Journal of Big Data

RESEARCH Open Access

A study on improving turnover intention forecasting by solving imbalanced data problems: focusing on SMOTE and generative adversarial networks

Jungryeol Park[1], Sundong Kwon[2] and Seon-Phil Jeong[3*]

*Correspondence:
spjeong@uic.edu.cn

[1] Technology Policy Research Division, Electronics and Telecommunications Research Institute (ETRI), Daejeon, South Korea
[2] MIS Department, Chungbuk National University, Chungbuk, South Korea
[3] Department of Computer Science, BNU-HKBU United International College, Zhuhai, Guangdong, China

Abstract

This study aims to improve the accuracy of forecasting the turnover intention of new college graduates by solving the imbalance data problem. For this purpose, data from the Korea Employment Information Service's Job Mobility Survey (Graduates Occupations Mobility Survey: GOMS) for college graduates were used. This data includes various items such as turnover intention, personal characteristics, and job characteristics of new college graduates, and the class ratio of turnover intention is imbalanced. For solving the imbalance data problem, the synthetic minority over-sampling technique (SMOTE) and generative adversarial networks (GAN) were used to balance class variables to examine the improvement of turnover intention prediction accuracy. After deriving the factors affecting the turnover intention by referring to previous studies, a turnover intention prediction model was constructed, and the model's prediction accuracy was analyzed by reflecting each data. As a result of the analysis, the highest predictive accuracy was found in class balanced data through generative adversarial networks rather than class imbalanced original data and class balanced data through SMOTE. The academic implication of this study is that first, the diversity of data sampling methods was presented by expanding and applying GAN, which are widely used in unstructured data sampling fields such as images and images, to structured data in business administration fields such as this study. Second, two refining processes were performed on data generated using generative adversarial networks to suggest a method for refining only data corresponding to a more minority class. The practical implication of this study is that it suggested a plan to predict the turnover intention of new college graduates early through the establishment of a predictive model using public data and machine learning.

Keywords: Imbalanced data, Turnover intention, SMOTE, Generative adversarial networks

Introduction

As public data disclosure began in many countries, public data have been used in various fields. Recognizing the importance of public data, thus, the South Korean government enacted the 'Act on Promotion of Provision and Use of Public Data' in 2013. In the past, public data was used in a limited and exclusive environment for research or policy development purposes. Recently, anyone can easily obtain public data from the government's portal websites or national institution websites in South Korea. These public data consist of various fields such as disaster safety, education, health and medical care, weather environment, public administration, public finance, national land management, social welfare, employment, food security, and cultural tourism. Therefore, various studies using these public data are being conducted. In particular, as youth unemployment has become a social problem in Korea, studies have been conducted to predict turnover intention and job performance using public data, machine learning, and deep learning to solve this problem [1, 2].

When conducting such a predictive study, the class of variables to be predicted in the data (ex: there is a turnover intention/there is no turnover intention) must be balanced. If the data with unbalanced classes are taught to the predictive model as it is, the model learns only information biased to a large number of classes, so the performance (prediction accuracy) of the model may decrease, and it is difficult to trust the predicted results [3–5].

This problem is called the data imbalance problem. Class imbalance data refers to at least one of its classes is usually outnumbered by the other classes. In the field of data mining, detecting events is a prediction problem, or, typically, a data classification problem. Rare events are difficult to detect because of their infrequency and casualness; however, misclassifying rare events can result in heavy costs. For financial fraud detection, invalid transactions may only emerge out of hundreds of thousands of transaction records, but failing to identify a serious fraudulent transaction would cause enormous losses. The scarce occurrences of rare events impair the detection task to imbalanced data classification problem. For example, for a data set where only 1% of the instances belong to the minority class, even if a model classifies all instances as the majority class, it still achieves an overall accuracy of 99%. However, the minority class instances, which we want to accurately classify, are all misclassified by this model though it achieves a very high accuracy.

This class imbalance problems have been reported to occur in a wide variety of real-world domains, such as facial age estimation [6], detecting oil spills from satellite images [7], anomaly detection [8], identifying fraudulent credit card transactions [9], software defect prediction [10], and image annotation [11]. Beurz and Van den Poel [3] suggested sampling techniques and algorithms to solve the problem, noting that class imbalance in learning data leads to negative results in modeling for customer departure prediction. Haixiang et al. [4] reviewed 527 articles related to unbalanced data published over the past decade. In particular, they found that although the imbalance data problem may occur in the field of business administration, there are not many papers that solve the imbalance problem in this area. Seliya et al. [5] mentioned imbalanced data as a unique problem that can occur in big data, and investigated how

the biased learning of imbalanced data leads to negative results in binary classification problems by investigating related studies for 10 years.

As Haxiang et al. [4] mentioned, predictive modeling using big data is also being carried out in the field of business administration. Applying this technique to the field of personnel organization can provide important insights in predicting the turnover intention of new employees, which has recently become a problem. Recently, the problem of youth employment, especially the deterioration of employment for college graduates, has become a social issue. However, companies are also complaining of a shortage of human resources, and one of the causes is the frequent turnover of new employees with college graduates [12]. According to the Ministry of Trade, Industry and Energy [13], the early retirement rate of experienced people is 13%, while the early retirement rate of new employees is 66%. According to a survey of economic activities by Statistics Korea [14], if young people aged 15–29 quit their first job, the average service period is 1.9 months. Most young people want to get a job, but those who succeed in getting a job are showing a paradoxical form of considering changing jobs or quitting. However, it is not easy for companies to measure the turnover intention of new employees. Because the turnover intention is a sensitive matter to individuals, and it is difficult to disclose it easily due to concerns about the personnel disadvantages that may arise when it is disclosed to the outside [15]. Also, even if it disclosed, it is not easy to determine whether the content is true. Companies are facing difficulties in selecting and managing new employees with turnover intentions.

