Investigating Optimal Number of Cross Validation on the Prediction of Postoperative Atrial Fibrillation by Voting Ensemble Strategy

Voting Ensemble Stratejisi ile Postoperatif Atriyal Fibrilasyonun Tahmininde Optimal Çapraz Geçerlilik Sayısının Belirlenmesi

M. Cengiz ÇOLAK,[a]
Cemil ÇOLAK,[b]
Nevzat ERDİL,[a]
Ahmet Kadir ARSLAN[b]

Departments of
[a]Cardiovascular Surgery, Malatya
[b]Biostatistics and Medical Informatics,
İnönü University Faculty of Medicine,
Malatya

Geliş Tarihi/*Received:* 23.01.2016
Kabul Tarihi/*Accepted:* 05.02.2016

Yazışma Adresi/*Correspondence:*
Ahmet Kadir ARSLAN
Inonu University Faculty of Medicine,
Department of Biostatistics and
Medical Informatics, Malatya,
TÜRKİYE/TURKEY
ahmetkadirarslan@gmail.com

ABSTRACT Objective: Atrial fibrillation (AF) is the most widely recognized cardiovascular arrhythmia. Symptomatic AF can decrease quality of life, functional condition, and performance of the heart. In order to make a clinical decision strategy, the present study attempts to investigate the optimal number of cross validation (CV) [NOCV] on the prediction of postoperative AF based on the several data mining and voting ensemble approaches. **Material and Methods:** The retrospective dataset included complete medical records of 2888 individuals after coronary artery bypass grafting. The subjects were divided into two groups: AF group (n=360) and non-AF group (n=2528), respectively. Data mining approaches including artificial neural networks (ANN), Naïve Bayes (NB) and logistic regression (LR) were constructed for the prediction of the presence or absence of AF. Additionally, voting ensemble strategy was employed in order to improve predictive accuracy. NOCV was optimized by Grid search. For evaluating the predictive performance, accuracy and area under curve (AUC) of the receiver operating characteristics (ROC) graph were considered as evaluation index. **Results:** After removing the missing values and outliers, this research consisted of 2694 subjects; 327 (12.1%) in AF group and 2367 (87.9%) in non-AF group, respectively. The largest accuracy of 87.8% for LR was observed in the 8-fold CV. Similarly, voting ensemble yielded the highest AUC value of 0.896 in the 10-fold CV. The highest accuracy and AUC were 87.2% and 0.896 for voting, 87.8% and 0.729 for LR, 87.1% and 0.668 for ANN, and 79.4% and 0.675 for NB, consecutively. **Conclusion:** The results of the current research demonstrated that the constructed models yielded different results for predicting postoperative AF in the determination of the optimal NOCV. It is recommended to determine the optimal NOCV by optimization techniques. The proposed voting was an acceptable and promising approach for predicting AF, and thence, can support clinicians in the clinical decision making.

Key Words: Atrial fibrillation; data mining methods; number of cross validation; parameter optimization, voting ensemble

ÖZET Amaç: Atriyal fibrilasyon (AF) en sık rastlanan kardiyovasküler bir aritmi olup, kalbin yaşam kalitesini, fonksiyonel durumunu ve performansını düşürebilir. Bu çalışmada çeşitli veri madenciliği ve voting ensemble yaklaşımlarına dayalı olarak postoperatif AF'nin tahmininde optimal çapraz geçerlilik (ÇG) sayısının [OÇGS] belirlenmesi amaçlanmıştır. **Gereç ve Yöntemler:** Koroner arter baypas cerrahisi sonrası 2888 bireyin geriye dönük olarak elde edilen veri seti, AF grubu (n = 360) ve AF olmayan grup (n = 2528) olarak ikiye ayrıldı. Yapay sinir ağları (YSA), Naive Bayes (NB), lojistik regresyon (LR) ve voting ensemble (VE) yaklaşımları AF'nin tahmininde kullanıldı. OÇGS'nin belirlenmesinde grid search algoritması kullanıldı. İşlem karakteristik eğrisi altında kalan alan (AUC) ve doğruluk değerleri, model performanslarını incelemek için dikkate alındı. **Bulgular:** Eksik ve aykırı değerler çıkarıldıktan sonra, bu araştırmada; 327 (% 12.1) birey AF grubunda ve 2367 (87.9%) birey AF olmayan grupta olmak üzere toplam 2694 kişi yer almaktaydı. Grid search optimizasyonu sonucunda; en yüksek doğruluk değeri % 87.8 olarak LR için 8-kat ÇG'de ve en yüksek AUC değeri 0.896 olarak VE için 10-kat ÇG'de gözlenmiştir. Ayrıca, en yüksek doğruluk ve AUC değerleri sırasıyla; VE için % 87.2 ve 0.896, LR için % 87.8 ve 0.729, YSA için % 87.1 ve 0.668, NB için % 79.4 ve 0.675 idi. **Sonuç:** Elde edilen bulgular, önerilen VE yaklaşımının AF'nin tahminde oldukça başarılı ve ümit verici bir yöntem olduğunu ve böylece klinisyenlere klinik karar vermede önemli destek verebileceğini göstermiştir.

