Journal of information and communication convergence engineering
http://jicce.org

J. Inf. Commun. Converg. Eng. 19(4): 228-233, Dec. 2021 Regular paper

Generating and Validating Synthetic Training Data for Predicting Bankruptcy of Individual Businesses

Dong-Suk Hong* and **Cheol Baik**

Big data Center, KCIS (Korea Credit Information Services), Seoul 04538, Korea

Abstract

In this study, we analyze the credit information (loan, delinquency information, etc.) of individual business owners to generate voluminous training data to establish a bankruptcy prediction model through a partial synthetic training technique. Furthermore, we evaluate the prediction performance of the newly generated data compared to the actual data. When using conditional tabular generative adversarial networks (CTGAN)-based training data generated by the experimental results (a logistic regression task), the recall is improved by 1.75 times compared to that obtained using the actual data. The probability that both the actual and generated data are sampled over an identical distribution is verified to be much higher than 80%. Providing artificial intelligence training data through data synthesis in the fields of credit rating and default risk prediction of individual businesses, which have not been relatively active in research, promotes further in-depth research efforts focused on utilizing such methods.

Index Terms: AI data, Synthetic data, Credit information, GAN, Bankruptcy prediction

I. INTRODUCTION

As cutting-edge fields, namely, big data and deep learning, have evolved and advanced, the demand for data disclosure for analysis has grown. However, studies have been conducted to preserve privacy by generating and disclosing synthetically processed data, rather than original data, because of the risk of revealing sensitive information about individuals when releasing original data [1]. In addition, building AI learning models, such as deep learning models, and improving their performance require voluminous data. Consequently, generating training data is a crucially significant research topic because collecting a large amount of data requires much time and effort [2].

On the contrary, predicting a corporation's bankruptcy possibility is a core forecasting problem in the financial sector, as financial bankruptcy because of corporate credit risk can result in high economic costs and, in extreme cases, economic downturns and corporate bankruptcy. However, training data for predicting corporate bankruptcy (including individual businesses) are difficult to access and are limited by the difficulties in using the actual data because of data imbalance issues.

In this paper, we propose a three-step procedure to generate synthetic training data for predicting the bankruptcy of individual businesses. We prepare the source data, generate synthetic data from the prepared source data, and evaluate the generated data to select the final dataset. We generate synthetic training data by applying a GAN variant, a representative model, and compare and evaluate them with respect to real data.

Considering previous studies on data synthesis with imbalanced data, only few studies target individual businesses. In this study, we first use the credit information of an actual individual business to generate data for predicting the bankruptcy of individual businesses. Second, we apply GAN vari-

Received 31 August 2021, **Revised** 22 October 2021, **Accepted** 02 November 2021
***Corresponding Author** Dong-Suk Hong (E-mail: dshong@kcredit.or.kr, Tel: +82-2-3705-5869)
Big data Center, KCIS(Korea Credit Information Services), Seoul 04538, Korea.

Open Access https://doi.org/10.6109/jicce.2021.19.4.228 print ISSN: 2234-8255 online ISSN: 2234-8883

ant to the individual business prediction problem and present the subsequent verification method. Third, we identify that increased information loss owing to full synthesis decreases the statistical significance through a comparison experiment between full and partial syntheses.

II. LITERATURE REVIEW

A. Synthetic Data

Synthetic data refers to virtual data that do not include sensitive information of individuals in the source data, are similar to real or processed data that preserve the statistical characteristics of variables of the source data, and are configured through certain modeling processes. Multiple alternatives, classification and regression trees (CARTs) [3], general adversarial networks (GANs) [4], and VAEs [5] are applied in data syntheses. Research on synthetic data was initially conducted to distribute the data publicly on behalf of the source data to address privacy constraints. However, utilizing synthetic data for data exploration and model establishment before analyzing actual data has garnered attention recently.

