Original Research *doi: 10.5455/medscience.2015.04.8274*

Application of Medical Data Mining on the Prediction of APACHE II Score

Cemil Colak[1], Mustafa Said Aydogan[2], Ahmet Kadir Arslan[1], Aytac Yucel[2]

[1] Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey

[2] Inonu University, Faculty of Medicine, Department of Anesthesia and Reanimation, Malatya, Turkey

Abstract

The Acute Physiology and Chronic Health Evaluation (APACHE II) is a beneficial tool for the estimation of risk and the comparison of the patients who received care with similar risk properties. Machine learning based systems can assist clinicians in the early diagnosis of diseases. This research aimed at predicting the APACHE II score using Support Vector Machine (SVM) from Medical Data Mining (MDM). The records of 280 patients from intensive care unit included the dataset containing the target variable (the APACHE II score), and 23 demographical/clinical predictor variables. Genetic algorithm based feature selection and 10-fold cross validation method were employed. SVM with radial basis (RBF) was constructed. The performance of the proposed approach was assessed using root mean squared error (RMSE), mean absolute error (MAE), correlation (R) and coefficient of determination (R^2). Mean age of the individuals was 51±23 years. 153 (54.6%) were females, and 127 (45.4%) were males. The proposed approach yielded the values of 1.037 for RMSE, 0.727 for MAE, 0.993 for R and 0.986 for R^2, respectively. The results demonstrated that the proposed approach had an excellent predictive performance of the APACHE II score. Additionally, ensemble approaches such as bagging, boosting, voting etc. can improve markedly the performance of the prediction and classification tasks.

Keywords: APACHE II, Medical Data Mining, Support Vector Machines (SVM)

(Rec.Date: Mar 13, 2015 Accept Date: Apr 02, 2015)

Corresponding Author*:* Cemil Colak, Inonu University, Faculty of Medicine, Department of Biostatistics and Medical Informatics, Malatya, Turkey
E-mail: cemilcolak@yahoo.com **Phone:** +90 505 887 04 98 **Fax**: +90 422 3410036

Introduction

The Acute Physiology and Chronic Health Evaluation (APACHE II) is a score for severity and mortality calculation tool created from a vast sample of intensive care unit (ICU) patients in the United States. The APACHE II score includes 12 physiological and 2 sickness-related variables. The "worst" evaluation was characterized as the measure that related to the largest number of points. APACHE II is a beneficial tool for the estimation of risk and the comparison of the patients who received care with similar risk properties in diverse units [1,2].

Knowledge Discovery (KD) is a process of exploring unknown and possibly interesting patterns in huge databases. Since the amount of accessible data increases day after day, important information from the databases can be obtained in order for improving the performance of prediction and classification [3].

Contemporary medicine produces a lot of information stored in the databases. Clinical decision-making for diagnosing and treating diseases from the medical databases is so important. Medical data mining (MDM) may handle these problems, and also can enhance the administration standard for hospital information [4].

MDM can be used to investigate hidden patterns in the medical data sets. In medicine, these patterns may be used for preliminary diagnosis. On the other hand, the accessible unprocessed medical data are mostly dispersed, inhomogeneous, and tremendous. Such data requires to be gathered in a systematic form. The gathered data sets may be then incorporated to structure a clinic data framework [5].

MDM based systems can support clinicians in the early diagnosis of diseases. Based on the concept, this research aimed at predicting the APACHE II score using Support Vector Machine (SVM) from MDM.

Materials and Methods

Dataset

The dataset planned for this study were collected from the ICU, Turgut Ozal Medicine Center, Inonu University, Malatya, Turkey. The dataset included the records of 280 ICU patients. The

target variable was APACHE II score, and the demographical and clinical predictor variables were age (years), iron (FE), ferritin, the iron-binding capacity (IBC), hospital length of stay (days), length of stay in ICU (days), C-reactive protein (CRP), hemoglobine (HB), hematocrit (HCT), maximum voluntary contraction (MVC), mean cell hemoglobin (MCH), platelets, prothrombin time (PT), international normalized ratio (INR), the partial thromboplastin time (PTT), fibrinogen, glucose, blood urea nitrogen (BUN), creatinine, aspartate transaminase (AST), alanine transaminase (ALT), sodium (NA) and potassium (K), respectively.