Anahtar Kelimeler: Atriyal fibrilasyon; veri madenciliği yöntemleri; çapraz geçerlilik sayısı; parametre optimizasyonu, voting ensemble

doi: 10.5336/biostatic.2016-50382

Turkiye Klinikleri J Biostat 2016;8(1):30-5

Atrial fibrillation (AF) is the most widely recognized cardiovascular arrhythmia. Symptomatic AF can decrease quality of life, functional condition, and performance of the heart. Postoperative AF is a common complication after cardiac surgery with a rate of 20-50%, and is related with higher medical costs as well as an increased risk of death. Prediction of AF is important to understand medical costs and disability burden in the population related to this disease. Age is a significant factor affecting the prevalence of postoperative AF. While some risk factors associated with AF may be controlled, the others cannot be modified. Progression of the AF may be prevented by controlling the modifiable risk factors.[1,2]

Knowledge Discovery Process (KDP), also known as data mining, is an approach to achieve patterns from vast datasets by incorporating techniques of statistics, mathematics and machine learning. Also, KDP is a very practical approach. It has been shown that iterative approach utilized in KDP for the extraction of the novel knowledge might significantly improve the accuracy of the complete system.[3]

Cross-validation (CV) is a commonly known method for technique selection. The primary concept behind CV is to separate dataset, once or many times, to calculate the error of each technique. A portion of the data (training dataset) is utilized for training each technique, and the remaining portion (validation sample) is employed for calculating the error of the technique. Afterwards, CV chooses the technique with the smallest calculated error.[4] To assess the performance of Knowledge Discovery Process based methods, k-fold CV is generally utilized. Especially, 10-fold CV is considered as the standard method for evaluating the performance of prediction and classification with respect to error rate.[5]

In this context, there has been no research investigating the optimal number of CV (NOCV) on the prediction of postoperative AF by various data mining and voting ensemble approaches. Several single approaches can be combined based on a strategy, which generated a single approach called as an ensemble technique. It has been reported that the ensemble strategy can improve the predictive performance and outperforms a single approach.[6]

In the current study, the primary objective is to investigate the optimal NOCV on the prediction of postoperative AF based on the several data mining and voting ensemble approaches. The second objective of the current study is to evaluate the predictive performance of the data mining approaches.

MATERIAL AND METHODS

2.1. DATASET

The retrospective dataset included complete medical records of 2888 subjects after coronary artery bypass grafting (CABG) from the cardiovascular surgery department of Turgut Ozal Medical Center of Inonu University, Malatya, Turkey. The subjects were divided into two groups: AF group (*n*=360) and non-AF group (*n*=2528), respectively. Electrocardiogram was employed for diagnosing postoperative AF. The target of the present research is the presence or absence of postoperative AF. The target and predictor variables including demographic and clinical characteristics among the risk factors for AF were given in detail (Table 1).[7,8]

DATA PREPROCESSING

The data preprocessing involved the stages explained below:

- *Data selection:* The appropriate dataset was determined and collected from the cardiovascular surgery department.

- *Missing values:* Missing values were removed.

- *Nominal to numerical transformation:* The data mining approaches employed in this study assume that the attributes have categorical scale.[9] Therefore, the dataset was discretized in the current study.

- *Outlier detection:* Outliers are excessively different from other values of the dataset. The Class Outlier Factor (COF) operator can be used

TABLE 1: Summary information of the attributes

Attributes	Abbreviation	Attribute type	Definition	Role
Atrial fibrillation	AF	Categorical	Present/absent	Target
Pleural effusion	PE	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
Valve surgery	VS	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
Bleeding revision	BR	Categorical	Present/absent	Input
Pleural effusion at 4 days	PE4	Categorical	Present/absent	Input
Pleural effusion at 7 days	PE7	Categorical	Present/absent	Input
Pleural effusion at 15 days	PE15	Categorical	Present/absent	Input
Pleural effusion at 30 days	PE30	Categorical	Present/absent	Input
Length of hospital stay (days)	LHS	Numerical	Positive integer	Input

to detect outliers. In the present study, the outliers were determined by the COF method.[10]

■ *Data filtering:* Outliers were removed in this step.

