Available online at www.sciencedirect.com

ScienceDirect

Procedia Computer Science 121 (2017) 160–165

Procedia Computer Science

www.elsevier.com/locate/procedia

CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies, CENTERIS / ProjMAN / HCist 2017, 8-10 November 2017, Barcelona, Spain

Empirical thresholding logistic regression model based on unbalanced cardiac patient data

Iris Reychav[a*], Lin Zhu[a,b], Roger McHaney[c], Yaron Arbel[d]

[a] *Ariel University, Ari'el, Israel*
[b] *Southwest Petroleum University, Chengdu, China*
[c] Kansas State University, Manhattan, USA
[d] Sourasky Medical Center, Tel-Aviv, Israel

Abstract

Cardiac disease causes widespread morbidity and mortality. Past research in this area focused on risk factors and treatment. Little exists on patient survival classification in emergency room situations with unbalanced data. The current study expanded knowledge in this area based on over 2,000 cardiac patient records. This unbalanced dataset was used to develop an empirical, thresholding logistic regression model which predicted patients' survival. The model was refined using stepwise and cost-efficient methods. The exploration revealed important factors that influenced patient survival and suggested a thresholding logistic regression model can provide a flexible and pragmatic way to handle unbalanced cardiac patient data. The model identified key factors to help doctors concentrate on important indicators related to patient survival. This study offers novel technical and practical insights for instant survival analysis of cardiac patients, using an unbalanced dataset.

Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.

Keywords: Survival Prediction; Unbalanced Cardiac Patients Data; Classification; Empirical Thresholding Logistic Regression

[*] Corresponding author. Tel.: 972-3-9066325 ; fax: 972-3-9066322.
E-mail address: irisre@ariel.ac.il

1877-0509 © 2017 The Authors. Published by Elsevier B.V.
Peer-review under responsibility of the scientific committee of the CENTERIS - International Conference on ENTERprise Information Systems / ProjMAN - International Conference on Project MANagement / HCist - International Conference on Health and Social Care Information Systems and Technologies.
10.1016/j.procs.2017.11.022

1. Introduction

Cardiac disease is a major cause of morbidity and mortality worldwide [1]. Hence, many studies seek to enhance prevention and suggest appropriate treatments [2]. Unfortunately, not many classification studies of cardiac disease patients currently exist, particularly those based on unbalanced datasets [3]. Unbalanced data can cause high Type II errors of classification models [4,5]. This becomes important because data regarding cardiac diseases commonly are unbalanced and studies rarely take this into consideration [6,7]. Therefore, enhancing classification research for unbalanced cardiac patient data is necessary.

Currently, two streams of research describe ways to approach unbalanced data in machine learning: a rebalance method, comprised of sampling-based and weight-based approaches [8]; and thresholding approaches [9] which use a cost-based principle to select the optimal threshold for prediction. In general, sampling-based methods are more popular due to flexibility and ease of usage and interpretation. These methods include oversampling [10] and hybrid approaches [11]. Cost-based thresholding methods are used relatively less, but are powerful tools for handling unbalanced data [12,13]. In this study, the empirical thresholding method was selected to improve classification performance. To implement it, logistic regression (LR) was chosen as the machine learning method due to its relative simplicity, stable theoretical basis, and ease of interpretation [14]. Therefore, an empirical thresholding logistic regression model (ETLRM), which incorporates LR was used to handle the unbalanced cardiac patient data in this study.

2. Method

2.1. Research structure

Our research design comprised three parts. The first was data selection and cleaning. The second was data classification which used the tuneThreshold function of the MLR package in R[15] to find the optimal threshold. The third part was model refinement where stepwise regression and the cost-efficient method selected the most significant variables in ETLRM. The step function in R was used, with a backwards direction [16].

2.2. Data collection and cleaning

Data for the study came from Tel-Aviv Sourasky Medical Center, a well-respected medical institute in Israel. It launched a long-term survey of cardiac patients to enhance understanding cardiac diseases. Cardiac patient data were gathered from December 28, 2007 to April 30, 2016. In total, 2,099 records were collected. This data related to a specific scenario: a cardiac patient is urgently sent to a hospital. Doctors provide emergency treatments and make an instant estimation for the patient as a lifesaving measure. This decision only can use basic medical indexes and information readily available to the doctor during the emergency. Therefore, this study only uses data available in these circumstances. See Table 1.

Table 1. Basic variable descriptions.

Variables	Description	Data Type[a]	Missing Values	Basic summary[b]
DM	Diabetes Mellitus (risk factors for heart disease and diabetes)	N2	0	0(1318),1(399)
Insulin use	Insulin use	N2	5	0(1629),1(88)
Glucose levels	Glucose levels	N	8	156.938(71.05)
Renal function (mlmin)	Renal function (ml/min or milliliters per minute)	N	4	73.624(22.432)
Baseline Creatinine	Baseline Creatinine (measuring blood test)	N	4	1.133(0.326)
All cause [c]	All-cause mortality	N2	361	0(1587),1(130)
Shock	Presence of shock	N2	1	0(1633),1(84)
Mech Vent	Mechanical ventilation	N2	1	0(1628),1(89)