ISSN 2288-4866 (Print)
ISSN 2288-4882 (Online)
http://www.jiisonline.org
J Intell Inform Syst 2019 March: 25(1): 127~137 http://dx.doi.org/10.13088/jiis.2019.25.1.127

A Recidivism Prediction Model Based on XGBoost Considering Asymmetric Error Costs

Ha-Ram Won
Graduate School of Business IT, Kookmin University
(*haramy44@kookmin.ac.kr*)

Jae-Seung Shim
Graduate School of Business IT, Kookmin University
(*simong_kwnau@naver.com*)

Hyunchul Ahn
Graduate School of Business IT, Kookmin University
(*hcahn@kookmin.ac.kr*)

Recidivism prediction has been a subject of constant research by experts since the early 1970s. But it has become more important as committed crimes by recidivist steadily increase. Especially, in the 1990s, after the US and Canada adopted the 'Recidivism Risk Assessment Report' as a decisive criterion during trial and parole screening, research on recidivism prediction became more active. And in the same period, empirical studies on 'Recidivism Factors' were started even at Korea. Even though most recidivism prediction studies have so far focused on factors of recidivism or the accuracy of recidivism prediction, it is important to minimize the prediction misclassification cost, because recidivism prediction has an asymmetric error cost structure.

In general, the cost of misrecognizing people who do not cause recidivism to cause recidivism is lower than the cost of incorrectly classifying people who would cause recidivism. Because the former increases only the additional monitoring costs, while the latter increases the amount of social, and economic costs. Therefore, in this paper, we propose an XGBoost(eXtream Gradient Boosting; XGB) based recidivism prediction model considering asymmetric error cost.

In the first step of the model, XGB, being recognized as high performance ensemble method in the field of data mining, was applied. And the results of XGB were compared with various prediction models such as LOGIT(logistic regression analysis), DT(decision trees), ANN(artificial neural networks), and SVM(support vector machines). In the next step, the threshold is optimized to minimize the total misclassification cost, which is the weighted average of FNE(False Negative Error) and FPE(False Positive Error). To verify the usefulness of the model, the model was applied to a real recidivism prediction dataset. As a result, it was confirmed that the XGB model not only showed better prediction accuracy than other prediction models but also reduced the cost of misclassification most effectively.

Key Words : Recidivism Prediction, Asymmetric Error Cost, Threshold Optimization, Data Mining, XGBoost

Received : January 28, 2019 Revised : March 28, 2019 Accepted : March 28, 2019
Publication Type : Concise Paper Corresponding Author : Hyunchul Ahn

1. Background

The Hollywood movie Minority Report, based on a 1956 short story by Philip K. Dick, shows a program that helps people predict and arrest unsuspecting criminals based on historical data on crime. At the time of publication, crime prediction was like a story in imagination., but recently, the

crime prediction is realizable due to the rapid development of information and communication technologies and the appearance of big data. Because crimes that have already been committed are irreversible, proactive prevention through crime prediction is much more efficient and effective in social aspects than post-counteraction(Jung, 2012).

In the field of data analysis, crime prediction is a very interesting subject in that it takes a scientific approach to a wide variety of data. There are various research fields in crime prediction, but this study focuses on a prediction of criminal recidivism. There is no consistent definition of recidivism, but it is defined as “reengaging in criminal behavior after receiving a sanction or intervention" in general(King and Elderbroom, 2014).

Since the 1990s, when the United States and Canada adopted jurisdiction risk assessment reports as crucial criteria for judges and parolees, the academic and practical interest in ‘recidivism prediction' has increased(Seong, 2006). Thus, to establish a correctional policy based on scientific evidence, and consider reasonable and effective prevention measures of recidivism, various factors have been studied based on the Criminology. Along with this, tools for predicting recidivism have also been developed(Nam and Park, 2011).

