Biomedical Research 2017; 28 (4): 1553-1556

ISSN 0970-938X
www.biomedres.info

An intelligent system for the classification of postoperative pleural effusion between 4 and 30 days using medical knowledge discovery.

Emek Guldogan[1], Ahmet Kadir Arslan[1*], M. Cengiz Colak[2], Cemil Colak[1], Nevzat Erdil[2]

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

[2]Department of Cardiovascular Surgery, Faculty of Medicine, Inonu University, Malatya, Turkey

Abstract

Objective: Pleural Effusion (PE) is a considerable and a common health problem. The classification of this condition is of great importance in terms of clinical decision making. The purpose of the study is to design an intelligent system for the classification of postoperative pleural effusion between 4 and 30 days after surgery by medical knowledge discovery (MKD) methods.

Materials and methods: This study included 2309 individuals diagnosed with coronary artery disease for elective coronary artery bypass grafting (CABG) operation. The results of chest x-ray were used to diagnose PE. The subjects were allocated to two groups: PE group (n=81) and non-PE group (n=2228), consecutively. In the preprocessing step, outlier analysis, data transformation and feature selection processes were performed. In the data mining step, Naïve Bayes, Bayesian network and Random Forest algorithms were utilized. Accuracy and area under receiver operating characteristics (ROC) curve (AUC) were calculated as evaluation metrics.

Results: In the preprocessing step, 85 outlier observations were removed from the study. The rest of the data consisted of 2224 subjects: 2149 of these individuals were in non-PE group, and the 75 were in PE group. Random Forest yielded the best classification performance with 97.45% of accuracy and 0.990 of AUC for 0.7 of the optimal split ratio by Grid search algorithm.

Conclusion: The achieved results pointed out that the best classification performance was obtained from the RF ensemble model. Therefore, the suggested intelligent system can be used as a clinical decision making tool.

Keywords: Bayesian network, Naïve bayes, Pleural effusion, Random forest, Risk factors.

Accepted on August 11, 2016

Introduction

Pleural Effusion (PE) occurs as a result of the deterioration of the balance of absorption and secretion in the pleura [1]. PE is a considerable and common health problem; nonetheless the exact pathogenesis for the accumulation of pleural fluid has not been fully explained [2,3]. Many local and systemic diseases can cause pleural effusion [4]. Knowledge discovery process (KDP) is exploring latent attributes and patterns from the enormous and complicated datasets [5]. KDP is the entire process of discovering beneficial knowledge from the dataset(s) while data mining (DM) is a specific step in the process [6,7]. In medicine, medical knowledge discovery (MKD) covers to identify the optimal determinations to consider different medical conditions [8]. The split-validation (SV) or holdout technique splits dataset into training and testing sets [9]. The dataset is divided by a specified ratio and the classification model is trained in training part and tested in the test set [10,11]. The purpose of the study is to design an intelligent system for the classification of postoperative pleural effusion between 4 and 30 days after surgery by medical knowledge discovery (MKD) methods.

Material and Methods

Dataset

This study was carried out as retrospective case control design in the cardiovascular surgery department, School of Medicine at Inonu University, Malatya, Turkey. This study included 2309 individuals diagnosed with coronary artery disease for elective coronary artery bypass grafting (CABG) operation. The results of chest x-ray were used to diagnose PE. The primary output variable of this research is the absence or presence of post-operative PE between 4 and 30 days. The subjects were allocated to two groups: PE group (n=81) and non-PE group (n=2228), consecutively. Power analysis suggested a minimum total of 848 individuals with the rate difference of 0.03, Type I error (α) of 0.05 and Type II error (β) of 0.20. However, this study included a total of 2309 individuals. The summary information of the attributes considered in the present study was given in Table 1.

Table 1. *Summary information of the attributes.*

Attributes	Abbreviation	Attribute type	Definition	Role
Pleural effusion at 4 and 30 days	PE	Categorical	Present/absent	Target
Atrial fibrillation	AF	Categorical	Present/absent	Input
Age (year)	-	Numerical	Natural number	Input
Gender	-	Categorical	Female/male	Input
Smoking	-	Categorical	Yes/no	Input
Diabetes mellitus	DM	Categorical	Present/absent	Input
Hypertension	HT	Categorical	Present/absent	Input
Obesity	-	Categorical	Present/absent	Input
Body mass index (kg/m^2)	BMI	Numerical	Positive real number	Input
Family history	FH	Categorical	Present/absent	Input
Chronic obstructive pulmonary disease	COPD	Categorical	Present/absent	Input
Myocardial infarction	MI	Categorical	Present/absent	Input
Renal dysfunction	RD	Categorical	Present/absent	Input
Past cryoglobulinemia vasculitis	PCV	Categorical	Present/absent	Input
Carotid stenosis	CS	Categorical	Present/absent	Input
The left main coronary artery	LMCA	Categorical	Present/absent	Input
Aneurysmectomy	-	Categorical	Present/absent	Input
Duration of stay in intensive care (days)	DSIC	Numerical	Positive integer	Input
Ventilation time (hours)	VT	Numerical	Positive integer	Input
Length of hospital stay (days)	LHS	Numerical	Positive integer	Input