Therefore, this study proposes a method to predict the turnover intention of new employees. If a predictive model is built using public data and machine learning algorithms that measure the turnover intention of new college graduates, and the independent variables used in model construction are measured on new employees and put into the model, their turnover intention can be predicted. This approach not only provides practical implications for promptly identifying new employees with turnover intentions in situations where it is challenging to easily obtain sensitive information such as turnover intention, but also supplements the limitations of traditional econometric models that solely focus on causal analysis. However, even in the process of building such a prediction model, class imbalance problems may still arise. For example, if there are 10 learning data, consisting of a ratio of 8 employees who are willing to change jobs and 2 employees who are not willing to change jobs, the model is likely to predict employees who are not willing to change jobs at 80%. This may result in fairness issues in the selection process for new employees who are willing to turnover, and could potentially have negative impacts on organizational management. Thus, it is crucial to address the class imbalance problem in the training data when building a predictive model. Therefore, in this study, a solution to the problem of imbalanced data was presented through oversampling using SMOTE and generative adversarial networks. In order to check the effect of solving the imbalanced data problem, the predictive model was analyzed by the original data with unbalanced classes and then modified data via two methodologies: SMOTE and generative adversarial network are also applied. Then, the results were compared. The contributions of this study are as follows.

- the diversity of data sampling methods was presented by expanding and applying generative adversarial networks, which are widely used in unstructured data sampling fields such as images and images, to structured data in business administration fields such as this study.
- two refining processes were performed on data generated using generative adversarial networks to suggest a method for refining only data corresponding to a more minority class.
- this study suggested a plan to predict the turnover intention of new college graduates early through the establishment of a predictive model using public data and machine learning.

Background and literature review

Imbalanced data

Imbalanced data refers to a case where the number of data belonging to one class is significantly larger or smaller than the number of data belonging to another class. At this time, a class with a larger number of data is called a majority class, and a class with relatively small data is called a minority class [16]. There is an imbalanced ratio (IR) to indicate the degree of data imbalance, and it is defined as follows.

$$IR = \frac{n_{maj}}{n_{min}} \tag{1}$$

Here, n_{maj} and n_{min} refer to the number of data belonging to the majority class and the minority class. If IR = 3, it means that the data belonging to the majority class is three times more than the data belonging to the minority class. Using these imbalanced data for prediction purposes can cause many problems [17]. For example, when learning a prediction model using machine learning, if the class of the target variable is imbalanced in the training data (e.g., when the class of the target variable is 0 or 1, the model is biased toward the class that occupies a large proportion [18]. As a result, this negatively affects the performance (prediction accuracy) of the model [3, 4]. This is called the class imbalance problem, and it is a problem that must be dealt with caution in prediction research, especially in finance, medical care, and manufacturing industry [19]. For example, in marketing, when making a model to determine whether or not a customer has churned, training the model through measured customer data, and then reflecting new customer data to predict the likelihood of their churn [20]. If most of the training data consist of customers who do not churn, the model that learned this has a problem of classifying new customers as customers who do not leave most of the time.

In general, a minority class is a subject of major interest in classification problems [21], However, when a class imbalance problem occurs, it is difficult to obtain satisfactory classification performance, and in some cases, user damage may occur due to incorrect classification of minority class data [22]. For example, if a patient with a rare disease is incorrectly classified as a normal person in the medical field, the opportunity to treat the patient is lost. In addition, additional damage may occur if payment history with a stolen credit card is incorrectly classified as normal in fraud detection in the financial field [23]. Since it is very important to accurately classify a minority class in this way, it

is necessary to solve the problem when the class of data used is imbalanced. Data sampling is a method to solve the problem of imbalanced data [24]. Data sampling is divided into under-sampling, which deletes data by matching the majority class to the size of the minority class, and over-sampling, which generates data by matching the minority class to the size of the majority class [24]. Undersampling can save time and cost of additional data collection but has the disadvantage of losing helpful information of the data [25]. On the other hand, oversampling may increase the time required to train the predictive model due to additional data generation, but it is possible to avoid the loss of important information in the collected data. Figure 1 shows the process of undersampling and oversampling.

Since one of various sampling techniques cannot be said to be better, and each technique has its pros and cons, it is necessary to utilize a technique suitable for a given data and task. But undersampling removes data, such methods risk the loss of important concepts. We focused on oversampling, not undersampling, to avoid loss of important information. Among the oversampling techniques, there is also a method of increasing the number of data through simple randomization [26]. However, this method can also lead to overfitting problems because it simply copies data. On the other hand, SMOTE generates data based on algorithms, so the likelihood of overfitting occurrence is less than that of simple random methods. As a consequence, in the field of imbalance data classification problem, a better performance can be easily achieved [27]. In addition, prior studies that solve the imbalance problem through oversampling have used SMOTE in their research, emphasizing that it is the most popular model among the proposed methods [27–35]. In particular, SMOTE is used as an imbalance class processing method for data measured on the Likert scale as used in this study [36], and there is evidence that it is also used as an oversampling method for the development of a turnover intention prediction model [37–39].

Synthetic minority oversampling technique (SMOTE)

SMOTE is a technique for generating data through bootstrapping and the K-nearest neighbor methods. For specific data belonging to a minority class, K nearest neighbors

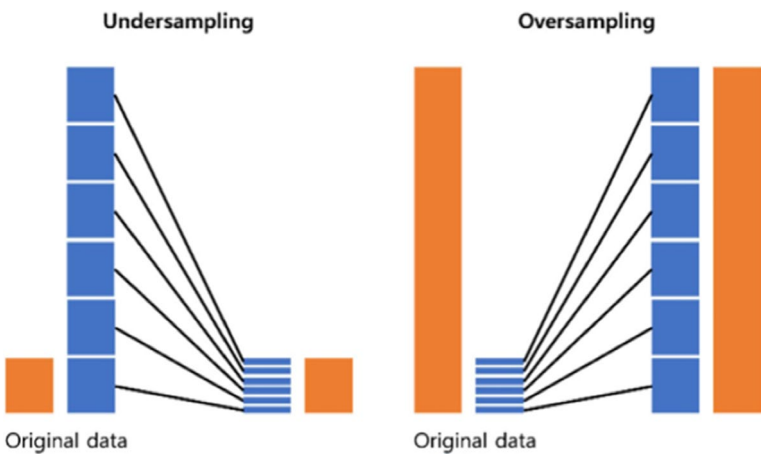

Fig. 1 Data sampling process [16]

of the same minority class are found, and new data are generated between them by creating a linear connection structure with the neighbors [40]. This is shown in the following Fig. 2.

