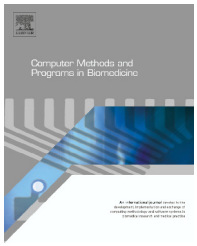
COMPUTER METHODS AND PROGRAMS IN BIOMEDICINE 119 (2015) 181–185

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



Application of knowledge discovery process on the prediction of stroke

Cemil Colak[a,*], *Esra Karaman*[b], *M. Gokhan Turtay*[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 18 December 2014
Received in revised form
17 February 2015
Accepted 4 March 2015

Keywords:
Artificial neural networks (ANN)
Knowledge discovery process (KDP)
Support vector machine (SVM)
Stroke

ABSTRACT

Objective: Stroke is a prominent life-threatening disease in the world. The current study was performed to predict the outcome of stroke using knowledge discovery process (KDP) methods, artificial neural networks (ANN) and support vector machine (SVM) models.
Materials and methods: The records of 297 (130 sick and 167 healthy) individuals were acquired from the databases of the department of emergency medicine. Nine predictors (coronary artery disease, diabetes mellitus, hypertension, history of cerebrovascular disease, atrial fibrillation, smoking, the findings of carotid Doppler ultrasonography [normal, plaque, plaque + stenosis ≥ 50%], the levels of cholesterol and C-reactive protein) were used for predicting the stroke. Feature selection based on the Cramer's V test was carried out for reducing the predictors. Multilayer perceptron (MLP) ANN and SVM with radial basis function (RBF) kernel were used for the prediction based on the selected predictors.
Results: The accuracy values were 81.82% for ANN and 80.38% for SVM in the training dataset ($n = 209$), and 85.9% for ANN and 84.62% for SVM in the testing dataset ($n = 78$), respectively. ANN and SVM models yielded area under curve (AUC) values of 0.905 and 0.899 in the training dataset, and 0.928 and 0.91 in the testing dataset, consecutively.
Conclusion: The findings of the current study pointed out that ANN had more predictive performance when compared with SVM in predicting stroke. The proposed ANN model would be useful when making clinical decisions regarding stroke.

© 2015 Elsevier Ireland Ltd. All rights reserved.

1. Introduction

Stroke is a significant cause for cognitive disorders around the world. In developing countries, there is a shortage of data in regard to the public health extent of stroke [1]. Stroke is a growing illness and is the third prevalent cause for death after coronary heart disease and cancer in the world, particularly in older people. Stroke regularly caused increased morbidity, mortality and decreased life quality in the community [2].

Knowledge discovery in databases (KDD), also called as knowledge discovery process (KDP) or data mining, is an approach of obtaining patterns from huge datasets by combining techniques of statistics and machine learning [3]. KDP involves distinctive ideas from machine learning, artificial intelligence, statistics, database query, and visualization. While the target of database advances is to discover productive ways of storing, retrieving, and controlling data, the chief concern of the machine learning and statistics is to build up techniques for extracting knowledge from datasets or

* *Corresponding author*. Tel.: +90 422 3410660; fax: +90 422 3410036; mobile: +90 505 8870498.
E-mail address: cemilcolak@yahoo.com (C. Colak).

http://dx.doi.org/10.1016/j.cmpb.2015.03.002
0169-2607/© 2015 Elsevier Ireland Ltd. All rights reserved.

