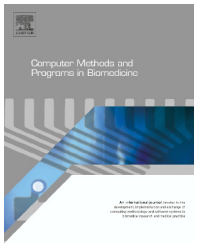
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 130 (2016) 87–92

journal homepage: www.intl.elsevierhealth.com/journals/cmpb



Different medical data mining approaches based prediction of ischemic stroke

Ahmet Kadir Arslan[a,*], *Cemil Colak*[a], *Mehmet Ediz Sarihan*[b]

[a] *Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey*
[b] *Inonu University, Faculty of Medicine, Department of Emergency Medicine, Malatya, Turkey*

ARTICLE INFO

Article history:
Received 10 February 2016
Received in revised form
8 March 2016
Accepted 18 March 2016

Keywords:
Ischemic stroke
Medical data mining
Penalized logistic regression
Stochastic gradient boosting
Support vector machine

ABSTRACT

Aim: Medical data mining (also called knowledge discovery process in medicine) processes for extracting patterns from large datasets. In the current study, we intend to assess different medical data mining approaches to predict ischemic stroke.

Materials and methods: The collected dataset from Turgut Ozal Medical Centre, Inonu University, Malatya, Turkey, comprised the medical records of 80 patients and 112 healthy individuals with 17 predictors and a target variable. As data mining approaches, support vector machine (SVM), stochastic gradient boosting (SGB) and penalized logistic regression (PLR) were employed. 10-fold cross validation resampling method was utilized, and model performance evaluation metrics were accuracy, area under ROC curve (AUC), sensitivity, specificity, positive predictive value and negative predictive value. The grid search method was used for optimizing tuning parameters of the models.

Results: The accuracy values with 95% CI were 0.9789 (0.9470–0.9942) for SVM, 0.9737 (0.9397–0.9914) for SGB and 0.8947 (0.8421–0.9345) for PLR. The AUC values with 95% CI were 0.9783 (0.9569–0.9997) for SVM, 0.9757 (0.9543–0.9970) for SGB and 0.8953 (0.8510–0.9396) for PLR.

Conclusions: The results of the current study demonstrated that the SVM produced the best predictive performance compared to the other models according to the majority of evaluation metrics. SVM and SGB models explained in the current study could yield remarkable predictive performance in the classification of ischemic stroke.

1. Introduction

Ischemic stroke (IS) is associated with high mortality worldwide and is considered among the most important public health problems [1]. IS influences the management, diagnosis, and outcome. Treatments for acute IS should be made according to subtype of IS. Classification of subtypes for IS was arranged utilizing medical/clinical characteristics and the finding of supplementary clinical studies. The classification of Trial of Org in Acute Stroke Treatment (TOAST) defines five subtypes of IS: (1) big-artery atherosclerosis, (2) cardioembolism, (3) little-vein occlusion, (4) stroke of other identified etiology/causes, and (5) stroke of unidentified etiology/causes. The proposed rating system can determine etiologic diagnosis of IS in high proportions [2]. The important inference demonstrates the determination and prediction of causes and markers for the diagnosis and prevention of IS [1,2].

* *Corresponding author*. Tel.: +90 546 932 8246; fax: +90 422 3410036.
E-mail address: ahmetkadirarslan@gmail.com (A.K. Arslan).
http://dx.doi.org/10.1016/j.cmpb.2016.03.022

Table 1 – The definition of the variables employed in the current study.