According to Seong(2006) at the beginning recidivism prediction was carried out without a scientific basis via the subjective evaluation of experts. From the 1970s, the scale was based on empirical studies, but most of them were static factors. Since the 1990s, the prediction accuracy has been improved by reflecting dynamic factors(Peter and Ann, 1987; Turgut, 2017).

Despite improved performance of recidivism prediction, recidivism is still a big problem in society According to Prison Education News(2014), Recidivism creates another victim, destroys lives of the recidivist's families, and cause the social costs by recidivism. Also, according to a new study by the Pew Charitable Trusts’ Center on the States, 43 percent of prisoners nationally were re-imprisoned within three years. And it is estimated that the 41 states in the U.S. would reap significant savings — $635 million in the first year — if they managed to cut their recidivism rates by just 10 percent(New York Times, 2011).

Therefore, social cost should be taken into consideration in the study of recidivism prediction, and social costs can be taken into consideration by reflecting the following two types of errors (Joo et al., 2003; Lee and Ahn, 2011). First, False-Positive Error (FPE) is misclassified error that a recidivism with a low realizable possibility. The second type, False-Negative Error (FNE), is a misclassification that a recidivism with a higher realizable possibility. In general, the cost of misclassifying noncommissioned person as recidivist is far less than the cost of misclassifying recidivist as non-recidivist. Because the former increases only the additional monitoring costs, whereas the latter causes social and economic huge costs by recidivism. It is very dangerous that the recidivism exists as a cause of potential problem to society at large. In that sense, proactive prevention is a much more efficient and effective method than

post-counteraction. Therefore, this paper investigated crime prediction, which is a proactive response through data analysis, and focused on recidivism prediction among crime predictions. In addition, the study focused on the asymmetric error cost, which is mainly used in the research field for detection model, in view of social cost of recidivism.

In the following sections, the suggested model is presented in Section 2. Empirical analysis including data structures and experimental results is described in Section 3. And the implications, limitation and future research plans are discussed in Section 4.