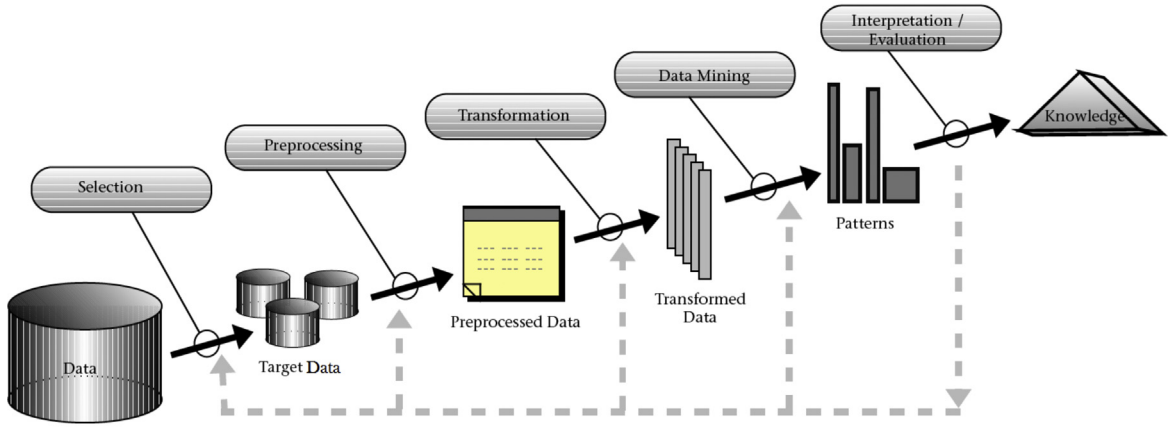


Fig. 1 – The KDP process [11].

databases [4]. Of the KDP methods, artificial neural networks (ANN) are one of the most widely used machine learning methods and have many distinguishing and valuable properties toward the traditional model-based methods [5]. Support vector machine (SVM) is one of the supervised machine learning strategies utilized generally in recognizing patterns and classification issues. The SVM accomplishes a classification by building a multi-dimensional hyperplane that ideally segregates between two classes by boosting the edge between two data groups [6].

As related with the prediction of stroke, a study used a machine learning method, SVM for predicting stroke thrombolysis outcome, and area under curve (AUC) of this SVM model was 0.744. The study demonstrated that SVM performed higher accuracy than radiological methods [7]. In another study, the outcome of ischemic stroke patients after intravenous thrombolysis was predicted using the two ANN models having the accuracy of 79.27% and 95.12%, respectively. The mentioned study reported that ANN had a good performance to predict thrombolysis outcomes [8]. In a different study, it was aimed to predict mortality in patients with stroke by utilizing ANNs trained with 6 various multilayer perceptron (MLP) algorithms. In the study, the MLP trained with the quick propagation algorithm produced the highest specificity of 81.3%, sensitivity of 78.4%, accuracy of 80.7% and AUC of 0.869 values [9].

The current study was performed to predict the outcome of stroke using KDP methods, artificial neural networks (ANN) and support vector machine (SVM) models.

2. Materials and methods

The protocol of this study was approved by the Ethical Review Board of the Medical Faculty of Inonu University, Malatya, Turkey (Protocol number: 2013/06). The current study was carried out in the department of emergency medicine, Turgut Ozal Medicine Center, Inonu University, Malatya, Turkey. The records of 130 stroke patients (patient group) and 167 healthy individuals (control group), 297 in total, were acquired from the databases of the department of emergency medicine. Brain computed tomography (CT) and/or magnetic resonance imaging (MRI) were employed in the diagnosis of stroke [10]. The variables of coronary artery disease, diabetes mellitus, hypertension, history of cerebrovascular disease, atrial fibrillation, smoking, the findings of carotid Doppler ultrasonography [normal, plaque, plaque + stenosis ≥ 50%], the levels of cholesterol and C-reactive protein (CRP) were used for predicting the stroke. KDP process was demonstrated in Fig. 1.

According to Fig. 1, the KDP process has five steps:

1. **Data selection:** Selecting data related to the analysis task from the database. In the current study, the target/response variable was absence/presence of stroke disease, and the predictor variables were coronary artery disease, diabetes mellitus, hypertension, history of cerebrovascular disease, atrial fibrillation, smoking, the findings of carotid Doppler ultrasonography, the levels of cholesterol and CRP.
2. **Data preprocessing:** Replacing missing observations and removing outliers, extreme values, noise and inconsistent data. In the present study, there were no missing observations, and multivariate outliers and extreme values in the data were detected using T^2 test based on the Mahalanobis distance.
3. **Data transformation and reduction:** Transforming data into convenient structures and finding useful features to implement data mining. In the present study, the continuous data were standardized (Mean = 0, Standard Deviation [SD] = 1). Feature selection based on the Cramer's V test was carried out for choosing the predictors.
4. **Data mining:** Choosing data mining or KDP algorithm(s) being suitable to pattern in the data; extracting data patterns. In this step, the following KDP methods were used in the training ($n = 209$) and testing datasets ($n = 78$), respectively.

