An expert system for the prediction of stroke disease by different least squares support vector machines models.

Mehmet Ediz Sarihan\textsuperscript{1*}, Davut Hanbay\textsuperscript{2}

\textsuperscript{1}Department of Emergency Medicine, Faculty of Medicine, Inonu University, Malatya, Turkey
\textsuperscript{2}Department of Computer Engineering, Faculty of Engineering, Inonu University, Malatya, Turkey

Abstract

Objective: One of the important life-threatening ailments is stroke across the world. The current paper was performed to classify the outcome of stroke by using Least-Squares Support Vector Machines (LS-SVMs) models.

Materials and methods: The medical dataset related to stroke disease was achieved from the clinical database of the emergency medicine department. 28 predictors were recorded in raw dataset. For dimension reduction, correlations between input and target (stroke) variables were evaluated. Different LS-SVMs models were performed with radial basis function (RBF), linear and polynomial kernels. 5-fold cross-validation was used in composing stages to achieve the best model using all of the data. The accuracy and the Area under Receiver Operating Curve (AUC ROC) values were used for performance assessment.

Results: At first, feature selection stage was performed. 14 input variables were determined after this stage. Whole dataset was partitioned into 5 sub-datasets ($D_1$, $D_2$, $D_3$, $D_4$, $D_5$) to use all data both training and testing. LS-SVMs models performance were evaluated by using 5-fold cross validation method. Accuracy and AUC values of the models were used as performance criteria. The best model performance was evaluated with LS-SVMs model using linear kernel. That model average accuracy was 86.6%. The best accuracy was evaluated with LS-SVM model using linear kernel on dataset $D_5$ was 94%. As a consequence, the LS-SVMs model can be used for predicting the outcome of stroke.

Conclusion: The results point out that LS-SVMs with linear kernel have much more accuracy and AUC values for predicting stroke disease. The suggested LS-SVMs with linear kernel may produce beneficial prediction results related to stroke disease. In future studies, several data mining techniques may be tested and assembled for better classification performance of stroke disease.

Keywords: Data mining, Stroke disease, Least square support vector machines (LS-SVMs).

Accepted on September 26, 2017

Introduction

Stroke is the significant reason of vascular behaviour and mentality disorderliness over the worldwide. In thriving countries, a shortage of information on the public health problem of stroke is present [1]. Stroke is an expanding illness and is an important reason of death worldwide following coronary heart disease and cancer ailment. Stroke frequently is the result of enhanced morbidity/mortality and lessened quality of life [2,3].

Data mining is a process of pattern discovery from a potentially large amount of data and is a multi-disciplinary topic that is conceived on the basis of logics in database systems. Examples of data mining techniques are Decision Trees, A priori Algorithm, Artificial Neural Networks (ANNs), Support Vector Machines (SVMs) and so on. Data mining can also be used in information technology evolving and subsequently branching off into sub-processes that include collecting data, creating database and management, analyzing data and finally interpreting data [4]. SVMs are one of the supervised machine learning (ML) approaches [5]. Since, then it is used widely in pattern recognition for regression and classification problems [6,7]. The LS-SVMs perform a classification by establishing a complex hyperplane optimally discriminating between two categories [6,7]. Kernel functions such as radial basis function (RBF), linear and polynomial are very powerful in mapping data into a larger dimensional domain and assist LS-SVMs to excellently separate data with very complex boundaries [8]. The LS version of the SVMs was described by Suykens and Vandewalle [9]. LS-SVMs are widely used in complex system studies [10].

In relation to the estimation of stroke, a study proposed SVMs in order to classify stroke thrombolysis, and the SVMs model yielded area under curve (AUC) of 0.744. The work showed that SVMs produced larger accuracy value than conventional,
radiology-based techniques [11]. Another work investigated whether functional magnetic resonance imaging allowed classification of personal motor damage following stroke employing linear SVMs, and forty acute stroke people and 20 controls underwent resting-state functional magnetic resonance imaging for their goal. 82.6-87.6% accuracy was reported from the study [12]. Another study made a comparative analysis in the usage of SVMs with several kernel functions for stroke patients. The study investigated classifications accuracies of RBF, quadratic and polynomial kernel functions for different SVMs models [13]. Another study investigated SVMs for classifying the walking conditions of individuals following stroke, and reported that the predictive performance of SVMs model was higher than that of alternative data mining approaches utilizing RBF ANNs and ANNs [14].