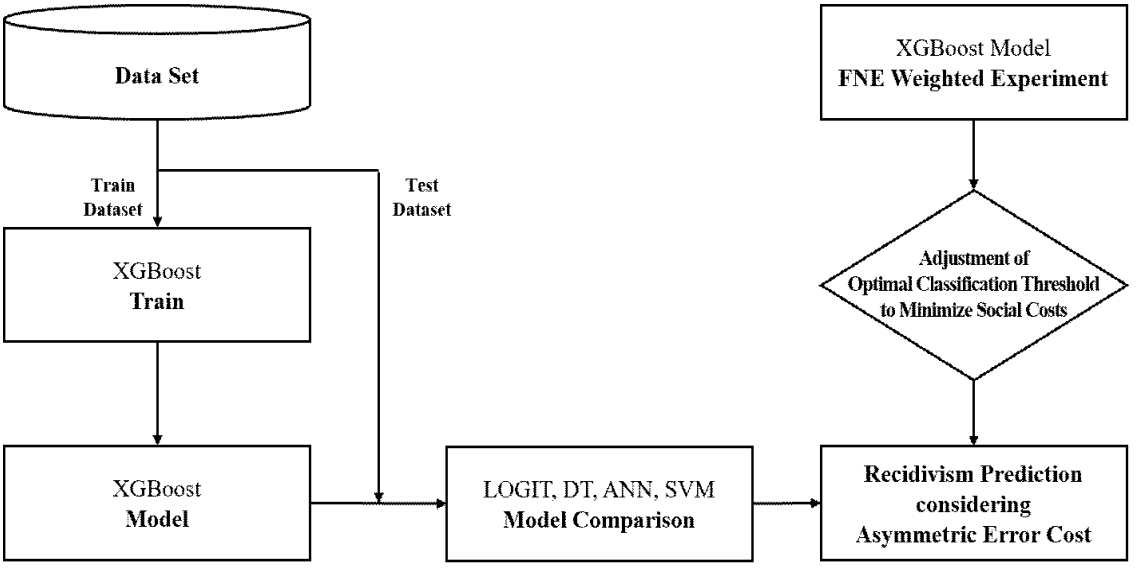
2. Suggested Model

XGBoost used in this paper is ensemble methods based on the idea that we can achieve better performance by combining Weak Learner rather than Single Learner, which is a way to learn several models using the same learning algorithm. And the ensemble method is divided into bagging and boosting. Bagging is a method developed by Breiman, which makes learning algorithms into multiple copies, then learns each of them and combines the results(Breiman, 1994). And, boosting is a method to construct a committee of weak learners that lowers the error rate in classification and prediction error in regression. Boosting works by iteratively constructing weak learners whose training set is conditioned on the performance of the previous members of the ensemble(Sharkey, 1999).

XGBoost, the abbreviated name for “eXtream Gradient Boosting”, which was used in this study, is a decision trees tree algorithm that uses a boosting method to reduce the error value by grouping several CART(classification and regression trees). It creates an optimized model that minimizes the training loss and controls the complexity of the tree to prevent overfitting. This model can solve complex classification problems and prediction problems. Practically, it is widely recognized as high performance ensemble method by many of the machine learning and data mining challenges(Chen and Guestrin, 2016).

<Figure 1> is a proposed model of this study presented in the form of a flow chart. First, a dataset was built up at 1: 1 ratio of recidivist and non-recidivist data. The dataset was classified into training and validation datasets after preprocessing and were apply to XGboost(XGB). To validate XGB performance after application, the results are compared with the statistical model, Logistic Regression Analysis(LOGIT), Decision Trees(DT), Artificial Neural Networks(ANN), and Support Vector Machine(SVM), which are machine learning models.

After comparing the results of XGB, an experiment to reflect the asymmetric error cost to the model is conducted. The method of the experiment is as follows. The usual classification method calculates the result of the model as a form of binary probability value between 0 and 1, and the predicted value is reinterpreted by the classification threshold, and the group is finally



〈Figure 1〉 Flow Chart of the Research Model

determined. In general, '0.5' is used as the classification threshold, but in this case, it is unlikely to be optimal from the viewpoint of the total cost since it basically does not consider differences in costs incurred by FPE and FNE. This means that the optimal classification threshold varies depending on the relative difference between FPE and FNE, and FNE was over-weighted in this experiment because FNE is usually more fatal than FPE. Thus, 10 scenarios are set up that change the weights for FNE from 1 to 10 times the FPE, and for the selected classification model, the optimal FPE and FNE values and the classification threshold to minimize the total cost are found. The calculation formulas of FPE, FNE and total cost used in this study are as follows (1), (2), (3) (Joo et al., 2003; Lee and Ahn, 2011).

$$\mathbf{FPE}(\%) = \frac{The\ number\ of\ misclassifying\ non-recidivists\ as\ recidivists}{The\ number\ of\ total\ non-recidivists} \times 100 \quad (1)$$

$$\mathbf{FNE}(\%) = \frac{The\ number\ of\ misclassifying\ recidivists\ as\ non-recidivists}{The\ number\ of\ total\ recidivists} \times 100 \quad (2)$$

$$\mathbf{Total\ Cost} = \frac{\omega_1 \cdot FPE + \omega_2 \cdot FNE}{\omega_1 + \omega_2} \times 100 \quad (3)$$

3. Empirical Analysis

3.1 Experimental Data

The data used in this project consisted of information from prisoners released from the North Carolina Prison in the United States from July 1, 1978 to June 30, 1979 and were collected at the ICSPR (Inter-university Consortium for Political and Social Research) website. To build the model, a total of 13,002 data were set with 1:1 ratio (6,501:6,501) of the recidivist and non-recidivist.

