Kuwait J. Sci. **43 (3)** pp. 86-90, 2016

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

Cemil Colak[1,*], Mehmet C. Colak[2], Necip Ermis[3], Nevzat Erdil[2], Ramazan Ozdemir[3]

[1]*Dept. of Biostatistics and Medical Informatics, Inonu University, Faculty of Medicine, Malatya, Turkey*

[2]*Dept. of Cardiovascular Surgery, Inonu University, Faculty of Medicine, Malatya, Turkey*

[3]*Dept. 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

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

2.2. Knowledge discovery in databases process

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

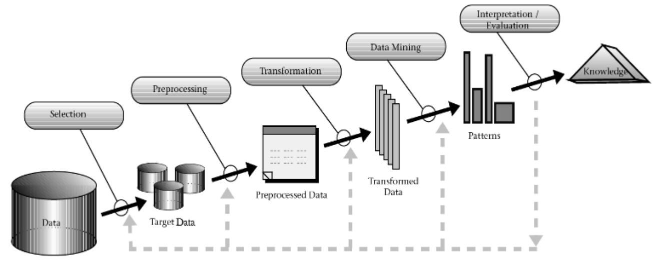

Fig. 1. An overview of the data mining step and additional steps in the KDD Process (Fayyad *et al.*, 1996)

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 (α) of 0.05 and type II error (β) 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.

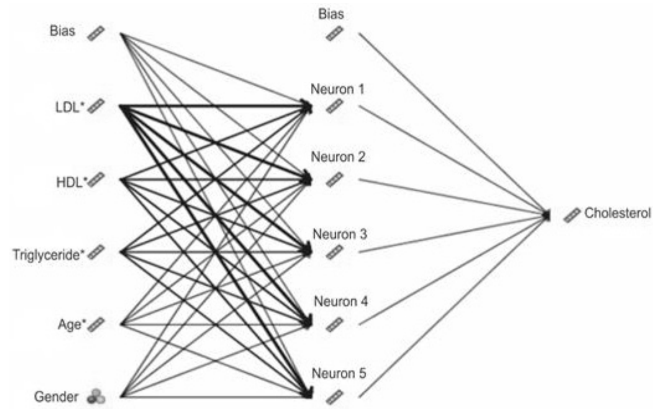

Fig. 2. The structure of applied ANNs model (*: transformed variable)

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 2. Relative predictor importance for the selected variables

	Relative predictor importance	
Variables	**ANNs**	**SVM**
LDL*	0.55	0.30
HDL*	0.22	0.24
Triglyceride*	0.18	0.11
Age*	0.03	0.19
Gender	0.02	0.16

*: 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 14.71in testing dataset, consecutively.

Table 3. The details of the model evaluation

	Training (n=376)		Testing (n=160)	
Model	correlation coefficient	mean absolute error	correlation coefficient	mean absolute error
ANNs	0.94	7.37	0.95	7.87
SVM	0.88	14.18	0.90	14.71

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

Arslan, A.K., Colak, C. & Sarihan, M.E. (2016). Different medical data mining approaches based prediction of ischemic stroke. Computer Methods and Programs in Biomedicine, **130**:87-92.

Colak, C., Karaman, E. & Turtay, M.G. (2015). Application of knowledge discovery process on the prediction of stroke. Computer Methods and Programs in Biomedicine, **119**(3):181-185.

Colak, M.C., Colak, C., Kocatürk, H., Sağiroğlu, S. & Barutçu, I. (2008). Predicting coronary artery disease using different artificial neural network models. Anadolu kardiyoloji dergisi: AKD= the Anatolian Journal of Cardiology, **8**(4):249-254.

Fayyad, U., Piatetsky-Shapiro, G. & Smyth, P. (1996). From data mining to knowledge discovery in databases. AI Magazine, **17**(3):37.

Hongfei, W., Yunyan, Z., Fei, Y. & Hui, L. (2013). Evaluation of an artificial neural network to ascertain why there is a high incidence of hepatitis B in the Chinese population after vaccination. Computers in Biology and Medicine, **43**(9):1167-1170.

Investigators, L.R.C. (1992). The lipid research clinics coronary primary prevention trial: results of 6 years of post-trial follow-up. Archives of Internal Medicine, **152**(7):1399.

Moses, D. (2015). A survey of data mining algorithms used in cardiovascular disease diagnosis from multi-lead ECG data. Kuwait Journal of Science, **42**(2):206-235.

Shin, J., Park, H., Cho, S., Nam, H. & Lee, K.J. (2014). A correction method using a support vector machine to minimize hematocrit interference in blood glucose measurements. Computers in Biology and Medicine, **52**(0):111-118.