In this work, an intelligent model is performed to predict the outcome of stroke disease using different LS-SVMs models. In section 2; the basic of the current study is described in details. In section 3; the application of the paper is explained. In section 4; conclusions are given.

**Materials and Methods**

**Database**

The current work was performed in the emergency medicine department, Turgut Ozal Medicine Center, Medical Faculty, Inonu University, Malatya, Turkey. Between January 2012 and January 2013, the medical enrollments of 104 individuals with stroke illness (patient group) and 104 healthy people (control group) were achieved from the database of the emergency medicine department. The recorded variables/factors were age (years), gender (1: female/2: male), educational status (1: primary school/2: middle school/3: high school/4: university/5: illiterate), marital status (MS; 1: married/2: single/3: widowed), application location for the emergency medicine service (LOC; 1: from home/ 2: from work/3: from any hospital), smoking status (SMO; 1: present/2: absent), coronary artery disease (1: present/2: absent), diabetes mellitus (1: present/2: absent), hypertension (1: present/2: absent), revascularization (REVAS; 1: angioplasty/2: stent/3: coronary bypass), hyperlipidemia (1: present/2: absent), electrocardiography (ECG; 1: present/2: absent), alcohol consumption (1: present/2: absent), congestive heart failure (CHF; 1: present/2: absent), systolic blood pressure (SBP; mmHg), diastolic blood pressure (DBP; mmHg), white blood cell (10^3/ML), hemoglobin (HB; g/dL), hematocrit (%), platelet (10^3/ML), glucose (mg/dL), blood urea nitrogen (mg/dL), creatinine (mg/dL), sodium (Na; mmol/L), potassium (K; mmol/L), chlorine (CL; mmol/L), calcium (mg/dL), and international normalized ratio (INR; %). After correlation based feature selection approach [15], 14 of 28 input variables were used for predicting the outcome (target; 1: present/2: absent) of stroke disease. These variables were gender, age, MS, LOC, HT, REVAS, ECG, SMO, CHF, SBP, DBP, HB, K, and CL. A brief explanation of the database used in this study is shown in Table 1. The predictors included in this study are similar with the risk factors of stroke disease reported by other clinical research articles [16-19].

<table>
<thead>
<tr>
<th>Target</th>
<th>Gender</th>
<th>Age</th>
<th>MS</th>
<th>LOC</th>
<th>HT</th>
<th>REVAS</th>
<th>ECG</th>
<th>SMO</th>
<th>CHF</th>
<th>SBP</th>
<th>DBP</th>
<th>HB</th>
<th>K</th>
<th>CL</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1</td>
<td>62</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>174</td>
<td>85</td>
<td>13</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>63</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>180</td>
<td>100</td>
<td>16</td>
<td>4</td>
<td>107</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>80</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>190</td>
<td>105</td>
<td>16</td>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>78</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>190</td>
<td>108</td>
<td>13</td>
<td>5</td>
<td>108</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>77</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>199</td>
<td>100</td>
<td>16</td>
<td>5</td>
<td>109</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>58</td>
<td>3</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>107</td>
<td>70</td>
<td>15</td>
<td>4</td>
<td>99</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>50</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>188</td>
<td>108</td>
<td>15</td>
<td>5</td>
<td>101</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>81</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>129</td>
<td>74</td>
<td>12</td>
<td>4</td>
<td>111</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>76</td>
<td>3</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>185</td>
<td>98</td>
<td>14</td>
<td>4</td>
<td>106</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>79</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>166</td>
<td>98</td>
<td>13</td>
<td>4</td>
<td>112</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>76</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>185</td>
<td>85</td>
<td>15</td>
<td>5</td>
<td>109</td>
</tr>
<tr>
<td>1</td>
<td>1</td>
<td>74</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>220</td>
<td>92</td>
<td>13</td>
<td>4</td>
<td>104</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>57</td>
<td>1</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>120</td>
<td>76</td>
<td>11</td>
<td>4</td>
<td>109</td>
</tr>
<tr>
<td>2</td>
<td>2</td>
<td>25</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>125</td>
<td>76</td>
<td>16</td>
<td>4</td>
<td>108</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>37</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>2</td>
<td>153</td>
<td>116</td>
<td>12</td>
<td>5</td>
<td>108</td>
</tr>
</tbody>
</table>
Least square support vector machines (LS-SVMs)

