Comparison of Machine Learning Based on Category Theory

Heng Zhao[1,*], Yixing Chen[2] and Xianghua Fu[1]

[1]*College of Big Data and Internet, Shenzhen Technology University, Shen Zhen, 518118, China*
[2]*College of Computer Science and Software Engineering, Shenzhen University, Shen Zhen, 518118, China*
E-mail: 15186958102@163.com
[*]*Corresponding Author*

Received 17 December 2022; Accepted 13 February 2023;
Publication 14 April 2023

Abstract

In recent years, machine learning has been widely used in data analysis of network engineering. The increasing types of model and data enhance the complexity of machine learning. In this paper, we propose a mathematical structure based on category theory as a combination of machine learning that combines multiple theories of data mining. We aim to study machine learning from the perspective of classification theory. Category theory utilizes mathematical language to connect the various structures of machine learning. We implement the representation of machine learning with category theory. In the experimental section, slice categories and functors are introduced in detail to model the data preprocessing. We use functors to preprocess the benchmark dataset and evaluate the accuracy of nine machine learning models. A key contribution is the representation of slice categories. This study provides a structural perspective of machine learning and a general method for the combination of category theory and machine learning.

Keywords: Web engineering, Big data, machine learning, preprocessing, category theory, accuracy.

Journal of Web Engineering, Vol. 22_1, 41–54.
doi: 10.13052/jwe1540-9589.2213

1 Introduction

Machine learning is the scientific study of algorithms and statistical models that are used by computer systems to perform specific tasks. These algorithms and statistical models are used for various purposes such as data mining, image processing, and predictive analytics [1]. A survey has evidenced that there are various algorithms and models in the digital arena [2]. The components of various algorithms and models need to be combined into the application domain. This combination of machine learning algorithms spans data normalization, unbalanced data processing, model training, and prediction. The combination with various machine learning algorithms has contributed to the development of various domains [3–5]. However, it is increasingly difficult to understand how the various parts of the combination interact with each other. For decades, methods for optimization of machine learning have been proposed [6]. Alves et al. [7] addressed the process fairness of machine learning models to improve classification fairness on the tabular and textual data. Zhu et al. [3] proposed an unsupervised representer mix2vec for mixed data with complex characteristics to represent the data, which is used to provide the interpretability of the representation. There are some limitations to these studies: In the case of multiple types of data and model, the interpretability of machine learning models is difficult [4].

To better manage the numerous models and data, this study introduces the category theory that was introduced as the mathematical structures in the middle of the 20th century. In recent years, the amount of research on category theory has been increasing. Yi et al. [10] used computer visual modeling of plant morphology as an object of study, and combined category theory, a formal mathematical tool for modeling, to explore a unified set of conceptual modeling methods and tools for partial plant measurement systems. Lu et al. [11] proposed data modeling for the integration between specifications and verification for cylindricity based on category theory, and various operations of the model are implemented through transformation of the functors. The combination of machine learning is constantly developing, and category theory provides a modeling method for the combination of various theories. The study of category theory in machine learning has been gradually introduced. Culbertson et al. [12] constructed parametric and nonparametric Bayesian inference models on function spaces by the category-theoretic approach to provide a foundation for supervised learning problems. They also showed how to view general stochastic processes by using the category of functor. Kamiya et al. [13] introduced a category

framework to formalize Bayesian inference and learning based on the notions of Bayesian inversions and the gradient learning functor GL constructed by Cruttwell et al. [14]. In addition, they obtained categorical formulations of batch and sequential Bayes updates. The study of machine learning from a category-theoretic perspective is a new field of research with a short history; however, the prospect is promising [15].

In this paper, we apply category theory to construct the representation of machine learning. In particular, it explains how slice categories can serve machine learning, and how functors represent the connection of dataset and model. The models are nine popular machine learning classifiers, including stand-alone classifiers (i.e. logistic regression [16], SVM [17], KNN [18], and decision tree [19]) and decision tree ensembles (i.e. random forest, AdaBoost [20], bagging [21], gradient boosting [22] (GBDT), and XGBoost [23].) Our aim is to compare the performance of machine learning classifiers trained on benchmark datasets that contain different oversampling approaches on a category-theoretic basis.

The remainder of this paper is organized as follows: Section 2 gives an overview of category theory. Section 3 explains how to define machine learning modeling from a category-theoretic perspective. The performance of classification models by applying the category theory is detailed in Section 4. In Section 5, we discuss the role of category theory in machine learning and detail directions for future work.

