Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 65 (2015) 459 – 468

Procedia Computer Science

International Conference on Communication, Management and Information Technology (ICCMIT 2015)

Feature Analysis of Coronary Artery Heart Disease Data Sets

Randa El-Bialy[1] , Mostafa A. Salamay[2], Omar H. Karam[3] and M.Essam Khalifa[4]

[1]*British University in Egypt (BUE), Cairo, Egypt, Randa.Elbialy@Bue.edu.eg*
[2]*British University in Egypt (BUE), Cairo, Egypt, mostafa.salama@bue.edu.eg*
[3]*British University in Egypt (BUE), Cairo, Egypt, omar.karam@bue.edu.eg*
[4]*Ain Shams University, Cairo, Egypt, esskhalifa49@gmail.com*

Abstract

Data sets dealing with the same medical problems like Coronary artery disease (CAD) may show different results when applying the same machine learning technique. The classification accuracy results and the selected important features are based mainly on the efficiency of the medical diagnosis and analysis. The aim of this work is to apply an integration of the results of the machine learning analysis applied on different data sets targeting the CAD disease. This will avoid the missing, incorrect, and inconsistent data problems that may appear in the data collection. Fast decision tree and pruned C4.5 tree are applied where the resulted trees are extracted from different data sets and compared. Common features among these data sets are extracted and used in the later analysis for the same disease in any data set. The results show that the classification accuracy of the collected dataset is 78.06% higher than the average of the classification accuracy of all separate datasets which is 75.48%.

Peer-review under responsibility of Universal Society for Applied Research

Keywords: Data Mining, Fast Decision Tree Learning Algorithm, Decision Trees.

1. Introduction

Data mining is the process of finding previously unknown patterns and trends in databases and using that information to build predictive models. In healthcare, Data mining is a field of high importance and has become increasingly effective and essential. The healthcare industry today generates large amounts of complex data concerning patients, hospitals resources, disease diagnosis, electronic patient records, medical devices, etc. The large amount of data is a key resource to be processed and analyzed for knowledge extraction and enabling support for cost-savings and decision-making. Data mining provides a set of tools and techniques that can be applied to the data to achieve these goals. Coronary heart disease is considered a fatal illness that causes death to over a million patients every year. Nearly half the patients diagnosed with CAD will eventually die from the disease. Devastatingly, 335,000 of CHD patients will die of a heart attack in an emergency department or before they even reach the hospital. According to the American Heart Association, over 7 million Americans have suffered a heart attack in their lifetime. When plaque is built inside the coronary arteries, they narrow these arteries making them unable to carry oxygenated blood to the heart muscle causing the well-known symptoms of CAD such as chest pain (angina) and shortness of

Peer-review under responsibility of Universal Society for Applied Research
doi:10.1016/j.procs.2015.09.132

breath [1, 2]. A system to identify the probability of the existence of coronary heart disease was presented in [3] by Patel et al. In that work, the parameters were divided into two levels, each with its weight. In order to derive a decision, a neuro-fuzzy integrated approach was implemented taking into consideration both parameter levels to reach a final decision. The choice of a fuzzy approach is advantageous in that it reduced error rate and in turn enhanced efficiency. In [4] Chitra R. et al. developed a computer aided heart disease prediction system. The system is intended to help physicians diagnose heart disease. Their conclusion was that, data mining can play a major role in heart disease classification. It is further noted that due to the nature of the issue, feature reduction techniques may be needed in order to enhance the efficiency and search time of the classifier in question. Nidhi Bhatlaet al. [5] discussed the results of applying various data mining techniques that have been closely associated with heart disease diagnosis in recent years. The techniques that were surveyed by that work include Neural Networks, Decision Trees and Genetic Algorithms. The work of Srinivas K. et al. [6] explored various data mining aspects that went beyond the scope of classification; namely clustering, association rule mining and time series analysis. The work focused on the prediction of various combinations of certain heart disease attributes. This work has a high potential for further expansion and enhancement. In recent decades, many experts have tried to make CAD diagnosis using computer-aided techniques, such as neural networks [7], the Bayesian model and decision tree 6, support vector machine [8], and the naive Bayes classifier [9]. The aim of this work is to apply an integration of the results of the machine learning analysis applied on different data sets targeting the CAD disease. This will avoid the missing, incorrect, and inconsistent data problems that may appear in the data collection, finding a set of attributes which are really important for this disease predication in order to improve the performance of the classifier, help physicians to successfully diagnose CAD. This current paper is organized as follows. Section 2 contains previous work related to the topic. In Section 3 the Decision Tree Integration Model / Methodology is presented. Results and Discussion are in Section 4. The conclusion is provided in Section 5.

2. Related Work

A Decision Tree is a decision support system that uses a tree-like graph where each node denotes a test on an attribute value and each branch represents an outcome of the test while tree leaves stand for classes or class distributions. Decision trees have many advantages such as i) construction of decision trees is unique for every dataset (not complicated), ii) end users can understand them easily, iii) a variety of input data can be handled e.g. nominal, numeric and textual, iv) the ability to process missing values or invalid/erroneous datasets and v) high performance is achieved with a little effort. The classification is performed by starting at the root node for each new record to be classified, and depending on the results at each consecutive node, a leaf node is reached thus determining the class for that record or determining a probability distribution for the possible classes [10, 11]. There are many algorithms for building decision trees such as ID3 and C4.5. There is also the Fast Decision Tree. Quinlan Ross introduced the ID3 (Iterative Dichotomiser) in 1986 [12]. It determine the splitting attribute according to the information gain measure; for each and every attribute in the data set the information gain the attribute with the highest information gain is identified and assigned as a root node for the tree. Declaring the selected attribute as a root node and representing the possible attribute values as arcs, all possible outcome instances are tested to check whether they fall under the same class or not. The node is represented with a single class name if the instances are falling under the same class; otherwise the splitting attribute to classify the instances is chosen. The main disadvantages facing ID3 are: accepting only categorical attributes, giving inaccurate results when noise exists, testing only one attribute at a time for making decisions and pruning is not supported [13,14].

2.1. C4.5 Algorithm

C4.5 is an extension and improved version of the ID3 algorithm developed by Quinlan Ross 12. Pruning is the key feature/step to overcome the over fitting problem in the ID3 algorithm. Both categorical and continuous attributes in the data set are handled. A Gain Ratio is used as the splitting criterion, and then according to the selected threshold the attribute values are split into two partitions: all the attributes with