2.1. *Artificial neural networks*

ANN models are commonly employed in pattern recognition and classification tasks by learning from data. Several neural network models are used for classifying or predicting the studied patterns [12]. In this study, the MLP ANN model was established, in which the neurons were organized in parallel layers and each layer was connected fully

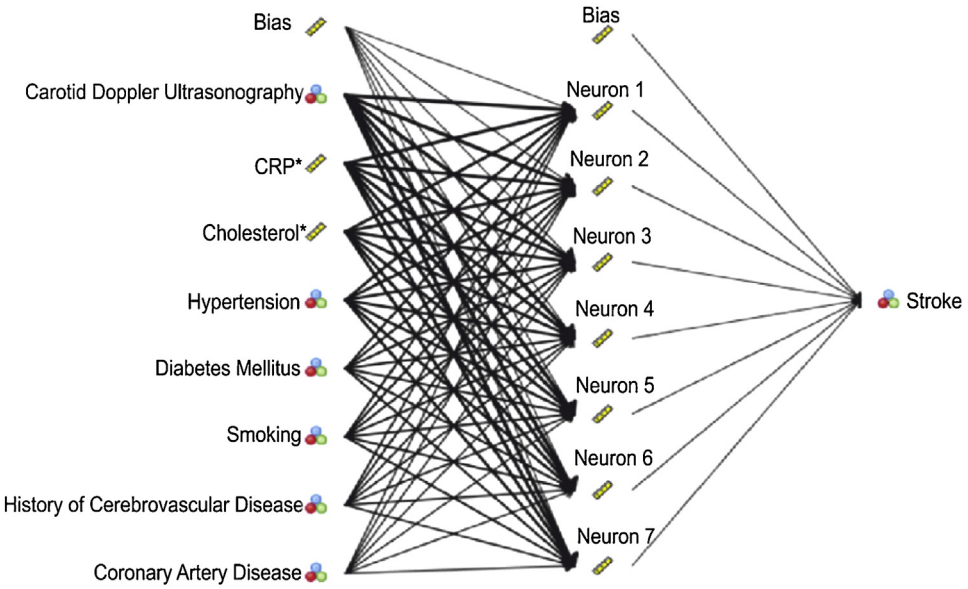


Fig. 2 – General structure for the MLP ANN (*: transformed predictor).

to the previous layer through synaptic connections. The input layer included the selected predictor variables, the findings of carotid Doppler ultrasonography, hypertension, history of cerebrovascular disease, CRP, diabetes mellitus, coronary artery disease, cholesterol and smoking. The hidden layer contained latent nodes, and the MLP ANN admitted one or two hidden layers. The output layer included the target variable, stroke disease [13,14]. MLP ANN had hidden layer with seven neurons, and hyperbolic tangent and softmax activation functions were used for the hidden and output layers, successively. General structure for the MLP ANN was shown in Fig. 2.

2.2. *Support vector machines*

SVM is one of the supervised learning methods for classification, prediction and regression analyses [15]. Also, SVM is used for nonlinear classification tasks and coping with high-dimensional data. SVM is employed in medical and biological researches and can map the input vectors to a high-dimensional space by utilizing several kernel functions (e.g., sigmoidal polynomial, hyperbolic tangent, radial basis and etc.) [16]. In the current study, SVM was constructed using radial basis function (RBF) kernel function with the regularization parameter (C) of 10 and gamma of 0.1. The grid search method was used for tuning of the optimal SVM parameters.

5. **Evaluation and interpretation:** Identifying the most suitable model(s) to obtain the targeted knowledge [17]. In this study, prediction performance of the models was evaluated based on the values of accuracy and AUC.

IBM SPSS Modeler Professional 16.0 for Windows was used in analyzing and modeling the data.

