate the search of nearest neighbors. For instance, KD-trees, ball trees and cover trees are further options than linear search. In this algorithm, the distance function is used a parameter of the search method.

**Performance Measures**

In this work, we used five performance measures such as accuracy, kappa, MAE, RMSE and F measures.

**Accuracy:** Accuracy is described as the percentage of correctly classified instances. Additionally, accuracy is one of the most popular classification metrics in classification data and given by,

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

where $TN$; true negative, $TP$; true positive, $FN$; false negative and $FP$; false positive.

**Kappa statistic**

This is a classifier performance measure between two sets of classified data. Kappa result values are between 0 to 1. The results become meaningful with increasing values of kappa, which is expressed as,

$$Kappa = \frac{P(A) - P(E)}{1 - P(E)} \quad (4)$$

**Mean absolute error (MAE)**

MAE is a quantity used to measure the predictions which deviate from the true probability. $P(i, j)$ is the estimated probability of $i$ coefficient to be of class $j$. MAE takes values of between 0 to 1, and is given by,

$$\text{MAE} = \frac{1}{m \prod_{c}} \sum_{i=1}^{k} \sum_{j=1}^{m} |f(i, j) - P(i, j)| \quad (5)$$

**Precision**

This is given as the correlation number between the correctly classified modules to entire classified fault-prone modules. Precision is the number of correctly prescribed as faulty:

$$\text{precision} = \frac{TP}{TP + FP} \quad (6)$$

**Recall**

Recall is given as the average probability of complete retrieval, and given by,

$$\text{recall} = \frac{TP}{TP + FN} \quad (7)$$

**Root Mean Square Error (RMSE)**

RMSE is the square root of the mean distance between predicted and observed data. Denoting the estimated probability of the suitable data $i$ and the target value for the suitable data $j$ by $P(i, j)$ and $T_j$, respectively, RMSE is given as follows:

$$E_i = \sqrt{\frac{1}{n} \sum_{j=1}^{n} (P(i, j) - T_j)^2} \quad (8)$$
F-measure

F-measure is the one combination of both precision and recall. F-measure uses the field of information retrieval in order to estimate classification performance, and is calculated by,

$$F - measure = \frac{2 \cdot precision \cdot recall}{precision + recall} \quad (9)$$

where precision and recall are as defined in (6) and (7).

Experimental Results

In this work, we perform seven different classification metrics to show the performance of classification systems. The result of the classification metrics is based on five performance measures, such as accuracy, kappa, MAE, RMSE and F measures, and these are shown in Table 3-6. All data is produced by the WEKA tool during the classification process. For the classification of healthy and sick people, “notckd” and “ckd” labels are used, respectively.

In our previous work, we used all features as well as the reduced features for the evaluation of two performance measures (accuracy and kappa values). Furthermore, we analyzed the performance measure differences of individual and ensemble classifiers. We showed that both ensemble algorithms and proposed feature selection methods are efficient tools to classify CKD. However, our performance measurement metrics changed differently [18, 19].

In this paper, a 70% training test is first completed for 24 features using 10-fold cross validation. Naive Bayes, Hoeffding Tree, RandomTree, REPTree, Random Subspaces, Adaboost and IBk classifiers are used to compare the CKD data. Performance measure metrics of the obtained data are shown in Tables 3-6. Then, we reduce 24 features to 12, 6 and 3 features and analyze the performance measurement metrics.

In Table 3, Random Tree and REPTree classifiers have the same quantity of accuracy, kappa and F measure values. At the same time, these values are the highest scores in all classifier methods. We also show that MAE and RMSE have the smallest values in other classifier methods.
In Table 4, we reduce 24 features to 12 by using Information Gain Attribute Evaluator Feature Selection method. Then, we analyze the data and use the following reduced features: specific gravity, diabetes mellitus, hypertension, red blood cells, albumin, and hemoglobin. As seen from Table 4, Random Tree and REPTree classifiers have the highest quantity of accuracy. Kappa and F measure values as shown in Table 3. In addition, these values are the highest scores in all reduced classifier methods.

In Table 5, we reduce 12 features to 6, which are specific gravity, diabetes mellitus, hypertension, red blood cells, albumin, and hemoglobin. Then, we analyze the performance metrics of the data with reduced features. As seen from Table 5, Naive Bayes, Random Tree and Random SubSpaces classifiers have the same and the highest quantity of accuracy, and kappa and F measure values. On the other hand, IBk classifier has smaller MAE and RMSE values than the other parameters.

Finally, in Table 6, in order to analyze the data, 6 features are reduced to 3, which are specific gravity, diabetes mellitus, and hypertension. Naive Bayes, Random Tree and Random SubSpaces classifiers have the same and the highest quantity of accuracy, and kappa and F measure values. On the other hand, IBk classifier has smaller MAE and RMSE values than the other parameters.

**Conclusion**

In this paper, chronic kidney diseases were classified using various features and classifier combinations. Initially, seven individual classifiers were applied to twenty-four features and the best results were obtained using individual REPTree and Random SubSpaces classifiers as 99.17%. Then, the effect of Information Gain Attribute Evaluator Feature Selection method was observed and analyzed on the CKD data. Consequently, twelve, six and three best features with the best information values were selected and then seven classifiers were applied to these reduced features.

The best accuracy and kappa values were calculated with Random Tree and IBk classifiers as 100% by using the six best features. More-
over, it was shown that Random SubSpace technique has the highest accuracy and kappa values in every reduced type of features.

The early detection stages of CKD can help in the treatment period of the patients, and at the same time, this may help to prevent the disease from getting worse.

Classification techniques considered in this paper can be used and evaluated to find rapid solutions for the patient. The main aim of this study is to reduce the number of classifiers used so that CKD can be diagnosed efficiently and rapidly.

Peer-review: Externally peer-reviewed

Conflict of Interest: The authors declared that this study has received no financial support.

Financial Disclosure: The authors declared that this study has received no financial support.

References