SVMs are one of the supervised ML techniques developed by Vapnik et al. at AT&T Bell Laboratories in 1995 [20]. It can be used for both classification and regression tasks in any discipline. The SVMs are based on the principle of structural risk minimization [20,21].

If a given training set \( \{x_k, y_k\}_{k=1}^{N} \) with input data \( x_k \in \mathbb{R}^n \) and output data \( y_k \in \mathbb{R} \) with class labels \( y_k \in \{-1, +1\} \) and linear classifier

\[
\hat{y}(x) = \text{sign}(w^T x + b) \rightarrow (1)
\]

If two classes can be separable then

\[
\begin{align*}
\text{If } w^T x_k + b \geq +1, \text{ if } y_k = +1 \\
\text{if } w^T x_k + b \leq -1, \text{ if } y_k = -1
\end{align*}
\]

These two equations can be combined and reduced to one equation as in Eq. 3.

\[
y_k = |w^T x_k| \geq 1 \quad k = 1, \ldots, N \rightarrow (3)
\]

SVMs subject is a concept of convex optimization theory. At first, the problem is stated as a constrained optimization problem. Then Lagrangian is formulated and the conditions for optimality are determined; finally, the problem is solved in the dual space of Lagrange multipliers with Eq. 4.

\[
y(x) = \text{sign}\left[\sum_{k=1}^{N} \alpha_k y_k x_k^T x + b\right] \rightarrow (4)
\]

Cortes & Vapnik were extended this linear SVMs classifier to non-separable case. It is done by adding slack variable in the problem formulation as in Eq. 5.

\[
y_k = w^T x_k + b \geq 1 - \xi_k \quad k = 1, \ldots, N \rightarrow (5)
\]

The SVMs have not been used only for linear function estimation; but also they have been used for nonlinear function estimation too.

The least square type of the SVMs methods were proposed Suykens and Vandewalle [21]. In the LS-SVMs methods equality type constraints are considered instead of inequalities [22]. This reformulation greatly simplifies a problem such that the LS-SVMs solution follows directly from solving a sequence of linear formulas rather than from a convex quadratic program. The LS-SVMs classifier, in the primal space can be described by Eq. 6.

\[
y(x) = \text{sign}(w^T \phi(x) + b) \rightarrow (6)
\]

Where \( \phi(.) \) map from input space to feature space and b is a real constant. For nonlinear classification, the LS-SVM classifier in the dual space it takes the form

\[
y(x) = \text{sign}\left[\sum_{k=1}^{N} \alpha_k y_k \phi(x)\phi(x_k) + b\right] \rightarrow (7)
\]

For function estimation, the LS-SVM model can be described by Eq. 8.

\[
y(x) = \sum_{k=1}^{N} \alpha_k \phi(x)\phi(x_k) + b \rightarrow (8)
\]

There are many types of kernels used with SVMs and LS-SVMs: the most known of them are linear, polynomial and radial basic function kernels. These kernels are tabulated in Table 2.

Table 2. Kernel Types of LS-SVMs.
Kernels Types | Equations
--- | ---
Linear | \( k(x, y) = x^T y \)
Polynomial | \( k(x, y) = (\alpha x^T y + c)^d \)
RBF | \( e^{-\gamma \|x - y\|^2} \), \( \gamma > 0 \)

The Linear kernel is the simplest kernel function. It is given by the inner product \( <x, y> \) plus an optional constant \( c \). When RBF kernels are used, two tuning parameters \( (\gamma, \alpha) \) are added. Where \( \Phi(.) \) map from input space to feature space and \( b \) is a real constant, \( K(.) \) is kernel function, \( \gamma \) is regularization constant, and \( \sigma \) is width of RBF kernel. The Polynomial kernel is a non-stationary kernel. Polynomial kernels are well suited for problems where all the training data are normalized [3].