Studies have been conducted to solve the class imbalance problem through such SMOTE. Wang et al. [28] applied SMOTE to address the problem of unbalanced data in healthcare. He et al. [29] proposed the use of SMOTE to solve the imbalance problem in various research fields. Chen [41] used several re-sampling techniques for finding the maximum accuracy of classification from fully labeled imbalanced training data set. SMOTE, Oversampling by duplicating minority examples, random under sampling, is mainly used to create new training data set. Rahman and Davis [30] tried to address class imbalance issue in medical datasets. They used undersampling techniques as well as oversampling techniques like SMOTE to balance the classes. Jishan et al. [31] proposed a method of preprocessing data to improve the accuracy of students' final grade prediction models using SMOTE. Douzas et al. [32] proposed an unbalanced data solution using SMOTE for the unbalanced dataset provided by UCI. Feng et al. [33, 34] proposed an SMOTE-based oversampling method to alleviate the class imbalance problem in software defect prediction. Nunez and Gatica [35] utilized SMOTE to solve the imbalance data problem in the Peruvian public investment prediction model.

Like this, SMOTE has been used in many fields to solve imbalanced data problems. However, when SMOTE generates data belonging to a minority class, it does not consider the location of data belonging to the adjacent majority class. Therefore, the problem of overlapping positions occurs [42]. In addition, since data is generated by relying only on the relationship between a few classes, overfitting may occur and prediction performance may deteriorate [43]. Therefore, it is necessary to generate new data that does not overlap with the existing data while considering the overall distribution of the data when generating data was derived [44]. To overcome these issues, we adopted the generative adversarial network (GAN) concept. GAN can solve overfitting and data superposition problems because it learns the actual data distribution of minor class and then generates similar data [45]. Through this, GAN can overcome the limitations of existing oversampling techniques.

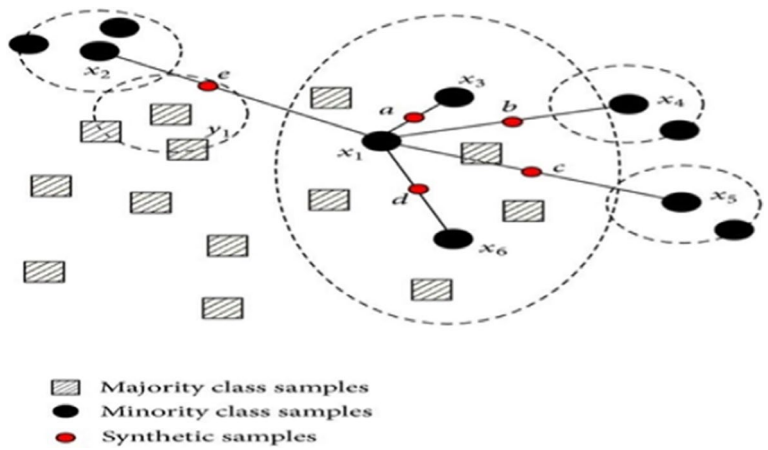

Fig. 2 SMOTE [40]

Generative adversarial network (GAN)

GAN is an unsupervised learning-based neural network that learns and optimizes by competing with a generator that generates data and a discriminator that compares the generated data with real data [46]. Since the two neural networks compete with each other, the quality of the generated data increases as learning takes place, and when the generator is optimized, the generated data becomes very similar to the quality of real data. As shown in the following Fig. 3, if a random type of noise (Z) is put into the generator, the generator generates fake data based on it. When this is put into the discriminator, the discriminator compares it with the existing actual data to determine whether the data generated by the generator is real or not. The result is expressed as a loss value, and when this value is sent to the generator, the generator proceeds with learning to generate fake data similar to real data based on this. After that, fake data is generated and sent back to the discriminator. This process proceeds until the discriminator does not clearly distinguish between real data and fake data, and when this state is reached, the generator is said to be optimized [45].

The goal of the generator is to maximize the probability (D(G(z))) that the discriminator will distinguish fake data into real data by 1. The goal of the discriminator is to maximize the probability (D(x)) of determining that the generated fake data is fake by comparing it with the actual data. The formula is as follows.

$$\underset{G}{Min}\ \underset{D}{Max}\ V(D,G) = E_{x\sim Pdata(x)}[logD(\mathrm{x})] + E_{z\sim Pz(Z)}[log(1-D(G(z)))] \tag{2}$$

GAN has been mainly used in the field of unstructured data such as images and videos [47]. However, it has also been used in the field of database sized structured data. In particular, studies applied to oversampling for solving the class imbalance problem have taken placed. Kim [48] presented a solution to the problem of data imbalance for predicting defaulted companies through generative adversarial networks based oversampling. Kim et al. [49] also suggested a solution to the problem of data imbalance for predicting credit card fraud through generative adversarial networks based oversampling. Mao et al. [50] presented a solution to the problem of data imbalance for predicting

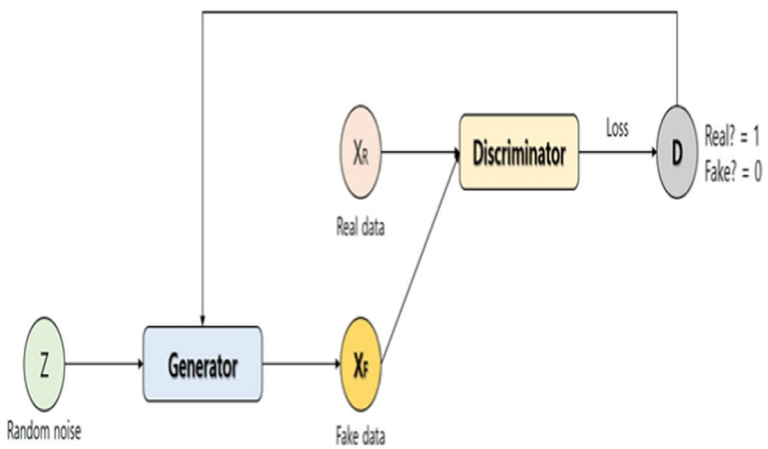

Fig. 3 Generative adversarial networks [46]