Data Preprocessing

The data preprocessing formed as a sub-processes contained multiple operations. The details of these operations were given below:

- ✓ *Data selection:* In this step, suitable data for the analysis were determined and selected randomly from the Turgut Ozal Medicine Center database.
- ✓ *Replacing missing value:* In this step, if there are missing values, these are replaced by a value of the related variables such as mean, minimum, maximum, etc.
- ✓ *Outlier detection:* Outliers are values which excessively differ from other values of the dataset. In the present study, the outliers were identified by the anomaly index values calculated from the group deviation index [6].
- ✓ *Data filtering:* Outliers are filtered and removed in this step.
- ✓ *Normalization:* Normalization is a method utilized for rescaling values of dataset to fit in a range [7]. There are assorted normalization methods in Statistics. In the present study, Z-transformation also known as statistical normalization method was utilized. In the Z-transformation method is intended to be transformed to a standard normal distribution, N(0,1) of the dataset [8].
- ✓ *Feature Selection (FS):* FS methodologies plan to choose a little subset of attributes that maximize pertinent to the target and minimize redundancy. FS can improve learning performance, reduce the complexity of calculation, construct more suitable models, and diminish obliged capacity [9]. Additionally, FS is a procedure that selecting a subset of attributes therefore, the dimention of the attribute is ideally

diminished by specific assessment standard [10]. There are various FS algorithms. In the present study, genetic algorithm (GA) based FS was used. GA is a heuristical search method imitating the natural evolution process. This strategy is particularly helpful for an investigation of issues being not totally deterministic or having few analogous solutions. Hence, it is critical to have a system, evaluating the quality for a solution. Additionally, this system is important to eliminate a few solutions and to acknowledge an another [11].

At the end of the data preprocessing step, the preprocessed dataset, well-formed, deducible and suitable for modeling, was extracted from raw dataset.

Support Vector Machines

In the modeling step, Support Vector Machines (SVM) [12] classifier was used as machine learning methods. SVM is a beneficial method for classification and regression [13,14]. The major concept of SVM is to elicit a hyperplane by maximizing the margin between two groups [15]. SVM is one of the approaches of supervised classification tasks and is very beneficial owing to the ability of its generalization. Substantially, SVM augments the margin between the training dataset and the decision bound. The subsets of patterns being nearest to the decision bound are referred as support vectors. Different kernel functions such as radial basis, linear, polynomial etc. are built with SVM [16]. The difficulties of non-linear classification may be resolved by kernel functions [17]. Kernel functions might be depicted as a bridge from linearity to non-linearity [18].

Data Mining

In the current study, SVM with radial basis function (RBF) kernel was employed in training and test sets. A grid search was utilized to optimize SVM parameters of the kernel gamma (γ) and complexity constant (C). For tuning these SVM parameters, 10-fold cross validation (CV) was used. RapidMiner Studio 6.3.0 software was used for machine learning and data mining tasks.

Performance Evaluation

In this study, root mean squared error (RMSE), mean absolute error (MAE), correlation (R) and coefficient of determination (R^2) were employed for performance evaluation as defined below:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - x_i)^2}{n}}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\frac{|x_i - y_i|}{x_i}$$

$$R = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{(n-1)\cdot S_x \cdot S_y}$$

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - x_i)^2}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

where x_i is the measured value, y_i is the predicted value, $\bar{x}$ and $\bar{y}$ are the sample means for the values of x_i and y_i, respectively, n is the number of samples, S_x and S_y are the sample standard deviations for the values of x_i and y_i, respectively.

Results

There were a total of 280 individuals at the beginning of the study. Mean age of the individuals was 51±23 years. 153 (54.6%) were females, and 127 (45.4%) were males. The main objective of the present study was to predict the APACHE II score SVM model from MDM. There were no missing values of the attributes in the present study. According to anomaly index mentioned earlier, 13 instances were filtered and removed from the dataset. Subsequently, standardization was applied to the selected attributes. Afterwards, 13 out of 24 attributes were chosen by GA based FS. The selected predictors were age, IBC, hospital length of stay (days), CRP, HB, MVC, MCH, platelets, PTT, glucose, creatinine, NA and K. Then, the SVM was trained on the basis of the selected attributes. When the grid search method was applied for the tuning of the SVM parameters, the kernel gamma (γ) and the complexity constant (C) values were chosen as 0.01 and 10, respectively.

The findings of performance evaluation were presented in Table 1. Based on the selected performance criteria, the values of 0.936 for *RMSE,* 0.437 for *MAE,* 0.992 for *R* and 0.983 for R^2 were calculated from the process.