**Results**

This study was launched to estimate the outcome of stroke disease using several LS-SVMs models. After correlation based feature selection, the aforementioned variables were used for predicting the outcome of stroke disease. Different LS-SVMs models with RBF, Linear and Polynomial kernels were composed. 5-fold cross-validation was used to evaluate the models performance by using all data. The instances having empty values were ignored too. All program codes were written in MATLAB.

In 5-fold cross-validation, the stroke database was partitioned randomly into 5 sub-datasets, and training and testing were repeated for 5 times. Average accuracy of 5 models was accepted as model accuracy. To gauge the classifiers’ performance, accuracy and area under receiver operating characteristic curve (ROC) were considered as performance metrics.

At first, accuracy and ROC values of LS-SVMs models were evaluated. Averaged accuracy percentages were 85.6% for LS-SVMs with RBF kernel, 86.6% for LS-SVMs with linear kernel, and 74.4% for LS-SVMs with polynomial kernel. As a consequence, the best LS-SVMs model was obtained with linear kernel.

Table 3 presents the results of LS-SVMs models. Based on the Table 3, the highest average accuracy is 86.6% for LS-SVM with linear kernel and 74.4% for LS-SVMs with polynomial kernel. As a consequence, the best LS-SVMs model was obtained with linear kernel.

At first, accuracy and ROC values of LS-SVMs models were evaluated. Averaged accuracy percentages were 85.6% for LS-SVMs with RBF kernel, 86.6% for LS-SVMs with linear kernel, and 74.4% for LS-SVMs with polynomial kernel. As a consequence, the best LS-SVMs model was obtained with linear kernel.

Table 3. The results of LS-SVMs models using different kernels.

<table>
<thead>
<tr>
<th>Testing Dataset</th>
<th>RBF Accuracy (%)</th>
<th>Linear Accuracy (%)</th>
<th>Polynomial Accuracy (%)</th>
<th>RBF AUC</th>
<th>Linear AUC</th>
<th>Polynomial AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>D_1</td>
<td>86</td>
<td>86</td>
<td>78</td>
<td>0.8591</td>
<td>0.9732</td>
<td>0.775</td>
</tr>
<tr>
<td>D_2</td>
<td>86</td>
<td>86</td>
<td>69</td>
<td>0.8875</td>
<td>0.9732</td>
<td>0.6905</td>
</tr>
<tr>
<td>D_3</td>
<td>81</td>
<td>78</td>
<td>69</td>
<td>0.8372</td>
<td>0.9881</td>
<td>0.694</td>
</tr>
</tbody>
</table>

**Figure 1.** \( D_3 \) test result with linear kernel and optimized gamma parameter of 4.962.

Simplex cost function optimization routine was used for tuning the LS-SVM kernel parameters. Sample ROC graphic for \( D_3 \) is shown in Figure 1 with linear kernel and optimized gamma parameter of 4.962.

**Conclusion**

In the first stage of the current study, we investigated the possible use of LS-SVMs models by different kernels in the prediction of stroke. In the second stage, performance of LS-SVMs models was compared for predicting the outcome of stroke and compared based on the accuracy rates and AUC values. The obtained results of this work indicated that LS-SVMs with linear kernel had more accuracy and AUC for the prediction stroke disease.

The current study demonstrated the possible use of LS-SVMs models by different kernels in the prediction of stroke considering a small set of clinical variables. When the suggested model includes larger data sets, containing many other demographical and clinical variables associated with stroke disease, the prediction performance may be higher.

The results point out that LS-SVMs with linear kernel have much more accuracy and AUC values as compared with other LS-SVMs models in predicting stroke disease. The suggested LS-SVMs with linear kernel may produce beneficial prediction results related to stroke disease. In future studies, several data mining techniques may be combined for better classification performance of stroke disease.

**References**

An expert system for the prediction of stroke disease by different least squares support vector machines models

Correspondence to
Mehmet Ediz Sarihan
Department of Emergency Medicine
Faculty of Medicine
Inonu University
Malatya
Turkey