■ *Normalization:* Among the normalization methods, the aim of the Z-transformation is to convert a data into standard normal distribution, N(mean=0, variance=1).[11] In this study, Z-transformation also known as the statistical normalization method was utilized.

■ *Feature Selection (FS):* In this study, Support vector machine (SVM) based attribute weighting method was used. The weighted attribute values were normalized and ranked in descending order.

ARTIFICIAL NEURAL NETWORKS

Artificial neural networks (ANN) are one of the important data mining and machine learning approaches and have many useful properties for the prediction and classification problems in medical applications. Different ANN models including multilayer perceptrons (MLP), radial basis function network, etc. are capable of predicting the desired target(s) owing to their adaptive and flexible structure [12]. In the current study, a feed-forward ANN trained with a back propagation algorithm (MLP) was constructed for predicting the presence or absence of AF. The input layer enclosed the selected attributes according to the FS method, and the output layer included the target attribute, AF. The ANN model had 14 neurons in the hidden layer, and the sigmoid activation function was employed in both of the hidden and output layers. The learning rate and momentum for the generated ANN were 0.3 and 0.2, consecutively.

NAÏVE BAYES

Naïve Bayes (NB) methods implement the theorem of Bayes to estimate the most possible target distribution based on the values of attributes (features). A Naive Bayes classifier is so fast for the calculation since it usually learns by seeking each attribute space and related target once and classifies by regarding each target and attribute-value pair once.[13] Naïve Bayes approach is a very efficient model for data mining owing to the supposition which all attributes are independent.[14,15] In the current study, in order for the prediction of AF, NB with Laplace correction was built based on the chosen predictor attributes. Laplace correction was utilized to avoid high influence of zero probabilities.[16]

LOGISTIC REGRESSION ANALYSIS

Logistic regression (LR) is one of the multivariate statistical methods that can be used in classification tasks. This method is applied when the dependent variable is discrete and the independent variables are both discrete and continuous. In addition, there is no assumption of normal distribution and continuity of preconditions in this technique. The effects of independent variables on the dependent variable is determined as the probability. LR is a method used for determining the association between the dependent and independent variables.[17,18] In this study, LR model was employed in predicting AF based on the selected predictor variables.

VOTING ENSEMBLE

An ensemble-based strategy constructs an arrangement of base approaches from the training data and makes estimations by combining the estimated outcome of each approach. The predictive performance of an ensemble approach is better than any of its base approaches. In addition, an ensemble-based strategy can avoid the overfitting issues and decreases the variability for the predictive errors. The voting chooses the assessed output by incorporating the predictive results of diverse approaches. Firstly, each base approach is constructed based on the training dataset, and each is related to a weight and evaluated by its predictive accuracy. Afterward, a particular combination form is chosen for the ultimate estimation for the output value.[19] In the present study, several base techniques, ANN, NB, and LRA, were considered to construct the voting ensemble strategy to improve the predictive performance of AF. To this end, the predicted target (i.e., the presence/absence of AF) was identified by the combinations of the aforementioned single approaches, and the predictive results were evaluated based on the performance measures.

DATA MINING

The CV method ensures to optimize for the better predictive performance.[20] In addition, in order to achieve the best predictive performance from the generated models, NOCV was optimized using Grid search (GS) technique in this study. GS technique is a conventional method to optimize parameters. Additionally, it is one of the most extensively utilized techniques to optimize hyper-parameter(s).[21,22] For the tuning of NOCV, all integers between 2 and 10 (including 2 and 10) were considered. Each result was evaluated according to the values of accuracy and area under curve (AUC). RapidMiner 6.3 for Windows and IBM SPSS Statistics 22.0 for Windows were used in all analyses.

PERFORMANCE EVALUATION

The receiver operating characteristic (ROC) curve determines the predictive performance of a two class approach for the several threshold settings, and it is formed by plotting one minus the specificity (the false positive rate) in the x-axis against the sensitivity (the true positive rate) in the y-axis. The overall accuracy is calculated by the AUC of the ROC graph. In the medical studies, AUC values >0.70 can be acceptable for the model evaluation.[23] In the present study, for assessing the predictive performance, accuracy and area under the ROC curve were considered as evaluation index.