3. Results

The current study consisted of 130 (43.8%) stroke patients and 167 (56.2%) healthy individuals, 297 in total, initially. While 159 (53.5%) of the individuals were males, 138 (46.5%) were females. Mean ages were 59.4 ± 16.3 years for males and 63.3 ± 14.4 years for females, respectively.

According to the results of T^2 test based on the Mahalanobis distance, the determined 10 inconsistent records were discarded from the analysis, and the remaining data ($n = 287$) were used in the subsequent analyses. The levels of cholesterol and CRP were standardized (Mean = 0, SD = 1). After feature selection, the selected predictors with importance higher than 0.90 were the findings of carotid Doppler ultrasonography, hypertension, history of cerebrovascular disease, CRP, diabetes mellitus, coronary artery disease, cholesterol and smoking. Atrial fibrillation was excluded owing to their importance values smaller than 0.90. Predictor importance values were presented in Table 1.

Table 1 – Importance values of the predictors based on the feature selection.

Variables	Importance values
Carotid Doppler ultrasonography	1.0
Hypertension	1.0
History of cerebrovascular disease	1.0
CRP[a]	1.0
Diabetes mellitus	1.0
Coronary artery disease	1.0
Cholesterol[a]	0.992
Smoking	0.959
Atrial fibrillation	0.115

[a] Transformed to standard units (Mean = 0, SD = 1); CPR: C-reactive protein.

Table 2 – The values of relative predictor importance for the models.

Variables	Relative Predictor Importance	
	ANN	SVM
Carotid Doppler ultrasonography	0.23	0.40
Hypertension	0.13	0.15
History of cerebrovascular disease	0.06	0.14
CRP[a]	0.20	0.05
Diabetes mellitus	0.09	0.07
Coronary artery disease	0.03	0.08
Cholesterol[a]	0.17	0.04
Smoking	0.09	0.07

[a] Transformed to standard units (Mean = 0, SD = 1); CPR: C-reactive protein.

After selecting the predictors based on the feature selection, MLP ANN and SVM with RBF were used for the prediction of non-colorectal cancer/colorectal cancer based on the selected predictors of the findings of carotid Doppler ultrasonography, hypertension, history of cerebrovascular disease, CRP, diabetes mellitus, coronary artery disease, cholesterol and smoking. The values of relative predictor importance for the models were given in Table 2.

The accuracy values were 81.82% for ANN and 80.38% for SVM in the training dataset ($n = 209$), and 85.9% for ANN and 84.62% for SVM in the testing dataset ($n = 78$), respectively. ANN and SVM models yielded AUC values of 0.905 and 0.899 in the training dataset, and 0.928 and 0.91 in the testing dataset, consecutively.

4. Conclusions

The current study aimed at predicting the outcome of stroke using KDP methods, ANN and SVM models. KDP was used for discovering patterns and knowledge in the studied data. The KDP provided better understanding of the stroke data, saved costs and time, and reduced unnecessary processes to be performed. At first, we selected the dataset including the stroke disease (target) and the other predictors from the database, and removed the inconsistent data identified by the test mentioned earlier. Subsequently, the data were standardized and reduced based on the feature selection method in order to apply data mining techniques. Hereby, we used ANN and SVM models for extracting data patterns. Ultimately, the extracted models were evaluated and interpreted. Based on the findings of accuracy and AUC, ANN model had more predictive performance for stroke as compared with SVM.

Since care of stroke disease is a difficult and complex process, and essential clinical decisions need to be made rapidly during the first treatment phase of stroke [18], both model might predict the stroke based on the selected predictors for early diagnosis as a clinical decision support system. This system can be extended to large datasets including other predictors and other data mining techniques to improve further prediction performance.

As a result, the findings of the current study pointed out that ANN had more predictive performance when compared with SVM in predicting stroke. The proposed ANN model would be useful when making clinical decisions regarding stroke.

REFERENCES

[1] A. Arauz, Y. Rodriguez-Agudelo, A.L. Sosa, M. Chavez, F. Paz, M. Gonzalez, J. Coral, C. Diaz-Olavarrieta, G.C. Roman, Vascular cognitive disorders and depression after first-ever stroke: the Fogarty-Mexico stroke cohort, Cerebrovasc. Dis. 38 (2014) 284–289.

