Prediction of cholesterol level in patients with myocardial infarction based on medical data mining methods

Cemil Colak1,*, Mehmet C. Colak2, Necip Ermis3, Nevzat Erdil2, Ramazan Ozdemir2

1Dept. of Biostatistics and Medical Informatics, Inonu University, Faculty of Medicine, Malatya, Turkey
2Dept. of Cardiovascular Surgery, Inonu University, Faculty of Medicine, Malatya, Turkey
3Dept. of Cardiology, Inonu University, Faculty of Medicine, Malatya, Turkey

*Corresponding author: cemilcolak@yahoo.com

Abstract

Myocardial infarction (MI) is a significant reason for death and disability over the world and might be the first sign of coronary artery disease. The current study was carried out to predict the cholesterol level in patients with MI using data mining methods, artificial neural networks (ANNs) and support vector machine (SVM) models. The data of 596 patients, who had been diagnosed with segment elevation MI were analysed in the present study. The retrospective dataset including gender, age, weight, height, pulse, glucose, creatinine, triglyceride, high-density lipoprotein, and low-density lipoprotein was used for predicting the cholesterol level. Correlation based feature selection was applied. Multilayer perceptron (MLP) ANNs and SVM with radial basis function kernel were used for the prediction based on the selected predictors. The performance of the ANNs and SVM models was evaluated on the basis of correlation coefficient and mean absolute error. The estimated correlation coefficients observed and predicted values were 0.94 for ANNs and 0.88 for SVM in training dataset (n=376), and 0.95 for ANNs and 0.90 for SVM in testing dataset (n=160), respectively. ANNs and SVM models yielded mean absolute error of 7.37 and 14.18 in training dataset, and 7.87 and 14.71 in testing dataset, consecutively. The results of the performance evaluation showed that MLP ANNs performed better for the prediction of cholesterol level in patients with MI in comparison to SVM. The proposed MLP ANNs model might be employed for predicting the level of cholesterol for MI patients in clinical decision support process.

Keywords: Artificial neural networks (ANNs); cholesterol level; medical data mining; myocardial infarction (MI); support vector machine (SVM).

1. Introduction

Myocardial infarction (MI) can be perceived by clinical peculiarities and is a real reason for death and handicap around the world. MI may be one of the first indications of coronary artery disease (CAD). MI may have major mental and legitimate ramifications for the society and is an indicator of one of the health problems over the world (Thygesen et al., 2012). Worldwide, cardiovascular sickness is assessed to be the main reason for death. Powerful prevention needs a worldwide strategy focused on the importance of risk factors for cardiovascular sickness in diverse geographic districts (Yusuf et al., 2004).

The reasons for CAD are multifactorial. Some of these variables are modified and are associated with ways of life; for example, tobacco smoking, absence of physical movement, and dietary propensities. Other variables are non-modifiable, such as age and male gender. Studies discovered an immediate relationship between levels of low-density lipoprotein (LDL) cholesterol and the rate of new-onset CAD in males and females, who have no CAD at first (Investigators, 1992; Stamler et al., 1986; Wilson et al., 1998). The same relation holds for repetitive coronary occasions in individuals with situated CAD.

Among the supervised machine learning methods, artificial neural networks (ANNs) are computer based programs, and accumulate their knowledge from input-output relationships in datasets (Colak et al., 2008). Support vector machine (SVM) is one of the supervised machine learning methods used widely in pattern recognition and classification problems and performs a classification by building a multidimensional hyperplane that ideally segregates between two classes (Yu et al., 2010).
Of the studies related with estimation of the probability of MI, a study used logistic regression and ANNs models based on patient clinical history variables, and reported that both models can predict successfully the likelihood of myocardial infarction according to some factors alone (Wang et al., 2001). Another study assessed an improvement achieved by ensemble-based methods, bootstrap aggregation of regression trees, random forests, and boosted regression trees in patients with either acute MI or congestive heart failure. The study reported that ensemble methods of data mining boosted the prediction performance of regression trees.