RESULTS

After removing the missing values and outliers, this research consisted of 2694 subjects; 327 (12.1%) in AF group and 2367 (87.9%) in non-AF group, respectively. Mean ages were 65.2 ± 8.6 for AF group and 60.4 ± 9.8 for non-AF group. While 86 (26.3%) in AF group and 576 (24.3%) in non-AF group were females, 241 (73.7%) in AF group and 1791 (75.7%) in non-AF group were males. In the current research, SVM based feature selection approach was utilized. The selected attributes after applying FS were presented (Table 2).

The results of accuracy and AUC for each optimal number of folds with respect to the models based on Grid search were summarized (Table 3). The largest accuracy of 87.8% for LR was observed in the 8-fold CV. Similarly, voting ensemble yielded the highest AUC value of 0.896 in the 10-fold CV. The highest accuracy and AUC were 87.2% and 0.896 for voting, 87.8% and 0.729 for LR, 87.1% and 0.668 for ANN, and 79.4% and 0.675 for NB, consecutively.

DISCUSSION

The primary objective of the current study was to investigate the optimal NOCV on the prediction of postoperative AF by several data mining approaches using Grid search technique. To accomplish this objective, various base techniques, namely ANN, NB, and LRA, and voting ensemble, were built for the prediction of postoperative AF. When the accuracy was considered in defining the optimal NOCV for all the models, LR had the highest accuracy of 87.8% for 8-fold CV. Similarly, among the optimal number of folds, Voting and ANN had the accuracy values (87.2% and 87.1%) for NOCVs of 10 and 3, respectively. In relation to NB, the accuracy (79.4%) was obtained in the 4-fold CV.

TABLE 2: The selected attributes after FS.

No	Attributes
1	Obesity
2	Age
3	Body mass index
4	Length of hospital stay
5	Duration of stay in intensive care
6	Smoking
7	Carotid stenosis
8	Ventilation time
9	Chronic obstructive pulmonary disease
10	Family history
11	Bleeding revision
12	Hypertension
13	Aneurysmectomy
14	The left main coronary artery

TABLE 3: The results of accuracy and AUC for each optimal number of folds with respect to the models based on Grid search.

Model	Optimal number of folds	Accuracy (%)	AUC
Voting	10	87.2	0.896
LR	8	87.8	0.729
ANN	3	87.1	0.668
NB	4	79.4	0.675

When regarding AUC values for investigating the optimal NOCV, the largest AUC (0.896) was calculated from the voting ensemble method for 10-fold CV. Other NOCV values having high AUC were 8 for LR, 4 for NB and 3 for ANN, respectively.

The second objective of the current study is to evaluate the predictive performance of the data mining approaches in the prediction of postoperative AF. Among the ANN, NB, and LR models, the largest accuracy (87.8%) was achieved from LR. When the results of LR were compared to Voting, LR had slightly higher accuracy; however, the AUC for voting method was considerably higher. The area under the ROC curve is an another indicator used for the performance evaluation [24]. Therefore, the voting ensemble had the best AUC (0.896). LR, NB and ANN models were followed the prediction performance with respect to the AUC values (0.729, 0.675 and 0.668), respectively. Because AUC values higher than 0.70 can be acceptable for model evaluation in the medical studies [23], the generated Voting and LR models may be more useful in predicting postoperative AF due to their AUC above 0.70.

The subjects included in the present study were retrospectively gathered from a medical center. In addition, if the number of subjects was much increased at the multiple medical centres, better prediction results can be obtained from the models. In addition, in future research, other ensemble schemes such as boosting, bagging and stacking, etc. can be tested for achieving more accurate prediction performance.

The results of the current research demonstrated that the constructed models yielded different results for predicting postoperative AF in the determination of the optimal NOCV. Thus, for achieving more definite results, it is recommended to determine the optimal NOCV by optimization techniques, and then, the identified optimal NOCV can be used for the prediction or classification tasks. When considering the predicted results, the voting ensemble method outperformed the other three models (ANN, NB and LR) for the prediction of postoperative AF based on the performance measures. The present prediction results clearly demonstrated that the proposed voting was an acceptable and promising approach for predicting postoperative AF, and thence, can support clinicians in the clinical decision making.

