Diagnosis of Chronic Kidney Disease by Using Random Forest

Abdulhamit Subas, Emina Alickovic, and Jasmin Kevric

College of Engineering, Effat University,
Jeddah, 21478, Saudi Arabia
E-mail:absubasi@effatuniversity.edu.sa
Department of Electrical Engineering, Linkoping University,
Linkoping, 58183, Sweden,
E-mail: emina.alickovic@liu.se
Faculty of Engineering and Information Technologies
International Burch University
Sarajevo, Bosnia and Herzegovina
E-mail: jasmin.kevric@ibu.edu.ba

Abstract. Chronic kidney disease (CKD) is a global public health problem, affecting approximately 10% of the population worldwide. Yet, there is little direct evidence on how CKD can be diagnosed in a systematic and automatic manner. This paper investigates how CKD can be diagnosed by using machine learning (ML) techniques. ML algorithms have been a driving force in detection of abnormalities in different physiological data, and are, with a great success, employed in different classification tasks. In the present study, a number of different ML classifiers are experimentally validated to a real data set, taken from the UCI Machine Learning Repository, and our findings are compared with the findings reported in the recent literature. The results are quantitatively and qualitatively discussed and our findings reveal that the random forest (RF) classifier achieves the near-optimal performances on the identification of CKD subjects. Hence, we show that ML algorithms serve important function in diagnosis of CKD, with satisfactory robustness, and our findings suggest that RF can also be utilized for the diagnosis of similar diseases.

Keywords: Chronic kidney disease (CKD); Machine learning; Artificial Neural Networks (ANNs); Support Vector Machines (SVM); k-Nearest Neighbour (k-NN); C4.5 Decision Tree; Random Forest (RF).

Introduction

In the real world environment, one of the most prominent decision making problems is the classification problem. Classification play a major role in a wide range of machine learning (ML) problems. The main aim of solving the classification problems is to find solution to a problem of prescribing the objects into predefined classes according to the number of detected attributes linked to the corresponding object, and offering several measures for deciding if a prescribed object belongs a specific group or not. Classification tasks are found in a large number of decision making task in different areas such as medicine, science, industry etc. Several approaches on how to solve the classification problems are suggested in the literature.

Probabilistic approach is a traditional approach having a clear core probability modeling, and, as such, this approach is established on the well-known Bayesian decision theory [1]. However, one caveat deserves attention: probabilistic approaches report high performances just in case when the primary assumptions are accurate, what is clearly the main disadvantage of these approaches. Thus, despite all their advantages, Probabilistic approaches have a serious shortcoming. So as to obtain salient performances, users should own an adequate understanding of both information characteristics and model abilities. Therefore, other types of classification approached may work better than probabilistic approaches in different data set. Another approach in solving the classification problems would be to exploit different ML techniques. Wide range of different ML algorithms have been proposed for these purposes. These algorithm, resulting in higher classification performances, are adequate to provide more evidence on how to assign the objects to certain groups (classes), and on how to enhance the model performances. Artificial neural networks (ANN) and its improved version termed as deep neural networks (DNN), support vector machines (SVM), k-Nearest Neighbor (k-NN), decision tree methods have been widely exploited in classification problems, and are still a hot research topic used in ongoing researches. Random forest (RF) have shown state-of-the-art performances in the classification tasks [2, 3, 4]. It is worth noting that many studies on the

A. Badnjevic (ed.), *CMBEBIH 2017*,
IFMBE Proceedings 62,
DOI: 10.1007/978-981-10-4166-2_89

classification problems demonstrated that random forest classifier is favorable replacements to different probabilistic approaches. On the other hand, many practical studies have used the conventional ML tools, ANN, k-NN and SVM, for different classification tasks.

The overall goal of ML is to generate automatic models being able to rapidly generalize (classify) from the examples observed a priori, and it generates by designing or learning functional dependencies among the selected input (features) and output (classes) domains. Diagnosis of Chronic kidney disease (CKD), which is aimed to translate the knowledge from the extracted features (symptoms) into meaningful groups (groups of healthy individuals, CDK individual or individuals with the some other type of disorder), is thus fundamentally a ML problem.

CKD is a chronic healthcare problem, affecting almost 10% of the population worldwide [5, 6]. In real life, CKD appears in many cases to be related to the increased risk of hospital admission, morbidity and death, because of the cardiovascular disease and the progressive loss of kidney function(s). Individual diagnosed with CKD have a high risk of being affected by atherosclerosis and other types of syndromes. These syndromes have substantial effects on their quality of life. The main implication of the CKD diagnosis is the kidney damage [7]. Several symptoms or risk factors are also associated with the CKD progress, so that, these variables could highly influence CKD recognition. By observing the progressive nature of CKD, new insights from the diagnostic computational models grounded on the ML paradigms offer the great promise to advance the diagnosis of CKD [8].

Several studies have suggested the models grounded on the fuzzy logic for diagnosis of CKD [8, 9]. The contribution of this study is to find a simple classifier which gives better classification accuracy. In our study, we consider well-known machine learning methods, artificial neural network (ANN), support vector machine (SVM), k-nearest neighbor (k-NN), C4.5 decision tree and random forest (RF) to produce the highly confident model for the CKD diagnosis. To prove the efficiency and effectiveness of these machine learning tools, we experimentally validated our proposed methods on the real data (CKD dataset) produced by the University of California at Irvine (UCI) machine learning (ML) database [10].

This paper is organized as follows. In Section 2, CKD dataset and classification methods were explained in more details. Section 3 presents the experimental results and discussion. Section 4 provides the conclusions.

Materials and Methods

2.1 Chronic kidney disease dataset

The chronic kidney disease (CKD) dataset used in this study is taken from the UCI Machine Learning Repository [10]. The data, collected during a nearly 2-month period and donated by Soundarapandian et al., includes a total of 400 samples represented by 14 numeric and 10 nominal attributes and a class descriptor. Out of 400 samples, 250 samples belong to the CKD group, and the other 150 samples belong to the non-CKD group. Details are more discussed in [8]

2.2 Artificial Neural Network (ANN)

Over the last few years, artificial neural networks (ANNs) developed into regular and salient ML paradigm. ANNs are the prominent type of quantitative modeling techniques and they received great reputation between researchers over the last two decades and have been effectively used for solving many different difficulties in almost all areas of business, industry, and science [11, 12]. Nowadays, ANNs are considered as a regular machine learning tool and employed for numerous data mining tasks like pattern classification, clustering, prediction and time series analysis. Actually, ANNs are core module for almost all profitable data mining software packages [12].

They are perfectly proper for modeling relations among a set of input variables and one or more output variables. In our study, we had 24 input variables (24 attributes) for each pattern and two possible output variables (CKD or No-CKD). MLPs are suitable for any functional mapping tasks where we are trying to find out output variable(s) are affected by a number of input variable(s). As almost all prediction and classification problems can be considered as function mapping tasks, the MLPs are very attractive to data mining [12]. Due to this reason, MLP networks were employed in this study and their performance results were investigated for CKD diagnosis.

MLP network consists of numerous greatly interconnected simple computing units named neurons or nodes, and they are organized in different layers. Every neuron does simple information processing task by translating received input values into processed outputs values. Knowledge can be obtained and stored as arc weights concerning the strength of the relationship among various nodes through the linking arcs between neurons. Even though every neuron performs its function gradually and improperly, jointly a neural net-