Synthetic data are classified into fully synthetic and partially synthetic data. No real data is included in neither of the categories, and reproducing the entire variable value results in fully synthetic data. Data created by synthesizing only a few values at a high risk of exposure denotes partially synthetic data. Because fully synthetic data often has a large bias in estimation, it is regarded as a significant method with high information loss, while partial synthetic data still exhibits less information loss but a high risk of exposure.

B. GAN (Generative Adversarial Network)

The GAN is a representative deep-learning technique used to generate synthetic data. It uses a discriminator and generator to determine whether the generated sample coincides with the real data distribution. Generative models aim to enhance the feasibility of generated samples until the discriminator cannot distinguish between actual and synthetic data [4].

The GAN framework mentioned above may exhibit a poor performance compared to other networks. When the discriminator learns the actual data distribution perfectly, Minmax game [6] contradictions occur owing to certain data distributions oscillating during learning.

Several advanced studies have proposed modified GANs to compensate for the constraints of a conventional GAN. TGAN [7] and CopulaGAN [8] are specialized in synthesizing tabular data with different data types, different distributions, and CTGAN [9] is specialized in synthesizing unbalanced data among tabular data. Various studies are conducted to generate GAN-based synthetic data in various fields that suffer from unbalanced data issues, such as intrusion and defect detection in manufacturing processes.

C. Bankruptcy Prediction

Predicting a corporate bankruptcy possibility is a core forecasting problem in the financial sector, as financial bankruptcy because corporate credit risk can result in high economic costs and, in extreme cases, economic downturns and corporate bankruptcy [10].

The first model in this field is Altman's model [11], which leverages corporate financial information to produce Z-scores based on statistical techniques. Other statistical-based models, such as the probit model, logit model, survival analysis, etc., have since been introduced. Machine learning and deep learning are state-of-the-art methodologies that lack both volume and quality compared to predictive studies using traditional statistical methods in the financial sector. Corporate bankruptcy forecasting studies traditionally focus on financial information and rarely on individual businesses, which is the main prediction goal in this study. An individual business is defined as a business that is obliged to pay VAT and income tax and differs from a corporate business in many ways, such as being sensitive to economic fluctuations, being regarded as an individual, and having a poor funding environment. Consequently, it is necessary to develop a prediction model that is different from the based on typical corporate entities for predicting individual business bankruptcy.

D. Imbalanced Data

In classification problems, when the degree of data imbalance is severe (the number of samples in one category is significantly less than that in other categories), the data imbalance issue causes the classifier fails to adequately classify the data, which has a large impact on the model performance. Data with imbalanced sampling are diverse in the real world, among which corporate data are sparse in the number of companies with a history of bankruptcy. The prediction model for individual businesses also requires addressing the degradation of learning performance caused by unbalanced data challenges. Zhou [10] confirmed that re-sampled datasets, such as those utilizing under-sampling (which eliminates the value of real data) and over-sampling (which generates virtual data from real data), influence the model performance.

III. GAN-BASED TRAINING DATA GENERATION

A. Credit Information

The GAN-based training data generated in this study are processed and partially synthesized data regarding credit

information, such as loans, overdue payments, and bankruptcies of individual businesses in Korean manufacturing fields managed by the Korea Credit Information Services (KCIS) [12]. Hence, we introduce credit information that constitutes the training data.

Credit information of individual businesses with credit transactions, such as loans, is collected and managed by the KCIS as corporate credit information, which is defined as the credit information of corporate businesses under the Credit Information Act. Corporate credit information includes essential information for pre-emptive corporate risk management, development of corporate financial products, forecasts provided by policy institution regarding financial markets and economic activities, credit assessment information, such as loans, and technology credit information.

The credit information used as input in the training data includes whether the owner is co-representative, elapsed period after registration of owner information, etc., and loan-related information includes the total balance of monthly credit contributions and total number of Korean loan institutions. Information related to overdue payments includes unresolved overdue balances in the base month and number of overdue experiences. In contrast, information related to technical credit includes whether a technical loan is held in the base month and a technical credit rating.