Data preprocessing

In the study, there was no missing value, so the preprocessing step started with outlier analysis. For detecting outliers, local density cluster-based outlier factor (LDCOF) [10] technique was used and the kernel based k-means was applied as clustering algorithm. In this technique, an outlier factor is assigned for each example and the outlier example(s) was/were determined according to this factor. Secondly, numeric values were normalized. In this study, standardization method was used among the various normalization techniques. Finally, the third step was formed by feature/variable selection (FS). In this step, genetic algorithm (GA) based FS method was utilized. In addition, NB classifier was used as learning algorithm for FS. According to Zhang and Gao, NB is immensely sensible to FS so that NB advances FS performance [12].

Data mining

Naïve bayes: NB is considered to be a Bayesian supervised model that has been employed in clinical applications [13]. NB is of excellent predictive results in the classification problems and is frequently taken into account as a reference approach [14,15]. The NB model can stochastically estimate the class of a hidden pattern by the existing training set to estimate the most possible outcome [16]. In the current study, PE between 4 and 30 days was classified by using NB. In the implementation of NB, Laplace correction was used to preclude high impact of zero possibilities [17].

Bayesian network: Bayesian Network (BN) describes as probabilistic graphical model that points out the relationship between attributes [18]. BN is a strong instrument in the representation of knowledge and appropriate for the MKD procedures with uncertainty [19]. Thence, BN has been successfully implemented in many clinical problems [20]. In this study, BN was constructed for classifying PE between 4 and 30 days.

Random forest: Random Forest (RF), presented by Breiman [21], is a well-known technique for classification and regression problems [22]. The RF technique utilizes and aggregates results of composition of classification and regression tree that formed using a few bootstrap samples of dataset [23]. In the present study, RF was built for the classification of PE between 4 and 30 days. In the application of RF, the parameters were 10 for the number of trees, 20 for minimal depth and 0.25 for confidence level.

Validation and optimization: Holdout (split) validation approach was used for assessing the predictive results of the constructed models [24]. The possible ranges for determining the optimal ratios for each model varied from 0.50 to 0.90 by 0.10 increments. In the current study, the grid search algorithm was utilized to tune the optimal ratios for split validation in order for achieving the best evaluation metrics [25].

Performance evaluation

In the study, accuracy and area under Receiver Operating Characteristics (ROC) curve (AUC) were calculated to evaluate performance of the constructed models for the classification of the target.

Results

In the preprocessing step, 85 outlier observations were removed from the study. The rest of the data consisted of 2224 subjects: 2149 of these individuals were in non-PE group, and the 75 were in PE group. The mean ages of PE and non-PE groups were calculated 63.13 ± 8.51 and 61.40 ± 9.19, respectively. While 16 (21.3%) in PE group and 524 (24.4%) in non-PE group were females, 59 (78.7%) in PE group and 1625 (75.6%) in non-PE group were males. The chosen attributes after implementing FS were presented in Table 2. The results of accuracy and AUC for optimal ratios determined by Grid search algorithm were given in Table 3 according to the examined models.

Table 2. *The chosen attributes after FS.*

Attributes Number	Attributes
1	Age
2	Body mass index
3	Smoking
4	Diabetes mellitus
5	Hypertension
6	Obesity
7	Family history
8	Myocardial infarction
9	Past cryoglobulinemia vasculitis
10	Carotid stenosis
11	The left main coronary artery
12	Aneurysmectomy

Table 3. *The results of accuracy and AUC for optimal ratios determined by Grid search algorithm according to the examined models.*

Model	Optimal number of split ratio	Accuracy (%)	AUC
NB	0.9	97.75%	0.689
BN	0.8	97.08%	0.618
RF	0.7	97.45%	0.990

Conclusions

In the current study, an intelligent system was constructed for the classification of postoperative pleural effusion between 4 and 30 days after surgery by Medical Knowledge Discovery (MKD) methods. In this context, we built three MKD approaches, NB, BN and RF. For the determination of optimal split ratio, grid search was utilized for each model. According to findings of grid search technique, RF yielded 0.7 of the optimal ratio with 97.45% of accuracy and 0.990 of AUC. When AUC and accuracy were considered, RF produced remarkable classification performance as compared to NB and BN. Since the RF is an ensemble learning algorithm, obtaining higher predictive results from RF may be attributed to the important property of ensemble learning.