Table 1 Research on generative adversarial networks

Author/s	Summary
Kim [48]	Presented a solution to data imbalance problem for predicting credit card fraud through generative adversarial networks based on oversampling
Kim et al. [49]	Presented a solution to the problem of data imbalance for predicting credit card fraud through generative adversarial networks based oversampling
Park et al. [51]	Presenting a solution to the problem of data imbalance using generative adversarial networks and KNN
Mao et al. [50]	Presented a solution to data imbalance problem for predicting defective products in the manufacturing process through generative adversarial networks based on oversampling
Engelmann and Lessmann [52]	Presenting a solution to the problem of data imbalance through Wasserstein GAN

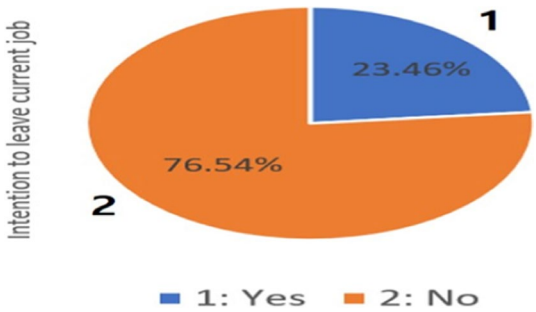

Fig. 4 Turnover intention rate

defective products in the manufacturing process through generative adversarial networks based oversampling. Table 1 shows the summary of these studies.

Research methodology

Introduction to the dataset

This study is to solve the class imbalance data problem and propose a method to improve the accuracy of the prediction model. For these purposes, data from the Graduate Occupations Mobility Survey (GOMS) of the Korea Employment Information Service was used. This data includes the turnover intention composed of class variables and personal characteristics, job characteristics, characteristics of the university, job search experience, and so on. The data consisted of 18,163 samples and 1270 variables. The data used in this study is freely available to anyone (Data download: https://github.com/jrpark16/GOMS.git). In order to predict turnover intention, 12,202 samples of respondents who worked in the past 4 weeks were extracted. The employee turnover intention, which is the dependent variable, was obtained from these samples as a discrete class. To the question of turnover intention, 2862 (23.46%) of the respondents answered 'Yes,' and 9340 (76.54%) answered 'No.' Fig. 4 shows that these classes are imbalanced data.

Variables

Thirteen explanatory variables were derived from the viewpoint of job choice motivation, needs satisfaction, and person-job fit based on previous studies. The following Table 2 shows the summarized variables.

Table 2 Variables

Main category	Middle category	Variables
Job choice motivation	Extrinsic motivation	Recognize the importance of workload
	Intrinsic motivation	Recognition of the importance related to the field of major
Needs satisfaction	Existence needs	Wages or income
		Job security
	Relatedness needs	Overall satisfaction at work
		group satisfaction
	Growth needs	HR system (promotion system)
		Individual development potential
		Social reputation of work
Person-job fit		The degree of agreement between the level of work and one's education level
		The degree of agreement between the skill level of the job and one's own skill level
		The degree of agreement between the content of work and one's major
		The degree of agreement between work and one's aptitude and interest

Oversampling

In this study, oversampling was performed through SMOTE and generative adversarial networks to match the imbalanced class ratio of the dependent variable in the original data. First, the original data was divided into 70% training and 30% test sets to avoid overfitting. Since then, the training set has divided data in response to "I have turnover intention" and data in response to "I do not have turnover intention." After that, the data in response to "I have turnover intention" was used as data for oversampling. Afterward, the class ratio of the dependent variable was matched by oversampling the previously classified data through the SMOTE algorithm provided by the Imbalanced-learn library. Next, GANs were built and trained on the previously classified data, and data oversampling was performed based on the optimized generator to match the class ratio of the dependent variable. The structure of the constructed GANs are as follows.

First, the generator receives a noise vector Z as input. Since the input Z is irrelevant to the output data, a dimension sufficient to contain the properties of the data may be arbitrarily set. This study was organized in 50 dimensions so that the data distribution in the latent space could be properly trained. The generator layer comprises $Z(50) \rightarrow 32 \rightarrow 64 \rightarrow 128 \rightarrow 13$ (number of explanatory variables). The discriminator layer is composed of the number of generator outputs $(13) \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 16 \rightarrow 1$ (sigmoid). This is shown in the following Fig. 5.

The batch size of the learning process was 50 and the epoch was performed 3,000 times. After learning 2000 times, the loss values of generator and discriminator did not change significantly around 0.7, so only 3000 times were performed and learning process was terminated. Then, data was generated through the learned generator. The learning process is shown in Fig. 6.

The generated data went through two purification processes. First, a prediction model for the turnover intention was trained through the original data, then the