To our knowledge, no study has been reported on the prediction of cholesterol level in MI patients using medical data mining methods. Consequently, the current study attempted to predict the cholesterol level in patients with MI using medical data mining methods, ANNs and SVM models.

2. Materials and Methods

2.1. Dataset

The studied data consisted of a sample without replacement from the database of Cardiology Department of Turgut Ozal Medical Center, Malatya, Turkey, between 2010 and 2013. The data of 596 patients, who had been diagnosed with segment elevation MI in pursuant of second universal definition of myocardial infarction guideline (Thygesen et al., 2007a; Thygesen et al., 2007b) were analysed in the present study. The retrospective dataset included cholesterol level, gender, age, weight, height, pulse, glucose, creatinine, triglyceride, high-density lipoprotein (HDL), and LDL.

Table 1 defines the details of the target and input attributes.

Table 1. The details of the target and input attributes

<table>
<thead>
<tr>
<th>Attribute</th>
<th>Attribute type</th>
<th>Role</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cholesterol level</td>
<td>Numerical</td>
<td>Target</td>
</tr>
<tr>
<td>Gender (female/male)</td>
<td>Categorical</td>
<td>Input</td>
</tr>
<tr>
<td>Age (years)</td>
<td>Numerical</td>
<td>Input</td>
</tr>
<tr>
<td>Weight (kg)</td>
<td>Numerical</td>
<td>Input</td>
</tr>
<tr>
<td>Height (m)</td>
<td>Numerical</td>
<td>Input</td>
</tr>
<tr>
<td>Pulse (beats per minute)</td>
<td>Numerical</td>
<td>Input</td>
</tr>
<tr>
<td>Glucose (mg/dL)</td>
<td>Numerical</td>
<td>Input</td>
</tr>
<tr>
<td>Creatinine (mg/dL)</td>
<td>Numerical</td>
<td>Input</td>
</tr>
<tr>
<td>Triglyceride (mg/dL)</td>
<td>Numerical</td>
<td>Input</td>
</tr>
<tr>
<td>HDL (mg/dL)</td>
<td>Numerical</td>
<td>Input</td>
</tr>
<tr>
<td>LDL (mg/dL)</td>
<td>Numerical</td>
<td>Input</td>
</tr>
</tbody>
</table>

2.2. Knowledge discovery in databases process

The knowledge discovery in databases (KDD) process is given in Figure 1.

Fig. 1. An overview of the data mining step and additional steps in the KDD Process (Fayyad et al., 1996)

According to Figure 1, KDD process contains five steps:

- Data selection: Selecting data related to the analysis task from the database.
- Data pre-processing: Removing outliers, extreme values, noise and inconsistent data.
- Data transformation: Transforming data into convenient structures to implement data mining.
- Data mining: Choosing data mining algorithm(s) being suitable to pattern in the data; extracting data patterns.
- Evaluation and interpretation: Identifying the most suitable model(s) to obtain the targeted knowledge (Silwattananusarn & Tuamsuk, 2012).

2.3. Power analysis and software

The power analysis calculated minimum 265 subjects with the supposed cholesterol difference of 5, assumed standard deviation of 25, type I error ($\alpha$) of 0.05 and type II error ($\beta$) of 0.10. For analysing and modelling the data, IBM SPSS Modeler Professional 16.0 for Windows was employed.

3. Results

The current study initially included 298 (50.0%) male and 298 (50.0%) female MI patients, 596 in total. Mean age of the patients was 68.3±12.3 y. The KDD is explained in the following steps.

3.1. Data Selection

The target was cholesterol level (numerical target attribute), and the predictors were gender, age, weight, height, pulse,
glucose, creatinine, triglyceride, high-density lipoprotein (HDL) and low-density lipoprotein (LDL).

3.2. Data pre-processing

Multivariate outliers and extreme values in the data were detected using $T^2$ test based on the Mahalanobis distance. This method uses a hypothesis testing based on the $T^2$ probability levels to test multiple extreme values. The determined 60 inconsistent instances were discarded, and further analyses were performed on the remaining instances ($n=536$).