After preprocessing, 22 variables with three types (population statistical, criminal, and prison-related) were derived as candidates for input variables in <Table 1>. And during the independent samples t-tests and chi-square tests, 17 of the 22 variables were extracted. Finally, we used forward feature selection of logistic regression to select 15 final input variables. The results are presented in the <Table 2>.

〈Table 1〉 Candidate Independent Variables

Variable Type		Variable Name	Description	
Independent variables	Demographic variables	RACE	Black	0: No / 1: Yes
		GENDER	Male or Female	0: Male / 1: Female
		AGE_10	In one's teens(10)	0: No / 1: Yes
		AGE_20	In one's twenties(20)	0: No / 1: Yes
		AGE_30	In one's thirties(30)	0: No / 1: Yes
		AGE_40	In one's forties(40)	0: No / 1: Yes
		AGE_50	In one's fifties(50)	0: No / 1: Yes
		AGE_60	In one's sixties(60)	0: No / 1: Yes
		MARRIED	Married	0: No / 1: Yes
		SCHOOL	Years of formal education	Integer (years)
	Crime-related variables	ALCHY	Alcohol problem	0: No / 1: Yes
		JUNKY	Drug problem	0: No / 1: Yes
		PERSON	Crime on person	0: No / 1: Yes
		PROPTY	Crime on property	0: No / 1: Yes
		FELON	Felony	0: No / 1: Yes
	Imprisonment -related variables	CRIME_EX	Previous imprisonment experience	0: No / 1: Yes
		SUPER	Parole	0: No / 1: Yes
		WORKREL	Participation of Prisoner work release Program	0: No / 1: Yes
		PRIORS	Number of previous imprisonments	Integer (times)
		RULE	Number of violations of discipline during sentence	Integer (times)
		TSERVD	Term of imprisonment	Integer (years)
		FOLLOW	A period of tracked down criminals	Integer (years)
Dependent variable		RECID	Recidivism	0: No / 1: Yes

〈Table 2〉 Selected Independent Variables Applied to the Model

Variable Type		Variable Name	Description	
Independent variables	Demographic variables	GENDER	Male or Female	0: Male / 1: Female
		AGE_10	In one's teens(10)	0: No / 1: Yes
		AGE_30	In one's thirties(30)	0: No / 1: Yes
		AGE_40	In one's forties(40)	0: No / 1: Yes
		AGE_50	In one's fifties(50)	0: No / 1: Yes
		AGE_60	In one's sixties(60)	0: No / 1: Yes
		MARRIED	Married	0: No / 1: Yes
	Crime-related variables	ALCHY	Alcohol problem	0: No / 1: Yes
		PROPTY	Crime on property	0: No / 1: Yes
		FELON	Felony	0: No / 1: Yes
	Imprisonment -related variables	CRIME_EX	Previous imprisonment experience	0: No / 1: Yes
		PRIORS	Number of previous imprisonments	Integer (times)
		RULE	Number of violations of discipline during sentence	Integer (times)
		TSERVD	Term of imprisonment	Integer (years)
		FOLLOW	A period of tracked down criminals	Integer (years)
Dependent variable		RECID	Recidivism	0: No / 1: Yes

3.2 Experimental Results

Using the selected 15 final variables in <Table 2>, we applied XGB model, and compared the results with LOGIT, DT, ANN, SVM models presented above. In <Table 3>, which shows the results of the classification models, the validation data set accuracy of XGB model was the highest at 69.52%.

After this, two-sample test for proportions was performed to determine whether the differences in prediction accuracy between XGB and the other methods are statistically significant. The null hypothesis (H0) for this test is PA = PB, and the alternative hypothesis (Ha) is PA > PB (PA: the average predicted accuracy rate for verification data sets in Model A).

