Available online at www.sciencedirect.com
ScienceDirect
Expert Systems with Applications 34 (2008) 366–374
Expert Systems with Applications
www.elsevier.com/locate/eswa

Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease

Imran Kurt [a,*], Mevlut Ture [b], A. Turhan Kurum [c]

[a] Eskisehir Osmangazi University, Medical Faculty, Department of Biostatistics, 26480 Eskisehir, Turkey
[b] Trakya University, Medical Faculty, Department of Biostatistics, 22030 Edirne, Turkey
[c] Trakya University, Medical Faculty, Department of Cardiology, 22030 Edirne, Turkey

Abstract

In this study, performances of classification techniques were compared in order to predict the presence of coronary artery disease (CAD). A retrospective analysis was performed in 1245 subjects (865 presence of CAD and 380 absence of CAD). We compared performances of logistic regression (LR), classification and regression tree (CART), multi-layer perceptron (MLP), radial basis function (RBF), and self-organizing feature maps (SOFM). Predictor variables were age, sex, family history of CAD, smoking status, diabetes mellitus, systemic hypertension, hypercholesterolemia, and body mass index (BMI). Performances of classification techniques were compared using ROC curve, Hierarchical Cluster Analysis (HCA), and Multidimensional Scaling (MDS). Areas under the ROC curves are 0.783, 0.753, 0.745, 0.721, and 0.675, respectively for MLP, LR, CART, RBF, and SOFM. MLP was found the best technique to predict presence of CAD in this data set, given its good classificatory performance. MLP, CART, LR, and RBF performed better than SOFM in predicting CAD in according to HCA and MDS.

Keywords: Logistic regression; Decision tree; Neural networks; Coronary artery disease; Multidimensional scaling; Hierarchical cluster analysis; ROC curve

1. Introduction

Coronary artery disease (CAD) is a major worldwide health problem with its incidence and mortality rates (Backer et al., 2003). Many risk factors are known to play role in pathogenesis of CAD and myocardial infarction. Family history, smoking, hypertension, hypercholesterolemia, diabetes mellitus and obesity have been described as the major risk factors for CAD (Burke et al., 1997; Celermajer et al., 1993; Chobanian et al., 2003; Expert Panel on Detection, Evaluation, & Treatment of High Blood Cholesterol in Adults, 2001; Fuster, 1994; Haskell et al., 1994; Herrington et al., 2003; Higgins et al., 1992; Rahman, Chaudhury, Malek, & Khaled, 2004; The Expert Committee on the Diagnosis & Classification of Diabetes Mellitus, 2003; Zeiher, Drexler, Saurbier, & Just, 1993). Identification of risk factors in CAD is essential for the management and follow-up of CAD. Numerous cross-sectional angiography studies have reported correlations of one or more major risk factors for myocardial infarction with the presence of CAD.

Predicting the outcome of a disease is one of the most interesting and challenging tasks in which to develop data mining applications. Some applications of neural networks in cardiology were performed for the analysis of heart sounds, analysis of cardiac arrhythmias, the detection of ventricular ectopic activity, and the detection of atrial fibrillation. Furthermore, the development of implantable

* Corresponding author. Tel.: +90 222 3227250; fax: +90 222 2393772.
E-mail addresses: ikurt@trakya.edu.tr (I. Kurt), ture@trakya.edu.tr (M. Ture), atkurum@trakya.edu.tr (A.T. Kurum).

0957-4174/$ - see front matter © 2006 Elsevier Ltd. All rights reserved.
doi:10.1016/j.eswa.2006.09.004