3.3. Data transformation and reduction

Feature selection based on correlation was carried out for reducing the predictors. The selected predictors were LDL, triglyceride, HDL, age and gender, respectively. The variables of weight, height, pulse, glucose and creatinine were not selected as a result of feature selection. The chosen numerical attributes of LDL, triglyceride, HDL and age were transformed to standard units (Mean=0, Standard Deviation=1) called Z-transformation.

3.4. Data mining

Multilayer perceptron (MLP) ANNs and SVM with radial basis function kernel were used for the prediction of cholesterol level based on the selected predictors of LDL, triglyceride, HDL, age and gender. The applied ANNs structure was displayed in the Figure 2. MLP is one of the most popular neural network architectures and is a supervised network, owing to the fact that it calls for a desired output to learn. MLP includes an input layer with neurons (input variables), an output layer with neurons (target variable), and one or more hidden layers containing neurons to discover the nonlinearity in the data (Hongfei et al., 2013; Süt & Çelik, 2012). The MLP ANNs included a hidden layer with 5 neuron, and activation functions for the hidden output layers were hyperbolic tangent and identity, respectively. SVM is a supervised machine learning technique in using classification and regression routines (Moses, 2015; Shin et al., 2014), and showed a good performance in solving medical and biological classification and prediction problems (Arslan et al., 2016; Colak et al., 2015; Zhou et al., 2014). Among the kernel functions, RBF for SVM was selected in the current study. As a result of the implementation of grid search algorithm, the regularization parameter (C), regression precision (epsilon) and RBF gamma were determined as 10, 0.1 and 0.1, respectively.

Relative predictor importance for the chosen variables is presented in Table 2 on the basis of the results of ANNs and SVM models; while relative predictor importance for ANNs model in descending order was LDL, HDL, triglyceride, age and gender. The importance arrangement for SVM model in descending order was LDL, HDL, age, gender and triglyceride.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Relative predictor importance</th>
</tr>
</thead>
<tbody>
<tr>
<td>LDL*</td>
<td>0.55</td>
</tr>
<tr>
<td>HDL*</td>
<td>0.22</td>
</tr>
<tr>
<td>Triglyceride*</td>
<td>0.18</td>
</tr>
<tr>
<td>Age*</td>
<td>0.03</td>
</tr>
<tr>
<td>Gender</td>
<td>0.02</td>
</tr>
</tbody>
</table>

*: transformed to standard units (Mean=0, SD=1)

3.5. Evaluation and interpretation

Using holdout technique, the dataset was divided into two sets: 70% of the dataset ($n=376$) for training the models, and 30% of the dataset ($n=160$) for testing the models. The performance of the ANNs and SVM models was assessed on the basis of correlation coefficient and mean absolute error. Table 3 tabulates the details of the model evaluation. The estimated correlation coefficients of observed and predicted values were 0.94 for ANNs and 0.88 for SVM in training dataset ($n=376$), and 0.95 for ANNs and 0.90 for SVM in testing dataset ($n=160$), respectively. ANNs and SVM models yielded mean absolute error of 7.37 and 14.18 in training dataset, and 7.87 and 14.71 in testing dataset, consecutively.
4. Conclusions

The current study attempted to predict the cholesterol level in patients with MI using medical data mining methods, ANNs and SVM models. To achieve this objective, we used the knowledge discovery process for extracting knowledge from data. In the first and second steps, the data related to cholesterol level were selected, and since there were outliers and extreme values, the inconsistent data were removed to increase the prediction performance of cholesterol level. In the third step, the data were transformed to convenient structures to implement data mining, and a subset of relevant features was selected in the model construction. As for the fourth step, we applied two medical data mining algorithms, ANNs and SVM for extracting data patterns. Finally, when the values of correlation coefficient and mean absolute error were evaluated and interpreted, ANNs yielded the best performance as compared with SVM in training and testing datasets.

Coefficient of determination ($R^2$) values were calculated as 0.902 for ANNs and 0.810 for SVM models. This finding also demonstrated that ANNs produced higher $R^2$ value in the prediction of the cholesterol level in patients with MI, when compared with SVM model. As for the selected predictors, both models used 5 features selected out of 10 features for optimal prediction of cholesterol.