〈Table 3〉 Experimental Results for each Classification Methods

Method	Training	Validation	Parameters
LOGIT	68.81%	67.06%	Forward Selection (Conditional)
DT	68.86%	66.33%	CART
ANN	69.12%	66.28%	H (# of the nodes in the hidden layer) = 10
SVM	72.83%	67.63%	RBF kernel, C = 100, σ2 = 100
XGB	71.93%	69.52%	learning_rate = 0.02, gamma = 5, max_depth = 5, min_child_weight = 6

<Table 4> presents the results of the two-sample test for proportions. As shown in this table, XGB model outperformed LOGIT and SVM at the 5% statistical significance level, and surpassed DT, and ANN at the 1% statistical significance level. Therefore, The XGB model was verified to be the optimal model, and the experiments reflecting the asymmetric error costs were performed using XGB model.

The experiments reflecting the asymmetric error costs were conducted according to the weighting scenarios described in Section 2. For each weighting scenarios, we increased the classification threshold by 0.01 from 0 to 1 in order to search for the optimal threshold that minimizes the total cost. Finally, the total costs threshold for each scenario were compared to the total costs of using the fixed classification threshold (i.e. 0.5). <Table 5> depicts the overall experimental results. The results show that when the classification threshold

〈Table 4〉 Two-Sample Test for Proportions (Z-values)

	DT	ANN	SVM	XGB
LOGIT	0.5589	0.5882	-0.4434	-1.9065**
DT		0.0293	-1.0022*	-2.4650***
ANN			-1.0316*	-2.4943***
SVM				-1.4633**

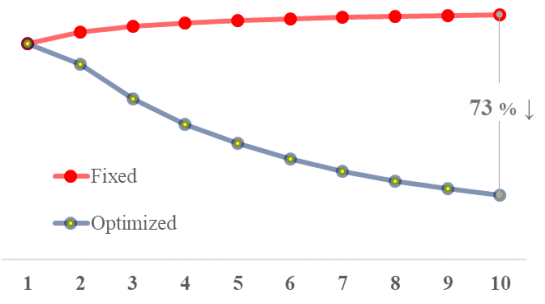
* Statistical significant at 10%, ** Statistical significant at 5%, *** Statistical significant at 1%

〈Table 5〉 Comparison of Results of Fixed and Optimized Classification Threshold

No	w_1	w_2	Fixed Classification Threshold			Optimized Classification Threshold				TC Savings (%)
			FPE(%)	FNE(%)	Total Cost(%)	Threshold(t)	FPE(%)	FNE(%)	Total Cost(%)	
1	1	1	25.60	35.36	30.48	0.50	25.21	36.05	30.48	0.00
2	1	2	25.60	35.36	32.10	0.32	58.42	12.14	27.57	14.11
3	1	3	25.60	35.36	32.92	0.28	68.64	7.30	22.64	31.23
4	1	4	25.60	35.36	33.41	0.20	83.94	2.84	19.06	42.95
5	1	5	25.60	35.36	33.73	0.20	83.94	2.84	16.36	51.50
6	1	6	25.60	35.36	33.96	0.15	94.85	0.77	14.21	58.16
7	1	7	25.60	35.36	34.14	0.11	99.15	0.08	12.46	63.50
8	1	8	25.60	35.36	34.27	0.11	99.15	0.08	11.09	67.64
9	1	9	25.60	35.36	34.38	0.11	99.15	0.08	9.98	70.97
10	1	10	25.60	35.36	34.47	0.09	99.92	0.00	9.08	73.65

is fixed, the total costs increase as the weight for FNE increases. On the other hand, as the FNE weight increases, the total cost decreases when the optimal classification thresholds are applied.

The column of 'TC savings' shows how the total cost is reduced when an optimal threshold is applied compared to when a fixed threshold is applied. As shown in <Figure 2>, our proposed model is found to reduce the total cost by up to 73%.