B. Training Data Generation

Artificial intelligence data refers to all associated data used to learn artificial intelligence models, such as machine learning and deep learning. Learning data is an important resource, along with the AI service model. However, data collection is often difficult, and building data for learning through a series of processes, such as refining, checking, and model validation, requires capital, time, and workforce resources.

Synthetic data are generated from actual data considering the level of anonymization, degree of data utility, operating costs, and loss of information compared to actual data. On the contrary, synthetic learning data can be distributed in appropriate formats as training data for various purposes, such as quick navigation of data and hypothesis verification of models before analyzing source data.

C. Proposed Method

The procedure for generating training data typically consists of preparing source data, generating synthetic data, and verification, in this order. Fig. 1 outlines the procedure for generating synthetic training data to predict the bankruptcy of individual businesses.

First, in the source data preparation phase, we generate samples from which the synthetic training data are generated through certain processes, such as feature selection, sampling, and data preprocessing from real data. The independent variables that make up the training dataset are individual business owners' unit characteristics, selected (p-value $\leq$ 0.05, R-square value $\geq$ 0.65) through independent sample T-test and logistic regression by step selection, from the initial candidate features defined by expert focus group discussions. After feature selection, random under-sampling is performed to meet the target size requirements of the generated and distributed training dataset, and data preprocessing is performed in a manner suitable for learning.

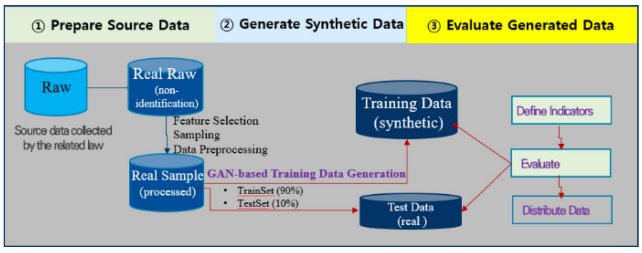

Fig. 1. Procedure for generating synthetic training data.

In the synthetic data generation phase, the prepared real data samples are separated into training and test datasets (with a ratio of 90:10). Then, they are randomly under-sampled such that the normal (major) record of the real data is 50% of the entire training dataset. On the contrary, CTGAN and other techniques are applied to generate synthetic data to ensure that the default (minor) record is 50% of the entire training dataset as well. In addition, partially synthetic data are generated by combining actual data from the normal and synthetic data from the default records.

In the model validation phase, the validation indicators used to evaluate the generated data are defined and verified. The training data generated can be distributed after determining whether it is feasibly deployable, where the training data are the generated partially synthetic data. The test data is the actual data.

To determine whether the data are deployable, we utilize the statistical distribution and important feature matching relative to the real data as a basic condition. A dataset has superior classification performance of a given data model among datasets that meet a certain level of basic conditions is deemed "deployable".

IV. EXPERIMENTAL ANALYSIS

A. Experimental Datasets

The dataset used for this experiment includes credit information, such as loans and the delinquency status of individual businesses (approximately 480,000 borrowers) operating in the Korean manufacturing industry. This information com-

prises a total of 41 types of corporate credit information selected through statistical verification of initial candidates in the source data preparation, where data from the preceding three months of the base year of default, June 2019, were input as an independent variable. The dependent variable, default, depends on the purpose of the credit risk predictive management. The concept of default defined in the new BIS convention includes not only defaults under the clearing agreement but also final defaults on household checks, current checks, promissory notes, defaults on debts, and immunity from bankruptcy.

The entire dataset was separated with a ratio of 90 (for training) to 10 (for the test), and the training data were again divided with a ratio of 80 (for training) to 20 (for validation and hyperparameter tuning). The target size of the synthetic training data was approximately 50,000 records, which corresponded to approximately 12% of the actual data.

B. Experimental & Evaluation Methods

As we have discovered through our analysis of preceding studies [2] that the instance size of both the majority and minority classes affects the model performance, to solve the data imbalance issue, the target value for the ratio of the number of majority class records to that of the minority class was initially set to 0.5, that is, a ratio of 50:50.