Silwattananusarn, T. & Tuamsuk, K. (2012). Data mining and its applications for knowledge management: A literature review from 2007 to 2012. International Journal of Data Mining & Knowledge Management Process, **2**(5):13-24.

Stamler, J., Wentworth, D. & Neaton, J.D. (1986). Is relationship between serum cholesterol and risk of premature death from coronary heart disease continuous and graded?: Findings in 356 222 primary screenees of the multiple risk factor intervention trial (MRFIT). Jama, **256**(20):2823-2828.

Süt, N. & Çelik, Y. (2012). Prediction of mortality in stroke patients using multilayer perceptron neural networks. Turkish Journal of Medical Sciences, **42**(5):886-893.

Thygesen, K., Alpert, J.S., Jaffe, A.S., White, H.D., Simoons, M.L., *et al.*(2012). Third universal definition of myocardial infarction. Journal of the American College of Cardiology, **60**(16):1581-1598.

Thygesen, K., Alpert, J.S. & White, H.D. (2007a). Universal definition of myocardial infarction. European Heart Journal, **28**(20):2525-2538.

Thygesen, K., Alpert, J.S., White, H.D., Jaffe, A.S., Apple, F.S., *et al.* (2007b). Universal definition of myocardial infarction. Circulation, **116**(22):2634-2653.

Wang, S.J., Ohno-Machado, L., Fraser, H.S.F. & Kennedy, R.L. (2001). Using patient-reportable clinical history factors to predict myocardial infarction. Computers in Biology and Medicine, **31**(1):1-13.

Wilson, P.W., D'Agostino, R.B., Levy, D., Belanger, A.M., Silbershatz, H., *et al.* (1998). Prediction of coronary heart disease using risk factor categories. Circulation, **97**(18):1837-1847.

Yu, W., Liu, T., Valdez, R., Gwinn, M. & Khoury, M.J. (2010). Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes. BMC Medical Informatics and Decision Making, **10**(1):16.

Yusuf, S., Hawken, S., Ôunpuu, S., Dans, T., Avezum, A., *et al.* (2004). Effect of potentially modifiable risk factors associated with myocardial infarction in 52 countries (the INTERHEART study): case-control study. The Lancet, **364**(9438):937-952.

Zhou, S., Li, G.B., Huang, L.Y., Xie, H.Z., Zhao, Y.L., *et al.* (2014). A prediction model of drug-induced ototoxicity developed by an optimal support vector machine (SVM) method. Computers in Biology and Medicine, **51**:122-127.

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 تلغيم البيانات الطبية

[1]سيميل كولاك، [2]مهمت سينجز كولاك، [3]نيسيب إرميس، [2]نيفزات إرديل، [3]رمضان أوزدمير
[1]جامعة إينونو – كلية الطب – قسم الإحصاء الحيوي والمعلومات الطبية – مالاتيا – تركيا
[2]جامعة إينونو – كلية الطب – قسم جراحة القلب والأوعية الدموية – مالاتيا – تركيا
[3]جامعة إينونو – كلية الطب – قسم طب القلب – 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 تلغيم البيانات، والشبكات العصبية الاصطناعية ونماذج 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 بأثر رجعي بما في ذلك الجنس والعمر والوزن والطول، والنبض، والجلوكوز، والكرياتينين، والدهون الثلاثية، البروتين الدهني عالي الكثافة، والبروتين الدهني منخفض الكثافة للتنبؤ بمستوى الكوليسترول في الدم. وتم اختيار الميزة على اساس الإرتباط واستخدام المستقبلات متعددة الطبقات (MLP) في الشبكات العصبية الصناعية و SVM. وقد تم تقييم أداء الشبكات العصبية الصناعية والنماذج على أساس معامل الارتباط ومتوسط الخطأ المطلق. وتم ملاحظة معاملات الارتباط المقدرة وكانت القيم المتوقعة 0.94 للشبكات العصبية الصناعية و 0.88 لSVM في تدريب مجموعة البيانات (ن = 376)، و 0.95 للشبكات العصبية الصناعية و 0.90 لSVM في اختبار مجموعة بيانات (ن = 160)، على التوالي. أسفرت الشبكات العصبية الصناعية والنماذج SVM عن قيم لمتوسط الخطأ المطلق من 7.37 و 14.18 في بيانات التدريب، و 7.87 و 14.71 في اختبار مجموعة البيانات، على التوالي. أظهرت نتائج تقييم الأداء أن الشبكات العصبية الصناعية MLP لها أداء أفضل للتنبؤ بمستوى الكوليسترول في الدم في المرضى الذين يعانون MI بالمقارنة مع SVM. من الممكن استخدام نموذج الشبكات العصبية الصناعية المقترحة للتنبؤ بمستوى الكوليسترول في الدم لمرضى MI في عملية دعم اتخاذ القرارات السريرية.

الكلمات المفتاحية:

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