〈Figure 2〉 Comparison of Total Social Cost using Fixed and Optimized Threshold

4. Concluding Remarks

This study proposed a novel recidivism prediction model that considers the asymmetric error cost structure. Using an open dataset from the ICSPR, we applied the recidivism prediction to the XGB model and compared it with other statistical and machine learning classification methods to verify that XGB is the best model for recidivism prediction accuracy. And then, we searched for the optimal classification threshold minimized the total cost, which is a weighted average of FPE and FNE. As a result, we found that the recidivism prediction model using the optimal classification threshold significantly reduces the total cost. Our study sheds a light on developing an effective recidivism prediction model, which minimizes the overall social costs caused by recidivism.

From the theoretical point of view, this study has the theoretical implication that the asymmetric error cost was reflected in the recidivism prediction and XGB, the latest classification prediction method, was applied to the recidivism prediction to consider the social cost. In addition, from the practical point of view, it is possible to utilize the proposed model in the present study as a reference for criminal judgment or review of parole, so that it is possible to proactively respond to the potential problem of recidivism. Also, it is possible to analyze the characteristics of criminals who are highly likely to have recidivism classified through learning. This can be used to study the causes of recidivism based on the theory of crime and to discuss ways to countermeasure about the recidivism.

In future research, our proposed model should be validated using other recidivism datasets. The dataset used in this study was generated in the early 1980s, so there is a limit to reflect the latest trends. In addition, though the cost for recidivism may be different according to the type of crime, our study did not consider it. Thus, it is needed to consider developing multiple recidivism prediction models according to the type of crime in the further study.

References

Breiman, L., "Bagging Predictors," *Machine Learning*, Vol.24, No.2(1996), 123~140.

Chen, T., and C, Guestrin, "Xgboost: A scalable tree boosting system," *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, (2016).

Joo, D., Hong, T., and I. Han, "The neural network models for IDS based on the asymmetric costs of false negative errors and false positive errors," *Expert Systems with Applications*, Vol.25(2003), 69~75.

Jung, S., "A Study on the Use of Big data in Criminal Law," *Journal of Public Policy Studies*, Vol.29, No. 2(2012), 161~184.

King, R. S., and B. Elderbroom, *Improving recidivism as a performance measure*, Washington, DC: Urban Institute, 2014.

Lee, H.-U., and H. Ahn, "An intelligent intrusion detection model based on support vector machines and the classification threshold optimization for considering the asymmetric error cost," *Journal of Intelligence and Information Systems*, Vol.17, No.4(2011), 157~173.

Nam, S., and S. Park, "Study on recidivism factors of prisoners," *Corrections Review*, Vol.50 (2011), 115~139.

New York Times, *Recidivism's high cost and a way to cut it*, 2011, Available at https://www.nytimes.com/2011/04/28/opinion/28thu3.html (Accessed 21 January 2019).

Prison Education News, The Cost of Recidivism: Victims, the Economy, and American Prisons, 2014, Available at https://prisoneducation.com/prison-education-news/the-cost-of-recidivism-victims-the-economy-and-american-pris-html (Accessed 21 January, 2019).

Schmidt, P., and A. D. Witte, "Predicting criminal recidivism using 'Split Population' survival time models", *Journal of Econometrics*, Vol.40, No.1(1989) 141~159.

Seong, H. G., "Methods and tasks in the prediction of criminal recidivism," *Proceeding of the 2006 Annual Conference of Korean Psychological Association*, (2006), 404~405.

Sharkey A.J., *Combining Artificial Neural Nets: ensemble and modular multi-net systems,* (Ed.), Springer Science & Business Media, 2012.

Turgut O., "Predicting recidivism through machine learning," *Ph.D. dissertation*, University of Texas at Dallas, 2017.