Acknowledgement

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

REFERENCES

1. Jessica YC, Hongshik A, James JC. On sequential closed testing dose groups with a control. Communications in Statistics-Theory and Methods 2000; 29(5-6): 941-56.
2. Go AS, Hylek EM, Phillips KA, Chang Y, Henault LE, Selby JV, et al. Prevalence of diagnosed atrial fibrillation in adults: national implications for rhythm management and stroke prevention: the AnTicoagulation and Risk Factors in Atrial Fibrillation (ATRIA) Study. JAMA 2001;285(18):2370-5.
3. Kurgan LA, Cios KJ, Tadeusiewicz R, Ogiela M, Goodenday LS. Knowledge discovery approach to automated cardiac SPECT diagnosis. Artif Intell Med 2001;23(2):149-69.
4. Arlot S, Celisse A. A survey of cross-validation procedures for model selection. Statistics Surveys 2010;4:40-79.
5. Kim M-Y, Lee DH. Data-mining based SQL injection attack detection using internal query trees. Expert Systems with Applications 2014;41(11):5416-30.
6. Santos-García G, Varela G, Novoa N, Jiménez MF. Prediction of postoperative morbidity after lung resection using an artificial neural network ensemble. Artif Intell Med 2004;30(1):61-9.
7. van der Hooft CS, Heeringa J, Brusselle GG, Hofman A, Witteman JC, Kingma JH, et al. Corticosteroids and the risk of atrial fibrillation. Arch Intern Med 2006;166(9):1016-20.
8. Erdil N, Gedik E, Donmez K, Erdil F, Aldemir M, Battaloglu B, et al. Predictors of postoperative atrial fibrillation after on-pump coronary artery bypass grafting: is duration of mechanical ventilation time a risk factor? Ann Thorac Cardiovasc Surg 2014;20(2):135-42.
9. Zaffalon M, Wesnes K, Petrini O. Reliable diagnoses of dementia by the naive credal classifier inferred from incomplete cognitive data. Artif Intell Med 2003;29(1):61-79.
10. Hewahi NM, Saad MK. Class outliers mining: Distance-based approach. International Journal of Intelligent Technology 2007;2(1):55-68.
11. Akthar F, Hahne C. RapidMiner 5 Operator Reference. 1st ed. Dortmund: Rapid-I GmbH; 2012. p.512-7.
12. Colak C, Karaman E, Turtay MG. Application of knowledge discovery process on the prediction of stroke. Comput Methods Programs Biomed 2015;119(3):181-5.
13. Barrett N, Weber-Jahnke J. A token centric part-of-speech tagger for biomedical text. Artif Intell Med 2014;61(1):11-20.
14. Zhang H. The optimality of naive Bayes. AA 2004;1(2):3.
15. Easton JF, Stephens CR, Angelova M. Risk factors and prediction of very short term versus short/intermediate term post-stroke mortality: a data mining approach. Comput Biol Med 2014;54:199-210.
16. Hofmann M, Klinkenberg R. RapidMiner: Data mining use cases and business analytics applications. 1sted. Boca Raton: CRC Press; 2013. p. 54.
17. Colak C, Colak MC, Orman MN. [The comparison of logistic regression model selection methods for the prediction of coronary artery disease]. Anadolu Kardiyol Derg 2007;7(1):6-11.
18. Samal L, Stavroudis T, Miller R, Lehmann H, Lehmann C. Effect of a laboratory result pager on provider behavior in a neonatal intensive care unit. Appl Clin Inform 2011;2(3):384-94.
19. Hu YH, Wu F, Lo CL, Tai CT. Predicting warfarin dosage from clinical data: a supervised learning approach. Artif Intell Med 2012; 56(1):27-34.
20. Baillie RT, Kapetanios G, Papailias F. Bandwidth selection by cross-validation for forecasting long memory financial time series. Journal of Empirical Finance 2014;29:129-43.
21. Bergstra J, Bengio Y. Random search for hyper-parameter optimization. The Journal of Machine Learning Research 2012;13(1):281-305.
22. Huang CL, Wang CJ. A GA-based feature selection and parameters optimization for support vector machines. Expert Systems with Applications 2006;31(2):231-40.
23. Silva Á, Cortez P, Santos MF, Gomes L, Neves J. Rating organ failure via adverse events using data mining in the intensive care unit. Artif Intell Med 2008;43(3):179-93.
24. Green M, Björk J, Forberg J, Ekelund U, Edenbrandt L, Ohlsson M. Comparison between neural networks and multiple logistic regression to predict acute coronary syndrome in the emergency room. Artif Intell Med 2006;38(3):305-18.