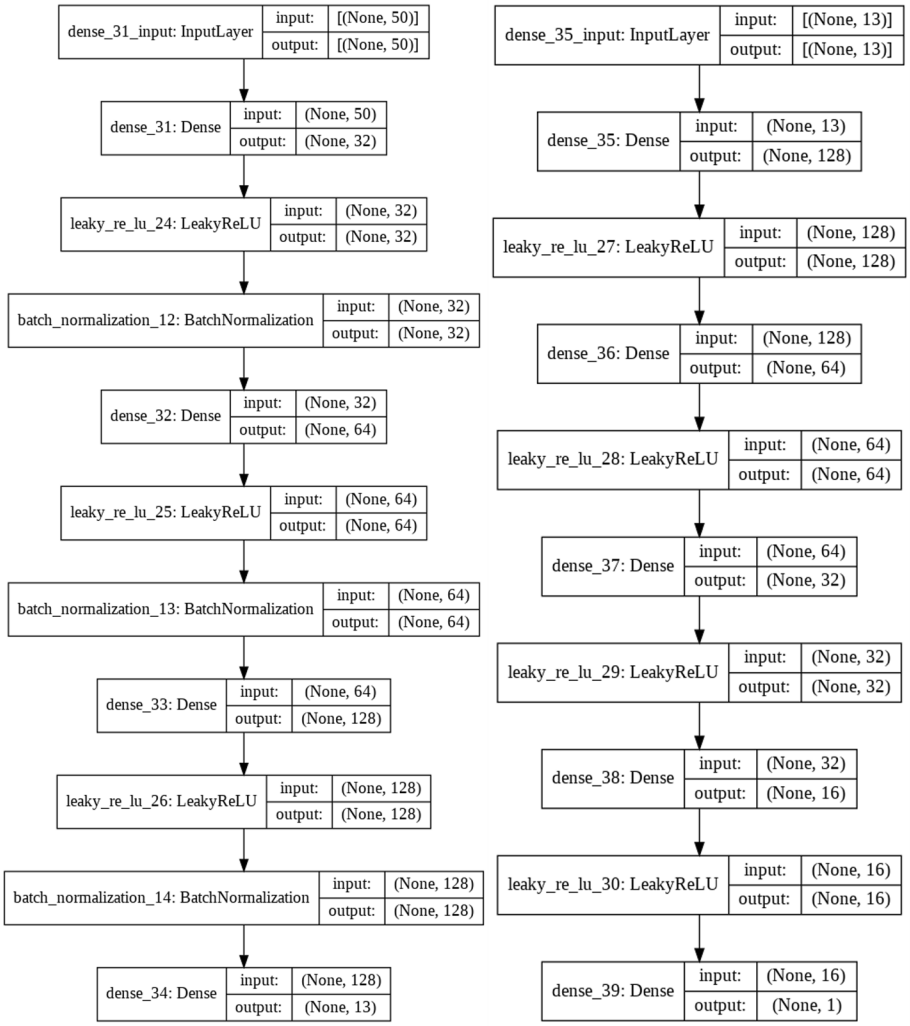

Fig. 5 Generator and discriminator layer

generated data was put in and analyzed through XGBoost, and only the data belonging to 'I have turnover intention' was primarily refined. After analyzing the refined data through logistic regression, only the data belonging to 'I have turnover intention' was secondarily refined. By including the 'I have turnover intention' data that had undergone two refinements in the imbalanced training set, the class ratio of the dependent variable in the training set was made the same. Through these processes, ① class imbalanced original data, ② class balanced data through SMOTE, and ③ class balanced data through generative adversarial networks were prepared.

Turnover intention prediction analysis

In order to predict turnover intention, a prediction model was trained using the supervised learning method during machine learning and then the prediction accuracy was analyzed. For this, the following method was performed. First, a prediction model for turnover intention was constructed in which all the variables presented in Table 2 were set as explanatory variables, and turnover intention was set as the

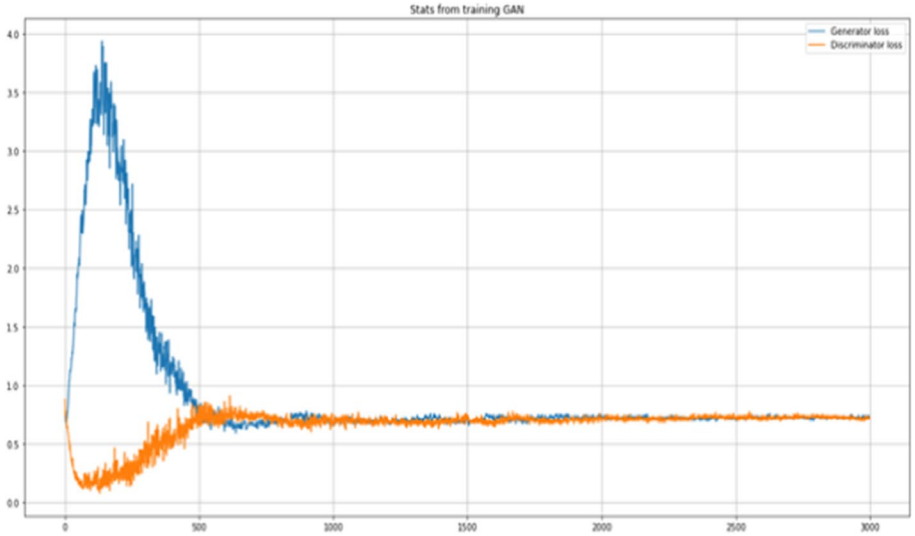

Fig. 6 Learning process

dependent variable. Second, the three previously prepared data were divided into 70% training set and 30% test set, respectively, and after learning the turnover intention prediction model with the training set, the model's prediction accuracy was analyzed with the test set. Third, the prediction model was analyzed using logistic regression (LR), K-nearest neighbor (KNN), and XGBoost (eXtreme Gradient Boosting: XGB). Accuracy, Precision, Recall, and F1-score were used as prediction performance evaluation indicators. In addition, cross-validation was performed four times to prevent the error of overfitting the prediction model only to a specific data set and to generalize the analysis results.

Meanwhile, Cross-validation was performed four times to prevent the model from overfitting only a specific data set and to generalize the analysis results. On the other hand, the widely used Accuracy, Precision, Recall, and F1-score are used as the performance evaluation indicators of the predictive model. The meaning and formula of each indicator are as follows.

Accuracy: Accuracy is the most intuitive performance measure and is simply a ratio of correctly predicted observations to the total observations.

$$(\text{Accuracy}) = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{FP} + \text{TN}} \tag{3}$$

Precision: Precision is the ratio of correctly predicted positive observations to the total predicted positive observations.

$$(\text{Precision}) = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

Recall: Recall is the ratio of correctly predicted positive observations to all observations in the actual "yes" class.

$$(\text{Recall}) = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{5}$$

F1-score: The F1-score is the weighted average of precision and recall.

$$(F1\text{-}score) = 2 \times \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{6}$$

Results and discussion

First, looking at the average of the results of 4 cross-validation after analysis using class imbalanced original data, LR showed accuracy of Accuracy = 0.783, Precision = 0.798, Recall = 0.956, and F1-score = 0.870. KNN showed accuracy of Accuracy = 0.761, Precision = 0.798, Recall = 0.915, and F1-score = 0.853. XGB showed accuracy of Accuracy = 0.785, Precision = 0.806, Recall = 0.942, and F1-score = 0.869.

Next, looking at the average of the results of 4 cross-validation after analysis using class balanced data using SMOTE, LR showed accuracy of Accuracy = 0.782, Precision = 0.797, Recall = 0.955, F1-score = 0.869. KNN showed accuracy of Accuracy = 0.765, Precision = 0.802, Recall = 0.914, and F1-score = 0.854. XGB showed accuracy of Accuracy = 0.784, Precision = 0.806, Recall = 0.941, and F1-score = 0.891. As a result of the analysis, it was found that there was no significant difference in prediction accuracy between the class imbalanced original data and the data with class balanced using SMOTE.