국문요약

비대칭 오류 비용을 고려한 XGBoost 기반 재범 예측 모델

원하람* · 심재승* · 안현철**

재범예측은 70년대 이전부터 전문가들에 의해서 꾸준히 연구되어온 분야지만, 최근 재범에 의한 범죄가 꾸준히 증가하면서 재범예측의 중요성이 커지고 있다. 특히 미국과 캐나다에서 재판이나 가석방 심사 시 재범 위험 평가 보고서를 결정적인 기준으로 채택하게 된 90년대를 기점으로 재범예측에 관한 연구가 활발해졌으며, 비슷한 시기에 국내에서도 재범요인에 관한 실증적인 연구가 시작되었다. 지금까지 대부분의 재범예측 연구는 재범요인 분석이나 재범예측의 정확성을 높이는 연구에 집중된 경향을 보이고 있다. 그러나 재범 예측에는 비대칭 오류 비용 구조가 있기 때문에 경우에 따라 예측 정확도를 최대화함과 동시에 예측 오분류 비용을 최소화하는 연구도 중요한 의미를 가진다.

일반적으로 재범을 저지르지 않을 사람을 재범을 저지를 것으로 오분류하는 비용은 재범을 저지를 사람을 재범을 저지르지 않을 것으로 오분류하는 비용보다 낮다. 전자는 추가적인 감시 비용만 증가되는 반면, 후자는 범죄 발생에 따른 막대한 사회적, 경제적 비용을 야기하기 때문이다. 이러한 비대칭비용에 따른 비용 경제성을 반영하여, 본 연구에서 비대칭 오류 비용을 고려한 XGBoost 기반 재범 예측 모델을 제안한다.

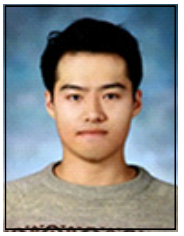
모델의 첫 단계에서 최근 데이터 마이닝 분야에서 높은 성능으로 각광받고 있는 앙상블 기법, XGBoost를 적용하였고, XGBoost의 결과를 로지스틱 회귀 분석(Logistic Regression Analysis), 의사결정나무(Decision Trees), 인공신경망(Artificial Neural Networks), 서포트 벡터 머신(Support Vector Machine)과 같은 다양한 예측 기법과 비교하였다. 다음 단계에서 임계치의 최적화를 통해 FNE(False Negative Error)와 FPE(False Positive Error)의 가중 평균인 전체 오분류 비용을 최소화한다. 이후 모델의 유용성을 검증하기 위해 모델을 실제 재범예측 데이터셋에 적용하여 XGBoost 모델이 다른 비교 모델 보다 우수한 예측 정확도를 보일 뿐 아니라 오분류 비용도 가장 효과적으로 낮춘다는 점을 확인하였다.

주제어 : 재범 예측, 비대칭 오류비용, 임계치 최적화, 데이터 마이닝, XGBoost

논문접수일 : 2019년 1월 28일 논문수정일 : 2019년 3월 28일 게재확정일 : 2019년 3월 28일
원고유형 : 단편논문 교신저자 : 안현철

* 국민대학교 비즈니스IT전문대학원
** 교신저자 : 안현철
국민대학교 비즈니스IT전문대학원
77 Jeongneung-ro, Seongbuk-gu, Seoul, 02707, Republic of Korea
Tel: +82-2-910-4577, Fax: +82-2-910-4017, E-mail: hcahn@kookmin.ac.kr

저 자 소 개



Ha-Ram Won

He holds a bachelor's degree in business administration from Halla University, Korea, and is currently a master's degree in Business Analysis Track at Business IT Graduate School, Kook-min University in Korea. His interests include business analytics and CRM and data-driven marketing analytics.

Jae-Seung Shim

is currently a master’s program at Graduate School of Business IT, Kookmin University, where he earned his bachelor’s degree in management information. His primary research interests include data mining and machine learning for the social sciences and business.

Hyunchul Ahn

is a professor of management information systems at the Graduate School of Business IT, Kookmin University, Republic of Korea. He obtained his Ph.D. from Korea Advanced Institute of Science and Technology. His major research interests are intelligent IS and IS adoption. His research has been published in Annals of OR, Computers in Human Behavior, International Journal of Electronic Commerce, International Journal of Information Management, International Journal of Production Research, Information & Management, etc.