Table 1. The findings of performance evaluation based on the *RMSE, MAE, R* and R^2

Criteria	SVM Model
RMSE	0.936
MAE	0.437
R	0.992
R^2	0.983

Conclusions

This study presented the prediction of the APACHE II score using SVM model from MDM. The results of the experiments pointed out that SVM model was so efficient in the prediction of the APACHE II score based on the results of the performance evaluation. The coefficient of determination, R^2, reveals how well a model fits to a given dataset and gives some clues on the goodness of fit of a model [19]. When considering the information, the proposed approach yielded a much higher value of R^2 (0.983).

Relatively small sample size for this study may be one of the main limitations. The other limitation might be to disuse other machine learning algorithms in the MDM process. Therefore, to address the limitations, further studies on huge data sets and other predictors not assessed in the current study are essential to achieve more reliable prediction results.

In conclusion, the achieved results demonstrated that the proposed approach had an excellent predictive performance of the APACHE II score. Additionally, ensemble approaches such as boosting, voting and stacking can improve markedly the performance of the prediction and classification tasks.

Acknowledgements

We would like to the RapidMiner Academia Team so much for providing RapidMiner Studio Enterprise free licence key.

References

1. Knaus WA, Draper EA, Wagner DP, Zimmerman JE. APACHE II: a severity of disease classification system. Crit Care Med. 1985;13(10):818-29.
2. Darrien J, Kasem H. Minimally invasive endoscopic therapy for the management of Boerhaave's syndrome. Ann R Coll Surg Engl. 2013;95(8):552-6.
3. Relich M, Muszyński W. The use of intelligent systems for planning and scheduling of product development projects. Procedia Computer Science. 2014;35:1586-95.
4. Zhu L, Wu B, Cao C. [Introduction to medical data mining]. Sheng Wu Yi Xue Gong Cheng Xue Za Zhi. 2003;20(3):559-62.
5. Soni J, Ansari U, Sharma D, Soni S. Predictive data mining for medical diagnosis: An overview of heart disease prediction. IJCA. 2011;17(8):43-8.
6. IBM SPSS Modeler 15 Algorithms Guide. 2012:3-7. ftp://public.dhe.ibm.com/software/analytics/spss/documentation/modeler/15.0/en/AlgorithmsGuide.pdf access date 12.02.2015
7. Raza K, Hasan AN. A comprehensive evaluation of machine learning techniques for cancer class prediction based on microarray data. arXiv preprint arXiv:13077050. http://arxiv.org/ftp/arxiv/papers/1307/1307.7050.pdf access date 12.02.2015
8. Akthar F, Hahne C. RapidMiner 5 Operator Reference. Rapid-I GmbH. 2012. https://rapidminer.com/wp-content/uploads/2013/10/RapidMiner_OperatorReference_en.pdf access date 12.02.2015
9. Tang J, Alelyani S, Liu H. Feature selection for classification: a review. data classification: algorithms and applications. In: Aggarwal C, ed, Chapman & Hall/CRC Data Mining and Knowledge Discovery Series. CRC Press. 2014.
10. Yu L, Liu H, eds, Feature selection for high-d imensional data: A fast correlation-based filter solution. Washington: ICML. 2003; 856-63.
11. Janc K, Tarasiuk J, Bonnet A, Lipinski P. Genetic algorithms as a useful tool for trabecular and cortical bone segmentation. Comput Methods Programs Biomed. 2013;111(1):72-83.
12. Vapnik VN, Vapnik V. Statistical learning theory: Wiley, New York, 1998.
13. Hsu C-W, Chang C-C, Lin C-J. A practical guide to support vector classification. 2003. http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf access date 12.02.2015
14. Shih FY, Zhang K. Support vector machine networks for multi-class classification. International Journal of Pattern Recognition and Artificial Intelligence. 2005;19(06):775-86.
15. Song Q, Hu W, Xie W. Robust support vector machine with bullet hole image classification. Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on. 2002;32(4):440-8.
16. Acır N, Güzeliş C. Automatic spike detection in EEG by a two-stage procedure based on support vector machines. Comput Biol Med. 2004;34(7):561-75.

17. Zararsiz G, Elmali F, Ozturk A. Bagging support vector machines for leukemia classification. Development. 2012. http://biorxiv.org/content/biorxiv/early/2014/07/28/007526.full.pdf access date 12.02.2015
18. Liu L, Shen B, Wang X. Research on kernel function of support vector machine. In: Advanced Technologies, Embedded and Multimedia for Human-centric Computing. Hannover: Springer. 2014;827-34.
19. Cameron AC, Windmeijer FA. An R-squared measure of goodness of fit for some common nonlinear regression models. Journal of Econometrics. 1997;77(2):329-42.