2 Introduction of Category Theory

To illustrate how machine learning can be represented by category theory, the basic concept of category theory is given at an abstract level. Category theory provides a framework in which a category can be regarded as a directed graph describing a mathematical structure, a functor, as the relationship between two categories, and natural transformation as the relationship between two functors.

2.1 Category

A category is defined as a set of objects, a class of morphisms or arrows between these objects, a composition of morphisms ($\circ$) and an identity morphism for the object (id) [24]. A category is a collection of objects that are related to each other by using morphism. Figure 1 reveals the schematic diagram of the category C. The a, b, and c indicate the objects $\in C$.

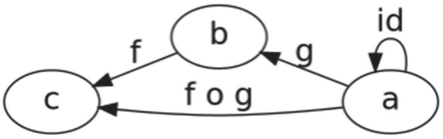


Figure 1 Schematic diagram of the category *C*.

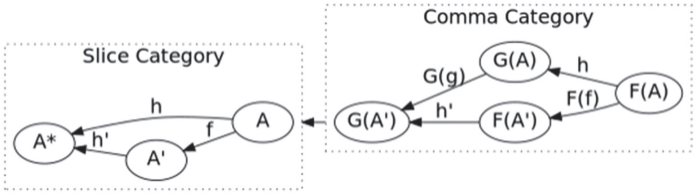


Figure 2 Schematic diagram of the slice category (slice category over A^*).

The morphism *g* denotes the relation of objects from *a* to *b*, and the composite morphism $a \xrightarrow{f \circ g} c$ is defined as $a \xrightarrow{g} b \xrightarrow{f} c$.

To illustrate how to represent the combination of machine learning by slice category, the basic concepts of slice category are given here. The slice category is a special case of the comma category.

A comma category refers to deriving a new category from three categories and a pair of functors associated with these categories, as shown in Figure 2. The comma category encompasses the functors *G*: $A \longrightarrow C$, the arrows *F*: $A' \longrightarrow C$, and the other arrows h and $h' \in C$. To define the comma category, we have a way of connecting the categories *A* and *A*' by looking at *F*(*A*), *G*(*A*) and the arrow *h*: $F(A) \longrightarrow G(A)$ between them.

The slice category is another way of getting new categories from the old, which is called slice category over A^*. Based on the comma category, the following conditions are specified: $C = A$, the functor F is the identity functor, and the category A' contains only one object * and a morphism. The slice category over A^* can be defined as: let *C* be a category, and *A* be a *C* object. The *C*-arrow *h* indicates $A \longrightarrow A^*$. The construction morphism f satisfies $h = f \circ h'$. The combination of machine learning can be interpreted in terms of slice categories. For example, the process of training data is considered as a *C*-arrow f: $A \longrightarrow A'$. The specific process is described in Section 3.

2.2 Functor

There are two kinds of operations done on the category by the functor, including the operations on objects and arrows [25]. The functor *F* is a

mapping relationship from one category to another, which preserves the structure between categories and the composition of morphism, including $F(f \circ g) = F(f) \circ F(g)$ and identities $F(\mathrm{id}) = \mathrm{id}_F$. In this research, the functor serves as the data preprocessing and model prediction.

3 The Category of Tabular Data Classification

The general workflow using the representation of category theory is displayed in Figure 3. In this section we utilize the monoidal category to define the combination of machine learning that includes data normalization, unbalanced data processing, model training, and prediction. We aim to create an explainable abstraction between the combination of machine learning.

3.1 Data Normalization and Unbalanced Data Processing

We first define the category of data normalization and unbalanced data processing. Given a dataset of input example X, we regard the z-score normalization as functor f and the normalization process can be expressed as f:X $\longrightarrow X'$; category X' is the normalized data. We can write this process