As a consequence, the results of the performance evaluation showed that MLP ANNs performed better for the prediction of cholesterol level in patients with MI in comparison to SVM. The proposed MLP ANNs model might be employed for predicting the level of cholesterol for MI patients in clinical decision support process.

References


Submitted : 29/12/2014
Revised : 15/02/2016
Accepted : 17/02/2016
التنبؤ بمستوى الكوليسترول في الدم في المرضى الذين يعانون من احتشاء عضلة القلب على أساس أساليب تلخيص البيانات الصناعية

سيمي كولاك، نبيل إزمير، نيفزا إردين، رمضان أوزديمیر

جامعة إسطنبول - كلية الطب - قسم الإحصاء الحيوي والعلوم الطبية - مالاتيا - تركيا

المؤلف:

cemilcolak@yahoo.com

خلاصة

احتشاء عضلة القلب (MI) هو سبب كبير للوفاة والعجز في العالم، وربما يكون أول بادرة من مرض الشريان التاجي. وتهدف الدراسة الحالية إلى التنبؤ بمستوى الكوليسترول في الدم في المرضى الذين يعانون MI باستخدام أساليب تلخيص البيانات، والشبكات العصبية الاصطناعية ومثير آلة الدعم الموجه (SVM). وقد تم في هذه الدراسة تحليل بيانات 596 مريضا، والذين كان قد تم تشخيصهم مع شريحة ارتفاع MI. تم استخدام بيانات أثر رجعي في ذلك الجنس والعمر والوزن والطول، والنضج والجلوكوز والكوليسترولات، والدهون الثلاثية، البروتين الدهني عالى الكثافة، والبروتين الدهني منخفض الكثافة للتنبؤ بمستوى الكوليسترول في الدم. تم اختبار SVM المربعة على أساس الارتباط واستخدام مستقبلات متعددة الطبقات (MLP) في الشبكات العصبية الاصطناعية وSVM المزدوج. وقد تم تقييم A وارتباط الشبكات العصبية الاصطناعية ونماذج على أساس معامل الارتباط ومتوسط الخطأ المطلق. وتتم ملاحظة معاملات الارتباط المقدرة وكانت القيمة المتوقعة 0.94 للشبكات العصبية الاصطناعية و 0.88 لـ SVM في اختبار مجموعة بيانات (N = 376) و 0.90 للشبكات العصبية الاصطناعية و 0.90 لـ SVM في اختبار مجموعة بيانات (N = 160). على التوالي. أظهرت النتائج تقييم أداء شبكات العصبية الاصطناعية SVM في قيم تحويل الخطأ المطلق من 3.97 و 7.38 في بيانات التدريب و 7.08 و 14.71 في اختبار مجموعة البيانات، على التوالي. أظهرت النتائج تقييم أداء شبكات العصبية الاصطناعية SVM في قيم تحويل الخطأ المطلق من 3.97 و 7.38 في بيانات التدريب و 7.08 و 14.71 في اختبار مجموعة البيانات، على التوالي. أظهرت النتائج تقييم أداء شبكات العصبية الاصطناعية SVM في قيم تحويل الخطأ المطلق من 3.97 و 7.38 في بيانات التدريب و 7.08 و 14.71 في اختبار مجموعة البيانات، على التوالي. أظهرت النتائج تقييم أداء شبكات العصبية الاصطناعية SVM في قيم تحويل الخطأ المطلق من 3.97 و 7.38 في بيانات التدريب و 7.08 و 14.71 في اختبار مجموعة البيانات، على التوالي. أظهرت النتائج تقييم A وارتباط الشبكات العصبية الاصطناعية النماذج للتنبؤ بمستوى الكوليسترول في الدم مرضي MI بالمقارنة مع MI في عملية دعم اتخاذ القرارات السريرية.

الكلمات النهائية:

الشبكات العصبية الاصطناعية، مستوى الكوليسترول في الدم، تلخيص (تنقيح) البيانات، احتشاء عضلة القلب (MI)، آلة الدعم الموجه (SVM).