Based on samples of 50,000 records extracted from real data, records of normal borrowers were under-sampled, while records of default borrowers were synthesized from real samples. The generated datasets were compared with respect to their synthesis techniques. The under-sampling technique applied in the experiment is random under sampling (RUS); whereas the synthesis techniques are GAN-based modified models: (1) CTGAN [9], (2) CopulaGAN [8], and (3) TVAE (which applies VAE [5] to the GAN framework).

We verify the performance of the generated synthetic training data compared to the actual data, as well as statistical characteristics. In other words, we assess whether there are significant variations in the critical model features. The area under the curve (AUC) and recall of the logistic regression model were used to analyze the classification performance, and the KS-Test score index was used for statistical verification (i.e., identifying the probability that both real data and generated data are sampled from the identical distribution).

To identify these important features, we used the permutation importance provided in the Scikit-learn library. In the trained model, when the order of specific feature values in the dataset was arbitrarily shuffled, the importance value, which influences the prediction result of the model, was calculated. Then, the rank of the features was determined according to the obtained importance value.

It is a key indicator that signifies how well the model can classify actual defaults as defaults. Among other indicators regarding classification performance, we mainly use the recall of the actual and synthetic data models and their performance ratio (relative comparison of the reproduction rate to the model learned from the actual data sample and the newly generated learning data).

In the experiment, the cross-validation was performed using GridSearchCV provided in the Scikit-learn library, which utilizes the average of repeatedly measured values that were repeated five times according to the number of divided folds (cv=5). In addition, hyperparameters, such as the C coefficient and max_iter, were tuned to maximize the generalization performance of the logistic regression model. CTGAN and CopulaGAN used the following hyperparameters: epochs of 300, learning rate of 0.0002, ADAM optimizer, and weight decay of 0.000001 for the constructor and discriminator. TVAE required epochs of 300, l2 scale of 0.00001, and loss factor of 2.

C. Evaluation Result

Table 1 reports the experimental results for three training datasets generated by synthesizing default records of real samples by (1) CTGAN, (2) CopulaGAN, and (3) TVAE, along with 380,000 record-based real data and 50,000 record-based real samples.

Considering the data imbalance of each dataset before verification, observe that the actual data comprises less than 0.01 and actual sample data less than 0.1 of the datasets, indicating that the data imbalance is considerably large. The synthetic training dataset matched to 0.5 through under-sam-

Table 1. Experimental results

	Real Raw	Real Sample	Partially Synthetic Data (1) CTGAN	(2) Copula-GAN	(3) TVAE
1 Record #	38	5	5	5	5
2 Imbalanced rate	0.006	0.05	0.5	0.5	0.5
3 Key variables match rate	-	1	0.7	0.7	0.5
4 KS-test score	-	1	0.85	0.81	0.87
5 AUC	0.61	0.70	0.82	0.73	0.81
6 Recall	0.23	0.40	0.70	0.47	0.67
7 Performance ratio(recall)	-	1	1.75	1.18	1.68

- Records # (number of records): number of training data records (in units: 10,000)
- Key variable match rate: the value representing the degree of redundancy by comparing the top 10 features of the synthetic dataset to the top 10 features of the real sample (if the same, 1 is applied).
- Performance ratio: the value of the synthetic data performance generated by the experiment compared to the performance of the source data sample (if the same, 1 is applied).

pling and synthesis is verified according to the synthesis technique.

