Exerting Cost-Sensitive and Feature Creation Algorithms for Coronary Artery Disease Diagnosis

Roohallah Alizadehsani, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
Mohammad Javad Hosseini, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
Reihane Boghrati, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
Asma Ghandeharioun, Department of Computer Engineering, Sharif University of Technology, Tehran, Iran
Fahime Khozeimeh, Mashhad University of Medical Science, Mashhad, Iran
Zahra Alizadeh Sani, Tehran University of Medical Science, Tehran, Iran

ABSTRACT

One of the main causes of death the world over is the family of cardiovascular diseases, of which coronary artery disease (CAD) is a major type. Angiography is the principal diagnostic modality for the stenosis of heart arteries; however, it leads to high complications and costs. The present study conducted data-mining algorithms on the Z-Alizadeh Sani dataset, so as to investigate rule-based and feature-based classifiers and their comparison, and the reason for the effectiveness of a preprocessing algorithm on a dataset. Misclassification of diseased patients has more side effects than that of healthy ones. To this end, this paper employs 10-fold cross-validation on cost-sensitive algorithms along with base classifiers of Naïve Bayes, Sequential Minimal Optimization (SMO), K-Nearest Neighbors (KNN), Support Vector Machine (SVM), and C4.5 and the results show that the SMO algorithm yielded very high sensitivity (97.22%) and accuracy (92.09%) rates.

Keywords: C4.5 Algorithm, Coronary Artery Disease, Cost-Sensitive Algorithms, Data Mining, Feature Creation, Naïve Bayes Algorithm

DOI: 10.4018/jkdb.2012010104
INTRODUCTION

The morality rates from diseases are much greater than those of accidents and natural disasters. The World Health Organization estimates that 17 million deaths worldwide each year occur due to cardiovascular diseases (Bonow, Mann, Zipes, & Libby, 2012). A major type of such diseases is coronary artery disease (CAD), which is reported to account for 7 million deaths over the world per annum (Bonow et al., 2012).

Mining is the extraction of knowledge from a set of data. In other words, data mining is a process that uses intelligent techniques whereby knowledge of a set of data can be extracted (Bickel & Scheffer, 2004).

Angiography is the modality of choice for the diagnosis of CAD. Angiography determines the location and extent of the stenotic arteries; nevertheless, its high costs and risks for the patient have prompted researchers to seek less expensive and more effective methods with the aid of data mining. Moreover, cost-sensitive algorithms can be of huge value in this field as misclassification of diseased or healthy patients has different costs. Pedreira et al. (2005), using the Neural Network on UC1 (UC Irvine Machine Learning Repository, 2012) datasets, attained an accuracy rate of 80% for CAD diagnosis. Das et al. (2009) applied the Neural Network on the datasets of Cleveland (UC Irvine Machine Learning Repository, 2012) and reported an accuracy rate of 89.01%. Babaoglu et al. (2010) utilized the Support Vector Machine (SVM) algorithm on an exercise test data and achieved an accuracy rate of 79.17%. Tsiourras et al. (2008) used the Fuzzy Model to detect CAD. Itchhaporia et al. (1995) drew upon the Neural Network to analyze an exercise test data for the diagnosis of CAD. Polat et al. (2007) by using fuzzy systems and KNN reached the accuracy of 87% for CAD diagnosis. Alizadehsani et al. (2012) proposed a new ensemble algorithm which diagnoses CAD by 88.5% accuracy. Lee et al. (2008) used Heart Rate Variability (HRV) features for diagnosing CAD. Karaolis et al. (2010) and Srinivas et al. (2010) used C4.5 and naive bayes algorithm respectively to diagnose CAD.

One of the purposes of the present study was to investigate rule based classifiers for CAD diagnosis. Resulted in low specificity rule based classifiers, other methods were sought in this paper. We use MetaCost, which is a cost-sensitive (Domingos, 1999) algorithm, so as to distinguish CAD patients from healthy individuals. The Sequential Minimal Optimization (SMO) (Platt, 1998), Naïve Bayes (Caruana, & Niculescu-Mizil, 2006), C4.5 (Quinlan, 1996), Support Vector Machine (SVM) (Ben-Hur & Weston, 2010), and K-Nearest Neighbors (KNN) (Larose, 2005) algorithms were employed to analyze the Z-Alizadeh Sani dataset with no feature normalization. The performance of all the mentioned algorithms was calculated using 10-fold cross-validation. This dataset contains information on 303 random visitors to Rajaei Cardiovascular, Medical and Research Center in Tehran, Iran. The dataset was enriched with three created features extracted from the other features prior to the application of the cost-sensitive algorithms on the datasets. The effect of the created features was investigated both theoretically and practically. First, an assumption was made about the created features. Then a lemma was stated to provide a subset of sample which satisfied the assumption. Afterwards, another lemma was presented using assumption 1 in order to discuss the effectiveness of the created features. In the experiments, the correctness of assumption 1 and the effectiveness of the created features were studied. As a result, high rates of both accuracy and sensitivity were obtained which, to the best of our knowledge, are superior to the existing studies in this area.
Related Content

Time-Aware Task Allocation for Cloud Computing Environment
[www.igi-global.com/article/time-aware-task-allocation-for-cloud-computing-environment/178603?camid=4v1a](www.igi-global.com/article/time-aware-task-allocation-for-cloud-computing-environment/178603?camid=4v1a)

Augmenting Medical Decision Making With Text-Based Search of Teaching File Repositories and Medical Ontologies: Text-Based Search of Radiology Teaching Files
Genetic Diagnosis of Cancer by Evolutionary Fuzzy-Rough based Neural-Network Ensemble

Information Retrieval and Access in Cloud
Punit Gupta and Ravi Shankar Jha (2017). *Library and Information Services for Bioinformatics Education and Research* (pp. 212-228).
[www.igi-global.com/chapter/information-retrieval-and-access-in-cloud/176146?camid=4v1a](www.igi-global.com/chapter/information-retrieval-and-access-in-cloud/176146?camid=4v1a)