Next, looking at the average of the results of 4 cross-validation after analysis using class balanced data using GAN, LR showed accuracy of Accuracy = 0.799, Precision = 0.824, Recall = 0.937, F1-score = 0.877. KNN showed accuracy of

Table 3 Prediction analysis result

Set	Classifier	Accuracy			Precision			Recall			F1-score		
		R	S	G	R	S	G	R	S	G	R	S	G
Set1	LR	0.788	0.787	0.803	0.806	0.805	0.829	0.951	0.952	0.954	0.873	0.872	0.887
	KNN	0.769	0.771	0.800	0.808	0.810	0.826	0.915	0.914	0.938	0.858	0.859	0.878
	XGB	0.785	0.787	0.816	0.811	0.813	0.834	0.936	0.936	0.967	0.869	0.870	0.896
Set2	LR	0.770	0.770	0.799	0.782	0.782	0.828	0.958	0.957	0.944	0.861	0.861	0.882
	KNN	0.748	0.750	0.798	0.779	0.783	0.820	0.921	0.916	0.937	0.844	0.844	0.875
	XGB	0.772	0.774	0.829	0.789	0.790	0.843	0.945	0.947	0.954	0.860	0.861	0.895
Set3	LR	0.800	0.799	0.798	0.817	0.816	0.822	0.957	0.956	0.948	0.881	0.880	0.881
	KNN	0.778	0.779	0.794	0.819	0.823	0.818	0.917	0.912	0.932	0.865	0.865	0.871
	XGB	0.802	0.797	0.824	0.828	0.824	0.841	0.940	0.938	0.968	0.880	0.877	0.900
Set4	LR	0.776	0.774	0.796	0.787	0.785	0.818	0.958	0.958	0.941	0.864	0.863	0.875
	KNN	0.752	0.760	0.786	0.789	0.794	0.818	0.916	0.916	0.921	0.848	0.851	0.866
	XGB	0.781	0.781	0.824	0.796	0.797	0.836	0.946	0.946	0.955	0.865	0.865	0.892
Average	LR	0.783	0.782	0.799	0.798	0.797	0.824	0.956	0.955	0.937	0.870	0.869	0.877
	KNN	0.761	0.765	0.794	0.798	0.802	0.820	0.915	0.914	0.929	0.853	0.854	0.871
	XGB	0.785	0.784	0.823	0.806	0.806	0.838	0.942	0.941	0.951	0.869	0.868	0.891

R Raw, *S* Smote, *G* Gan

Accuracy = 0.794, Precision = 0.820, Recall = 0.929, and F1-score = 0.871. XGB showed accuracy of Accuracy = 0.823, Precision = 0.838, Recall = 0.951, and F1-score = 0.891.

Comparing the average accuracy with the class imbalanced original data, in the case of LR, the balanced data with GAN showed higher prediction accuracy of Accuracy = 1.6%, Precision = 2.6% and F1-score = 0.7% than the original data. However, in Recall, the original data showed high prediction accuracy by 1.9%. In the case of KNN, the balanced data with GAN showed higher prediction accuracy of Accuracy = 3.3%, Precision = 2.2%, Recall = 1.4%, and F1-score = 1.8% than the original data. In the case of XGB, the balanced data with GAN showed higher prediction accuracy of Accuracy = 3.8%, Precision = 3.2%, Recall = 0.9%, and F1-score = 2.2% than the original data. As a result of the analysis, the highest prediction accuracy was shown when class balanced data using GAN was used, and there was no significant difference in prediction accuracy between the original data and class-balanced data using SMOTE. To summarize this, Table 3 is as follows.

According to the results of this study, it was found that there was no significant difference in prediction accuracy when the original data in which the class ratio of the dependent variable was unbalanced and the data in which the class ratio was balanced through oversampling using SMOTE were used for prediction model analysis. However, this analysis result does not consider the data values belonging to multiple classes when oversampling through SMOTE, so the generated data overlap [42] and may overfit [43] consistent with the result. Therefore, when oversampling data using SMOTE, it is necessary to consider the overall distribution of prime class data, such as increasing the number of oversampling times or increasing the number of K that can be considered. In addition, after data oversampling, it is necessary to determine whether there are overlapping values through comparison with data belonging to multiple classes.

On the other hand, it was found that the prediction accuracy was higher than that of the original data when data balanced with class ratios were used for prediction model analysis through oversampling using GANs. These analysis results are consistent with previous studies [49, 51, 53] that prediction accuracy is higher when data with imbalanced class ratio is used for prediction model learning. When the model is trained on data with disproportionate class proportions, the accuracy is very high. However, the recall of classes with a small number of data rapidly decreases. One of the practical implications of this study is to suggest a way to predict employee turnover intention based on a pre-trained turnover intention prediction model. Therefore, it can be seen that the recall of the model is a very important study. Previous studies dealing with the class imbalance problem claim that class imbalance through oversampling can consequently increase the model recall [54–56]. Khoda et al. [54] compared oversampling and undersampling, and confirmed that solving the class imbalance problem through oversampling resulted in a higher F1 score. Prasetiyo et al. [55] found that the recall value was highest when the classes were balanced 50:50 through oversampling. Jo and Kim [56] suggested that in binary classification problems in various industries such as business, medical, and marketing, models learned from data with class balance through oversampling show the highest recall. In addition, it is believed that this is because similar but non-overlapping data were generated while considering the overall distribution

of minority data during data oversampling. Therefore, when analyzing, it is necessary to check the class ratio of the dependent variable in the data if the dependent variable consists of a discrete class. If the class ratio is imbalanced, it is necessary to balance it through data oversampling methods. In addition, if the analysis is performed by applying a method of generating non-overlapping data while considering the overall distribution of minority data, higher prediction accuracy can be derived.