Variables	Abbreviation	Variable type	Definition	Role
Ischemic stroke	Is	Categorical	Present/absent	Target
Age (year)	–	Numerical	Natural number	Input
Gender	–	Categorical	Female/male	Input
Educational status	Es	Categorical	Elementary school/middle school/high school/university	Input
Marital status	Ms	Categorical	Single/married/widowed	Input
Alcohol consumption	Ac	Categorical	Present/absent	Input
White blood cell	Wbc	Numerical	Positive real number	Input
Hematocrit	Htc	Numerical	Positive real number	Input
Hemoglobin	Hb	Numerical	Positive real number	Input
Platelet	Plt	Numerical	Positive integer	Input
Glucose	Glc	Numerical	Positive integer	Input
Blood urea nitrogen	Bun	Numerical	Positive integer	Input
Creatinine	Cr	Numerical	Positive real number	Input
Sodium	Na	Numerical	Positive integer	Input
Potassium	K	Numerical	Positive real number	Input
Chlorine	Cl	Numerical	Positive integer	Input
Prothrombin time	Inr	Numerical	Positive real number	Input
Calcium	Ca	Numerical	Positive real number	Input

Data mining (also called knowledge discovery process) is a methodology for discovering hidden patterns from enormous datasets by using statistical approaches [3]. This methodology has many advantages compared to classical methods. For instance, in contrast to traditional statistical methods, data mining approaches require less presumptions in the classification and regression applications [4].

Alexopoulos et al. [5] applied inductive machine learning (ML) approaches in the diagnosis of stroke disease and used C4.5 algorithm by building a decision tree. These authors reported that inductive ML is a promising approach for computer-aided diagnosis of stroke. Linder et al. [6] used logistic regression (LR) and artificial neural networks (ANNs) for classifying acute ischemic stroke from the Database of German Stroke, and suggested that LR was the gold standard for the classification of acute ischemic stroke in comparison with ANNs, which may be employed as an alternative multivariate analysis. Khosla et al. [7] presented the comparison of the Cox proportional hazards model with a ML method for the prediction of stroke on the dataset of the Cardiovascular Health Study, and determined that combined with their suggested feature selection algorithm combined with support vector machine (SVM) achieved a higher area under the ROC curve when compared to the Cox proportional hazards model. In our previous study, ANNs and SVM were utilized to predict stroke disease using knowledge discovery process (KDP) approaches, and the results of the study determined that ANNs yielded more predictive performance as compared with SVM for the prediction of stroke and that the suggested ANNs might be beneficial for predictive purposes concerning stroke illness [3]. Additionally, there are some studies on ischemic stroke lesion segmentation using data mining or ML procedures [8–10].

The use of data mining approaches in many disciplines, especially in medicine, is increasing day by day. The medical application of data mining is called as medical data mining (MDM). Thence, MDM (also called knowledge discovery process in medicine) processes for extracting patterns from large datasets. In the current study, we intend to assess medical data mining approaches to predict ischemic stroke.

2. Material and methods

2.1. *Dataset*

This study which included 80 ischemic stroke patients (group I) and 112 healthy individuals (group II) was conducted in the department of emergency medicine, Turgut Ozal Medicine Center, Inonu University, Malatya, Turkey. Power analysis revealed that each group encapsulated minimum 68 individuals considering mean difference of creatinine for ischemic stroke patients and healthy individuals groups of 0.6, estimated standard deviations of 1.01 and 1.43, type I error (alpha) of 0.05 and type II error (beta) of 0.20. The definition of the variables that may associate with ischemic stroke [3,11,12] is summarized in Table 1.

2.2. *Preprocessing of the dataset*

In the current study, outliers were detected by local density cluster-based outlier factor [13]. This technique employs a cluster algorithm and allocates clusters into small and big ones. The outlier factor was calculated by dividing minimum sample distance to average cluster distance of all samples to the big cluster [14]. X-means was utilized as clustering algorithm in this technique. Also, z-transformation (standardization) was applied to the dataset.

2.3. *Support vector machines*

SVM is a supervised learning approach for classification and regression tasks and is utilized in order for linear/nonlinear classification problems with high-dimensional datasets [15]. To solve nonlinear classification problem, SVM maps the input sets to a high-dimensional space by applying various kernel functions [3]. A detailed explanation of SVM approach can be achieved in [16]. In this paper, SVM was employed with radial basis function (RBF) kernel function. SVM with RBF was applied by kernlab package [17] in R.