[2] A.O. Ogbera, O.O. Oshinaike, O. Dada, A. Brodie-Mends, C. Ekpebegh, Glucose and lipid assessment in patients with acute stroke, Int. Arch. Med. 7 (2014) 45.

[3] V. Aguiar-Pulido, J.A. Seoane, M. Gestal, J. Dorado, Exploring patterns of epigenetic information with data mining techniques, Curr. Pharm. Des. 19 (2013) 779–789.

[4] H. Kim, L. Soibelman, F. Grobler, Factor selection for delay analysis using knowledge discovery in databases, Autom. Constr. 17 (2008) 550–560.

[5] M. Khashei, A. Zeinal Hamadani, M. Bijari, A novel hybrid classification model of artificial neural networks and multiple linear regression models, Expert Syst. Appl. 39 (2012) 2606–2620.

[6] W. Yu, T. Liu, R. Valdez, M. Gwinn, M.J. Khoury, Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes, BMC Med. Inf. Decision Mak. 10 (2010) 16.

[7] P. Bentley, J. Ganesalingam, A.L. Carlton Jones, K. Mahady, S. Epton, P. Rinne, P. Sharma, O. Halse, A. Mehta, D. Rueckert, Prediction of stroke thrombolysis outcome using CT brain machine learning, NeuroImage: Clin. 4 (2014) 635–640.

[8] C.-A. Cheng, Y.-C. Lin, H.-W. Chiu, Prediction of the prognosis of ischemic stroke patients after intravenous thrombolysis using artificial neural networks, Stud. Health Technol. Inform. 202 (2013) 115–118.

[9] N. Süt, Y. Çelik, Prediction of mortality in stroke patients using multilayer perceptron neural networks, Turk. J. Med. Sci. 42 (2012) 886–893.

[10] J.A. Chalela, C.S. Kidwell, L.M. Nentwich, M. Luby, J.A. Butman, A.M. Demchuk, M.D. Hill, N. Patronas, L. Latour, S. Warach, Magnetic resonance imaging and computed tomography in emergency assessment of patients with suspected acute stroke: a prospective comparison, Lancet 369 (2007) 293–298.

[11] U. Fayyad, G. Piatetsky-Shapiro, P. Smyth, From data mining to knowledge discovery in databases, AI Mag. 17 (1996) 37.

[12] M. Hariharan, R. Sindhu, S. Yaacob, Normal and hypoacoustic infant cry signal classification using time–frequency analysis and general regression neural network, Comput. Methods Programs Biomed. 108 (2012) 559–569.

[13] W. Hongfei, Z. Yunyan, Y. Fei, L. Hui, Evaluation of an artificial neural network to ascertain why there is a high incidence of hepatitis B in the Chinese population after vaccination, Comput. Biol. Med. 43 (2013) 1167–1170.

[14] M.C. Colak, C. Colak, H. Kocatürk, S. Sağiroğlu, I. Barutçu, Predicting coronary artery disease using different artificial neural network models, Anadolu kardiyol. derg.: AKD 8 (2008) 249–254.

[15] H. Mohamed, M.S. Mabrouk, A. Sharawy, Computer aided detection system for micro calcifications in digital mammograms, Comput. Methods Programs Biomed. 116 (2014) 226–235.

[16] S. Korkmaz, G. Zararsiz, D. Goksuluk, Drug/nondrug classification using support vector machines with various feature selection strategies, Comput. Methods Programs Biomed. 117 (2014) 51–60.

[17] T. Silwattananusarn, K. Tuamsuk, Data mining and its applications for knowledge management: a literature review from 2007 to 2012, Int. J. Data Min. Knowl. Manag. Process (IJDKP) 2 (2012) 13–24.

[18] O.J. Gibson, J.S. Balami, G.A. Pope, L. Tarassenko, I.P. Reckless, Stroke Nav: a wireless data collection and review system to support stroke care delivery, Comput. Methods Programs Biomed. 108 (2012) 338–345.