In summary, the achieved results pointed out that the best classification performance was obtained from the RF ensemble model. Therefore, the suggested intelligent system can be used as a clinical decision making tool.

Acknowledgement

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

References

1. Batırel FH, Yüksel M. Plevral Efüzyona Yaklaşım: Cerrahi Perspektif. Türk Toraks Dergisi. 2002; 3: 13-19.
2. Momi H, Matsuyama W, Inoue K, Kawabata M, Arimura K, Fukunaga H. Vascular endothelial growth factor and proinflammatory cytokines in pleural effusions. Respirat Med 2002; 96: 817-822.
3. Putnam JB, Jr. Malignant pleural effusions. Surg Clin 2002; 82: 867-883.
4. Gönlügür TE, Gönlügür U. 454 plevral efüzyonun retrospektif analizi. İnönü Üniv Tıp Fak Derg 2007; 14: 21-25.
5. Khan DM, Mohamudally N, Babajee D. A unified theoretical framework for data mining. Proced Comput Sci 2013;17: 104-113.
6. Arslan AK, Colak C, Sarihan ME. Different medical data mining approaches based prediction of ischemic stroke. Comput Method Program Biomed 2016;130: 87-92.
7. Colak C, Karaman E, Turtay MG. Application of knowledge discovery process on the prediction of stroke. Comput Method Program Biomed 2015; 119: 181-185.

8. Roddick JF, Fule P, Graco WJ. Exploratory medical knowledge discovery: Experiences and issues. ACM SIGKDD Explor Newslett 2003; 5: 94-99.
9. Dawson CW, Abrahart RJ, Shamseldin AY, Wilby RL. Flood estimation at ungauged sites using artificial neural networks. J Hydrol 2006; 319: 391-409.
10. Hofmann M, Klinkenberg R. RapidMiner: Data mining use cases and business analytics applications: CRC Press; 2013.
11. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI 1995.
12. Zhang W, Gao F. An improvement to naive bayes for text classification. Procedia Eng 2011; 15: 2160-2164.
13. Nordyke RA, Kulikowski CA, Kulikowski CW. A comparison of methods for the automated diagnosis of thyroid dysfunction. Comput Biomed Res 1971; 4: 374-389.
14. Miasnikof P, Giannakeas V, Gomes M, Aleksandrowicz L, Shestopaloff AY, Alam D. Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. BMC Med 2015; 13: 1.
15. Berchialla P, Foltran F, Gregori D. Naïve Bayes classifiers with feature selection to predict hospitalization and complications due to objects swallowing and ingestion among European children. Safety Sci 2013; 51: 1-5.
16. Setsirichok D, Piroonratana T, Wongseree W, Usavanarong T, Paulkhaolarn N, Kanjanakorn C. Classification of complete blood count and haemoglobin typing data by a C4. 5 decision tree, a naïve Bayes classifier and a multilayer perceptron for thalassaemia screening. Biomed Signal Process Contrl 2012; 7: 202-212.
17. Colak M, Colak C, Erdil N, Arslan A. Investigating Optimal Number of Cross Validation on the Prediction of Postoperative Atrial Fibrillation by Voting Ensemble Strategy. Turkiye Klinikleri J Biostat 2016; 8: 30-35.
18. Zhu Y, Liu D, Jia H. A new evolutionary computation based approach for learning Bayesian network. Procedia Eng 2011; 15: 4026-4030.
19. Liu Z, Liu Y, Cai B, Zheng C. An approach for developing diagnostic Bayesian network based on operation procedures. Expert System Appl 2015; 42: 1917-1926.
20. López-Cruz PL, Larrañaga P, DeFelipe J, Bielza C. Bayesian network modeling of the consensus between experts: An application to neuron classification. Int J Approximate Reasoning 2014; 55: 3-22.
21. Breiman L. Random Forests. Machine Learning 2001; 45: 5-32.
22. Song J. Bias corrections for Random Forest in regression using residual rotation. J Korean Stat Soc 2015.
23. Díaz-Uriarte R, De Andres SA. Gene selection and classification of microarray data using random forest. BMC Bioinforma 2006; 7: 3.
24. Ebbes P, Papies D, Van Heerde HJ. The sense and non-sense of holdout sample validation in the presence of endogeneity. Market Sci 2011; 30: 1115-1122.
25. LaValle SM, Branicky MS, Lindemann SR. On the relationship between classical grid search and probabilistic roadmaps. Int J Robot Res 2004; 23: 673-692.

*Correspondence to

Ahmet Kadir Arslan

Department of Biostatistics and Medical Informatics

Faculty of Medicine

Inonu University

Turkey