devices for treatment of life-threatening arrhythmias has simulated intracardiac rhythm classification using neural networks. Neural networks have been trained to recognize ST-T segment changes, to recognize CAD in general, to predict the number of vessels involved and to identify three-vessel and mainstem disease, even at rest (Dassen, Egmont-Petersen, & Mulleneers, 1998). Few works have been published on the comparison of classification techniques in different areas. Moisen and Frescino (2002) compared linear models, generalized additive models, classification and regression tree (CART), Multivariate Additive Regression Splines (MARS), and artificial neural networks for mapping forest characteristics in the Interior Western United States using forest inventory field data and ancillary satellite-based information. Ture, Kurt, Kurum, and Ozdamar (2005) compared various classification techniques to predict control and hypertension groups. They created models using logistic regression (LR), flexible discriminant analysis (FDA), FDA with MARS (FDA/MARS), chi-squared automatic interaction detector (CHAID), quick unbiased efficient statistical tree (QUEST), CART, radial basis function (RBF) and multilayer perceptron (MLP) to predict control and hypertension groups. Delen, Walker, and Kadam (2004) compared LR, decision tree (C5) and artificial neural networks for predicting the survivability of diagnosed cases for breast cancer. Stark and Pfeiffer (1999) compared LR, classification tree algorithms (ID3, C4.5, CHAID, CART) and artificial neural networks to solve classification problems in complex data sets in veterinary epidemiology. Colombet et al. (2000) evaluated the implementation and performance of CART and artificial neural networks comparatively with a LR model, in order to predict the risk of cardiovascular disease in a real database. King, Feng, and Sutherland (1995) compared symbolic learning (CART, C4.5, NewID, AC^2, ITrule, Cal5, and CN2), statistics (Naïve Bayes, *k*-nearest neighbor, kernel density, linear discriminant, quadratic discriminant, LR, projection pursuit, and Bayesian networks), and neural networks (back-propagation and RBF) algorithms on twelve datasets with respect to large real-world problems.

The purpose of this study is to compare performances of classification techniques in order to predict the presence of CAD. We have created models using LR, CART, neural networks algorithms (RBF, MLP, and self-organizing feature maps (SOFM)) that they are often used for classification problems. LR is useful for situations in which you want to be able to predict the presence or absence of a characteristic or outcome based on values of set of independent variables which are continuous, categorical, or both. Furthermore, it assumes that measures of dependent variables are independently and randomly sampled, all potentially relevant independent variables are in the model and all independent variables in the model are relevant (Hosmer & Lemeshow, 2000; Kleinbaum, 1994). CART is inherently non-parametric that no assumptions are made regarding the underlying distribution of values of the predictor variables. Thus, CART can handle numerical data that are highly skewed or multi-modal, as well as categorical predictors with either ordinal or non-ordinal structure (Breiman, Friedman, Olshen, & Stone, 1984). Neural networks have been used to model medical and functional outcomes of dangerous disease. They have become a popular tool for classification, as they are very flexible, not assuming any parametric form for distinguishing between categories (Lee, 2001).

Performances of classification techniques were compared using ROC curve, Hierarchical Cluster Analysis, and Multidimensional Scaling.

2. Material and methods

2.1. Data

A retrospective analysis was performed in 1245 subjects (865 presence of CAD and 380 absence of CAD). Clinically relevant CAD was defined by the presence of at least one vessel with a stenosis $\geqslant 50\%$ in coronary angiography to be done due to angina associated with evidence for myocardial ischemia either by stress electrocardiography, stress Tc99 MIBI scintigraphy or pathologic resting electrocardiography. All angiograms were assessed by two cardiologists not participating in the study. All patients with suspected CAD were seen in the Cardiology Clinic of Trakya University Medical Faculty in Turkey between January 2002 and February 2003. The patients had coronary angiography because of stable angina pectoris or an acute coronary syndrome or atypical angina was included in the study. Patients with non-atherosclerotic CAD were excluded from the study.

Independent variables included age, sex, family history of CAD, smoking status, diabetes mellitus, systemic hypertension, hypercholesterolemia, and BMI. Hypertension was diagnosed when the systolic blood pressure (BP) was $\geqslant 140$ mm Hg and/or diastolic BP was $\geqslant 90$ mm Hg on at least three separate occasions and was established by the absence of clinical findings suggestive of secondary form of hypertension (Chobanian et al., 2003). BP was measured in the sitting position in a quite room, using a mercury sphygmomanometer, after the patient had rested for at least 10 min. Systolic BP was recorded at the appearance of sounds (korotkoff phase I) and the diastolic at their disappearance (korotkoff phase V). Patients who currently smoked or discontinued during the last 6 months were categorized as smokers. Diabetes was considered confirmed if the questionnaire indicated one of the following National Diabetes Data Group criteria: (1) Symptoms of diabetes plus casual plasma glucose concentration $\geqslant 200$ mg/dl (11.1 mmol/l). Casual is defined as any time of day without regard to time since last meal. The classic symptoms of diabetes include polyuria, polydipsia, and unexplained weight loss. (2) FPG $\geqslant$ 126 mg/dl (7.0 mmol/l). Fasting is defined as no caloric intake for at least 8 h. (3) 2-h PG $\geqslant$ 200 mg/dl (11.1 mmol/l) during an OGTT. The test should be