First, the degree of important feature redundancy is largely consistent with the important features of the actual sample dataset in three newly generated synthetic learning datasets with both CTGAN and CopulaGAN, equaling 0.7, and TVAE, equaling 0.5. This implies that the experimental data were generated in a similar construction to the actual data. Second, statistical verification exhibited that all three newly generated synthetic training datasets scored 0.8 or higher and were superior to the original dataset with the increasing order of TAVE, CTGAN, and CopulaGAN. Then, we recognized that synthetic training data exhibited a statistical distribution similar to real data, with a probability of 80% or higher, meaning that the generated data were sampled from the same distribution as the real data. Third, the classification performance (recall, AUC) of models that learned through the three synthetic training datasets improved compared to those of models that learned via the actual sample datasets. This showed that the obtained performance was excellent in the order of CTGAN, TVAE, and CopulaGAN. Therefore, the model established using 50,000 CTGAN-based synthetic training data demonstrates better performance compared to learning with 50,000 records of real sample data (and even 380,000 records of real source data).

Table 2 presents the important features of the actual sample and generated datasets.

Furthermore, we compared the partially synthesized dataset based on the default and normal owner records, such as the proposed method, to a fully synthesized dataset that was synthesized using the entire record data without any data discrimination.

In the fully synthetic dataset that was under the same conditions as the proposed method, the statistical index value decreased by approximately 80% compared to the partial synthesis-based method. Therefore, as mentioned in previous studies, we identified that full synthesis induced an increased information loss that, in turn, decreased the statistical significance.

Furthermore, the fully synthesized dataset exhibited lower matching rates among key variables and classification performance than those of the proposed dataset, especially with an AUC value lower than 0.5, which proved to be much less useful as training data. In addition, these experimental results imply that it is desirable to apply partial synthesis rather than full synthesis when generating learning data of credit information consisting of voluminous features and records in terms of data utility.

In general, training data can be created and evaluated based on certain priorities, such as the degree of data utility, level of anonymization, and operational costs.

In this study, we propose CTGAN-based synthetic training data with a highest recall performance ratio (LR, approximately 1.75) as the final dataset, with statistical indicators and important feature matching rates that meet or surpass a certain level.

Table 2. Top10 features

Dataset		Important Features (Top10)
Real	Real sample	Total credit balance Total number of lenders (Korean won loan) Technical credit rating Late payment rate (not released) Late payment balance (not released) Number of overdue cases (not released) Longest overdue period (release case) Average overdue period (release case) Payment guarantee payment Longest overdue period (not released)
Generated	CTGAN based	Total number of lenders (Korean won loan) Effective collateral amount interval Total credit balance Late payment balance (not released) Late payment rate (not released) Payment guarantee payment Total number of credit card overdue Number of overdue cases (not released) Average overdue period (not released) Average overdue period (release case)
	Copula GAN based	Effective collateral amount Total number of lenders (Korean won loan) Total number of credit card overdue Effective collateral amount interval Longest overdue period (release case) Total credit balance Average overdue period (release case) Number of overdue cases (not released) Technical credit rating Payment guarantee payment
	TVAE based	Effective collateral amount interval Total credit balance Loan Balance Interval (non-bank) Longest overdue period (not released) Total number of credit card overdue Number of overdue cases (not released) Longest overdue period (release case) Average overdue period (release case) Total number of lenders (new loan in non-bank) Average overdue period (not released)

Table 3. Comparison of partial synthesis-based proposed method and full synthesis

	Real Sample (RUS)	Partially Synthetic Data (CTGAN)	Fully Synthetic Data (CTGAN)
Record #	5	5	5
Imbalanced rate	0.05	0.5	0.5
Key variables match rate	1	0.7	0.2
KS-test score	1	0.85	0.68
AUC	0.70	0.82	0.30
Recall	0.40	0.70	0.51

V. CONCLUSION

In this study, we generated synthetic learning data via CTGAN, a GAN variant technique, to address the data imbalances that occur owing to sparse credit information data related to the bankruptcy of individual businesses. In addition, statistical verification, key feature verification, and classification performance validation of the model were performed via through assessments done on the newly generated training data. The verification results indicate that the generated synthetic training data preserves the statistical distribution characteristics of the actual source data, and important features in the real data model can be estimated (approximately 70% at accuracy) when a synthetic training data model is constructed. Even regarding recall verification, which is considered the most important factor in bankruptcy prediction, we confirm that synthetic training data models can enhance recall performance by 1.75 times compared to real data models, which can resolve the degradation caused by data imbalance issues.