Conclusion

This study focused on solving unbalanced data problems to derive higher predictive results in situations where data disclosure and sharing are active and artificial intelligence technology to analyze them is being developed and used for prediction in various fields. In addition, it was suggested that the performance of the predictive model can be improved when solving class imbalance using GANs. The academic implications of this study are that first, the diversity of data sampling methods was presented by expanding and applying GANs, which are widely used in unstructured data sampling fields such as images and videos, to structured data in business administration fields. Second, we proposed to perform two refinement processes through logistic regression analysis and XGBoost as a way to check whether the sampled data contains a small number of data properties well. The practical implication of this study is that it suggested a plan to predict the turnover intention of new college graduates early through the establishment of a predictive model using public data and machine learning. The main purpose of this study was to propose a method to improve the accuracy of the prediction model by resolving data imbalance through SMOTE, a data oversampling method, and GANs. There are limitations to its use. There are various parameters that can be applied to each classifier. Therefore, future research can be performed with an optimization algorithm such as grid search to find and apply optimal parameters, and in this way, higher prediction accuracy is expected.

Abbreviations

GAN	Generative adversarial networks
SMOTE	Synthetic minority oversampling technique
LR	Logistic regression
KNN	K-nearest neighbor
XGB	EXtreme Gradient Boosting

Acknowledgements
Not applicable.

Author contributions
JRP: designed the idea and experiments, and he wrote the manuscript and implemented the code. SDK: provided ideas during the analysis process. JSP has finalized the structure and content of the manuscript. All authors read and approved the final manuscript.

Authors information
JRP is a post-doctoral researcher at Electronics and Telecommunications Research Institute (ETRI), Technology Policy Research Division. He acquired a doctorate in Management Information Systems from Chungbuk National University. His main areas of interest are Machine Learning and Deep Learning, Data Analysis, and Business Prediction.
SDK is currently a professor in MIS program at Chungbuk National University. He acquired a doctorate in Management Information systems from Seoul National University. The main areas of interest are SCM-based Smart Factory, Machine Learning, and Deep Learning-based Business Predictions.
SPJ is currently an associate professor at the Division of Science and Technology of BNU-HKBU United International College. His current interests are Recommender Systems, Business Intelligence Systems, and Data Analysis.

Funding
Not applicable.

Availability of data and materials
https://github.com/jrpark16/GOMS.git.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
The publisher has the author's permission to publish this paper.

Competing interests
The authors declare that they have no competing interests.

Received: 7 April 2022 Accepted: 13 March 2023
Published online: 22 March 2023

References

1. Lee E-J, Cho H-S, Song Y-S. An exploratory study on determinants predicting university graduate newcomers' early turn over. J Corporate Educ Talent Res. 2020;22(1):163–93.
2. Choi J-W, Shin D-W, Lee H-J. Turnover rate prediction among IT firms according to job satisfaction and dissatisfaction factors: using topic modeling and machine learning. J Korean Data Inf Sci Soc. 2021;32(5):1035–47.
3. Burez J, Van den Poel D. Handling class imbalance in customer churn prediction. Expert Syst Appl. 2009;36:4626–36.
4. Haixiang G, Yijing L, Shang J, Mingyun G, Yuanyue H, Bing G. Learning from class-imbalanced data: review of methods and applications. Expert Syst Appl. 2017;73:220–39.
5. Seliya N, Zadeh A-A, Khoshgoftaar T-M. A literature review on one-class classification and its potential applications in big data. J Big Data. 2021;8:122.
6. Chao WL, Liu JZ, Ding JJ. Facial age estimation based on label-sensitive learning and age-oriented regression. Pattern Recogn. 2013;46(3):628–41.
7. Kubat M, Holte RC, Matwin S. Machine learning for the detection of oil spills in satellite radar images. Mach Learn. 1998;30:195–215.
8. Khreich W, Granger E, Miri A, Sabourin R. Adaptive ROC-based ensembles of HMMs applied to anomaly detection. Pattern Recogn. 2012;45(1):208–30.
9. Fawcett T, Provost F. Adaptive fraud detection. Data Min Knowl Disc. 1997;1(3):291–316.
10. Pelayo L, Dick S. Applying novel resampling strategies to software defect prediction. In: NAFIPS 2007–2007 annual meeting of the North American fuzzy information processing society. New York: IEEE; 2007. p. 69–72.
11. Zhang D, Islam MM, Lu G. A review on automatic image annotation techniques. Pattern Recogn. 2012;45(1):346–62.
12. Chung DB. Major factors affecting turnover intention of college graduates: comparison and analysis according to regular workers. Q J Labor Policy. 2019
13. Ministry of Trade, Industry and Energy. A survey on the supply and demand trend of industrial technology personnel in industrial technology. 2017.
14. Statistics Korea. The results of an additional survey of young people in the May 2018 economically active population survey. 2018.
15. Mobley WH. Some unanswered questions in turnover and withdrawal research. Acad Manag Rev. 1982;7(1):111–6.
16. Sun Y, Kamel M-S, Wong A-K-C, Wang Y. Cost-sensitive boosting for classification of imbalanced data. Pattern Recogn. 2007;40(12):3358–78.
17. Johnson J-M, Khoshgoftaar T-M. Survey on deep learning with class imbalance. J Big Data. 2019;6:27.
18. O'Brien R, Ishwaran H. A random forests quantile classifier for class imbalanced data. Pattern Recogn. 2019;90:232–49.
19. Tanha J, Abdi Y, Samadi N, Razzaghi N, Asadpour M. Boosting methods for multi-class imbalanced data classification: an experimental review. J Big Data. 2020;7:70.
20. Ahmad A-K, Jafar A, Aljoumaa K. Customer churn prediction in telecom using machine learning in big data platform. J Big Data. 2019;6:28.
21. Leevy J-L, Khoshgoftaar T-M, Bauder R-A, Seliya N. A survey on addressing high-class imbalance in big data. J Big Data. 2018;5:42.
22. Hasanin T, Khoshgoftaar T-M, Leevy J-L, Bauder R-A. Severely imbalanced Big Data challenges: investigating data sampling approaches. J Big Data. 2019;6:107.
23. Benchaji I, Douzi S, Ouahidi B-E, Jaafari J. Enhanced credit card fraud detection based on attention mechanism and LSTM deep model. J Big Data. 2021;8:151.
24. Hulse J-V, Khoshgoftaar T-M, Napolitano A-N. Experimental perspectives on learning from imbalanced data. In: Proceedings of the ACM international conference on machine learning. 2007. p. 935–42.
25. Kim H-Y, Lee W-J. On sampling algorithms for imbalanced binary data: performance comparison and some caveats. Korean J Appl Stat. 2017;30(5):681–90.
26. Kaggle. https://www.kaggle.com/code/rafjaa/resampling-strategies-for-imbalanced-datasets/notebook.
27. Chawla N-V, Bowyer K-W, Hall L-O, Kegelmeyer W-P. SMOTE: synthetic minority over-sampling technique. J Artif Intell Res. 2002;16:321–57.
28. Wang J, Xu M, Wang H, Zhang J. Classification of imbalanced data by using the SMOTE algorithm and locally linear embedding. In: 2006 8th international conference on signal processing, vol. 3. New York: IEEE; 2006.

29. He H, Bai Y, Garcia EA, Li S. ADASYN: adaptive synthetic sampling approach for imbalanced learning. In: 2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence). New York: IEEE; 2008. p. 1322–8.
30. Rahman MM, Davis DN. Addressing the class imbalance problem in medical datasets. Int J Mach Learn Comput. 2013;3(2):224.
31. Jishan ST, Rashu RI, Haque N, Rahman RM. Improving accuracy of students' final grade prediction model using optimal equal width binning and synthetic minority over-sampling technique. Decis Anal. 2015;2:1–25.
32. Douzas G, Bacao F, Last F. Improving imbalanced learning through a heuristic oversampling method based on k-means and SMOTE. Inf Sci. 2018;465:1–20.
33. Feng S, Keung J, Yu X, Xiao Y, Bennin KE, Kabir MA, Zhang M. COSTE: complexity-based OverSampling TEchnique to alleviate the class imbalance problem in software defect prediction. Inf Softw Technol. 2021;129:106432.
34. Feng S, Keung J, Yu X, Xiao Y, Zhang M. Investigation on the stability of SMOTE-based oversampling techniques in software defect prediction. Inf Softw Technol. 2021;139:106662.
35. Nunez NA, Gatica G. Applying profit-driven metrics in predictive models: a case study of the optimization of public funds in Peru. J Syst Manag Sci. 2022;12(2):52–65.
36. Aggarwal S, Saluja S, Gambhir V, Gupta S, Satia SPS. Predicting likelihood of psychological disorders in PlayerUnknown's Battlegrounds (PUBG) players from Asian countries using supervised machine learning. Addict Behav. 2020;101:106132.
37. de Oliveira JM, Zylka MP, Gloor PA, Joshi T. Mirror, mirror on the wall, who is leaving of them all: predictions for employee turnover with gated recurrent neural networks. Collaborative innovation networks: latest insights from social innovation, education, and emerging technologies research. 2019. p. 43–59.
38. Tao Z, Wu C, Zhao S. Research on the prediction of employee turnover behavior and its interpretability. In: Proceedings of the 2021 5th international conference on electronic information technology and computer engineering. 2021. p. 760–7.
39. Şahinbaş K. Employee promotion prediction by using machine learning algorithms for imbalanced dataset. In: 2022 2nd international conference on computing and machine intelligence (ICMI). New York: IEEE; 2022. p. 1–5.
40. Hu F, Li H. A novel boundary oversampling algorithm based on neighborhood rough set model: NRSBoundary-SMOTE. Math Probl Eng. 2013;2013:1–10.
41. Chen Y. Learning classifiers from imbalanced, only positive and unlabeled data sets. Ames: Department of Computer Science, Iowa State University; 2009.
42. Santoso B, Wijayanto H, Notodiputro K-A, Sartono B. Synthetic over sampling methods for handling class imbalanced problems: a review. In IOP Conference series: earth and environmental science, vol. 58. 2017. p. 1–8.
43. Krawczyk B. Learning from imbalanced data: open challenges and future directions. Prog Artif Intell. 2016;5(4):221–32.
44. Bagui S, Li K. Resampling imbalanced data for network intrusion detection datasets. J Big Data. 2021;8:6.
45. Goodfellow I-J, Pouget-Abadie J, Mirza M, Xu B, Warde-Farley D, Ozair S, Courville A, Bengio Y. Generative adversarial networks, In Proceedings of the neural information processing systems. 2014. p. 2672–80.
46. Kalin J. Generative adversarial networks cookbook, Packt. 2018.
47. Sampath V, Maurtua I, Aguilar Martín J-J, Gutierrez A. A survey on generative adversarial networks for imbalance problems in computer vision tasks. J Big Data. 2021;8:27.
48. Kim YL. GAN-based oversampling technique for imbalanced bankruptcy data processing, Ewha Womans University Master's thesis. 2020.
49. Kim Y-W, You Y-L, Choi H-Y. Fraud detection system model using generative adversarial networks and deep learning. Inf Syst Rev. 2020;22(1):59–72.
50. Mao Q, Lee H-Y, Tseng H-Y, Ma-S, Yang M-H. Mode seeking generative adversarial networks for diverse image synthesis. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR). 2019. p. 1429–37.
51. Park J-S, Ahn G-S, Heo S. Oversampling based on k-NN and GAN for effective classification of class imbalance dataset. J Korean Inst Ind Eng. 2020;46(4):365–71.
52. Engelmann J, Essmann S. Conditional Wasserstein GAN-based oversampling of tabular data for imbalanced learning. Expert Syst Appl. 2021;174(15):114582.
53. Son M-J, Jung S-W, Hwang I-J. A deep learning based over-sampling scheme for imbalanced data classification. KIPS Trans Softw Data Eng. 2019;8(7):311–6.
54. Khoda M-E, Kamruzzaman J, Gondal I, Imam T, Rahman A. Malware detection in edge devices with fuzzy oversampling and dynamic class weighting. Appl Soft Comput. 2021;112:107783.
55. Prasetiyo B, Muslim M-A, Baroroh N. Evaluation performance recall and F2 score of credit card fraud detection unbalanced dataset using SMOTE oversampling technique. In: Journal of physics: conference series, vol. 1918, no. 4. Bristol: IOP Publishing; 2021. p. 042002.
56. Jo W, Kim D. OBGAN: minority oversampling near borderline with generative adversarial networks. Expert Syst Appl. 2022;197:116694.

Publisher's Note