Our results indicate that CTGAN-based partial synthesis techniques could be used for generating reliable and virtual data that can be leveraged regarding individual businesses' credit information training data, which is difficult to access in practice. We expect that further research on predicting credit risk for individual businesses (e.g., credit rating for individual businesses, predicting bankruptcy rates by industry [13], etc.) will be promoted. In the future, we will be able to generate training data that includes more time-series data, such as representative personal credit data of individual businesses, and refine the evaluation and validation of generated data.

REFERENCES

[1] J. Y. Kang, S. Y. Jeong, D. W. Hong, and C. H. Seo, "A study on synthetic data generation based safe differentially private GAN," *Journal of The Korea Institute of Information Security & Cryptology*, vol. 30, no. 5, pp. 945-956, 2020. DOI: 10.13089/JKIISC.2020.30.5.945.

[2] D. S. Hong and C. Baik, "Comparison of resampling methods for generating learning data for predicting the default of individual business," *Proceedings of KIIS Spring Conference 2021*, vol. 31, no. 1, pp. 61-62, 2021.

[3] J. P. Reiter, "Using CART to generate partially synthetic public use microdata," *Journal of Official Statistics*, vol. 21, no. 3, pp. 441-462, 2005.

[4] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," In *Proceedings of the 27th Neural Information Processing Systems*, vol. 2, pp. 2672-2680, 2014.

[5] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," In *International Conference on Learning Representations*, pp. 1-14, 2013.

[6] T. Chen and C. Guestrin, "Xgboost: Ascalable tree boosting system," In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 785-794, 2016. DOI: 10.1145/2939672.2939785.

[7] L. Xu and K. Veeramachaneni, "Synthesizing tabular data using generative adversarial networks," *arXiv:1811.11264*, 2018.

[8] SDV (Synthetic Data Vault) CopulaGAN Model [Internet], Available: https://sdv.dev/SDV/user_guides/single_table/copulagan.html.

[9] L. Xu, Modeling tabular data using conditional GAN. *Massachusetts Institute of Technology* [Online], 2017, Available: https://dai.lids.mit.edu/wp-content/uploads/2020/02/Lei_SMThesis_neo.pdf.

[10] L. Zhou, "Performance of corporate bankruptcy prediction models on imbalanced dataset: The effect of sampling methods," *Knowledge-Based Systems*, vol. 41, pp. 16-25, 2013. DOI: 10.1016/j.knosys.2012.12.007.

[11] E. I. Altman, "Financial ratios, discriminant analysis and the predication of corporate bankruptcy," *Journal of Finance*, vol. 23, no. 4, pp. 589-609, 1968. DOI: 10.1111/j.1540-6261.1968.tb00843.x.

[12] Korea Credit Information Services (KCIS), [Internet], Available: http://www.kcredit.or.kr/eng/index.do.

[13] D. S. Hong, H. J. Baeck, and H. J. Shin, "The credit information feature selection method in default rate prediction model for individual businesses," *Journal of The Korea Society for Simulation*, vol. 30, no. 1, pp. 75-85, 2021. DOI: 10.9709/JKSS.2021.30.1.075.

Dong-Suk Hong

received her Ph.D. in computer science and engineering from the Konkuk University, Seoul, Korea and now works in the big data center of Korea Credit Information Services (KCIS). Her main research area is data analysis of credit information, and she is particularly interested in evaluating the performance of prediction models based on machine learning and deep learning. https:/orcid.org/0000-0003-0236-9357

Cheol Baik

received his Ph.D. in statistics from Yonsei University, Seoul, Korea. Currently, he leads data strategy team of big data center of Korea Credit Information Services (KCIS). His research interests include actuarial risk management, empirical data analysis, and their interpretation in the finance and insurance industries. https:/orcid.org/0000-0001-6789-4123