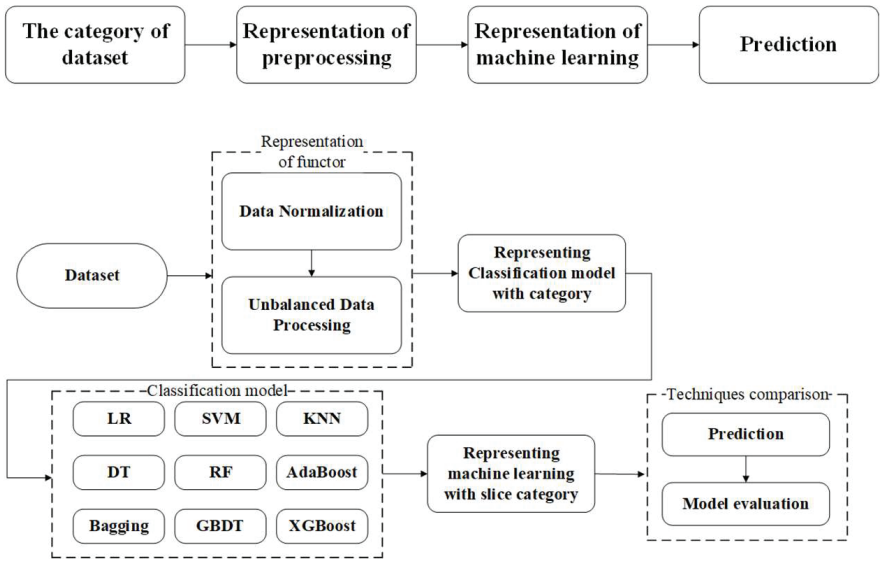


Figure 3 The workflow of proposed research.

as follows:

$$\left\{ f(\mathrm{x}, \mathrm{x}') = \frac{(x - x.\mathrm{mean})}{x.\mathrm{std}} \middle| \, x \in \mathrm{X}, \; \mathrm{x}' \in X' \right\} \tag{1}$$

where *x*.mean is the is the mean of the input data, and *x*.std is the standard deviation of the training samples.

After processing the data set for missing values and normalizing, The dataset is oversampled to produce category imbalance. The class distribution of the dataset used in this study is: probability for the label ">50K": 23.93%, probability for the label "<=50K": 76.07%, which indicates that the class distribution of data is unbalanced. The easiest way to oversample data is to replicate the instances from the few classes; however, these instances do not add any new information to the dataset. Instead, some new instances are synthesized from existing examples. The data augmentation by adding samples to minority classes is called the synthetic minority oversampling technique (SMOTE) [26]. The data must be of numeric type and have no missing value when using the SMOTE technique. After handling missing values and data normalization, we used three smote-based technologies for oversampling, including distance SMOTE [27], random SMOTE [28], and Gaussian SMOTE [29]. Given three SMOTE-based technologies, we have the functor *F*:

$$F = X \longrightarrow X'. \tag{2}$$

3.2 Model Representation by Slice Category

After preprocessing the input data, the slice category is used to represent the combination of machine learning. Figure 4 shows the canonical construction of model training and prediction from slice category and functors.

The combination of machine learning and its internal complexity is assumed to be a slice category. The *X* represents the category of raw input data, which is a set of examples. *F* is the functor used for the preprocessing of raw data. The preprocessed data and its internal relations form the

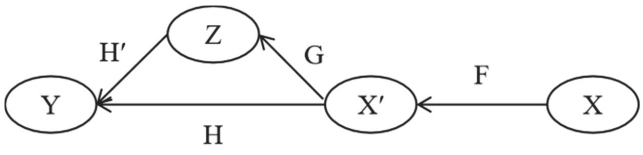


Figure 4 The representation of machine learning by using slice category.

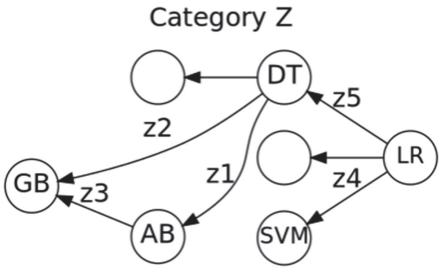


Figure 5 Sample models of the category *Z*.

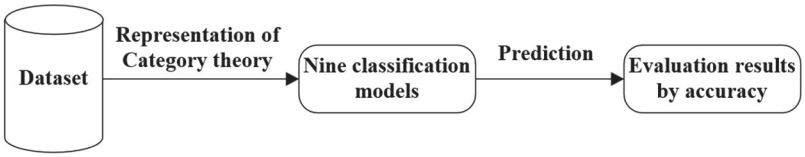


Figure 6 Evaluation of the models.

category X'. We assume that H: $X' \longrightarrow Y$ is a functor, and Y is the label category with the data type of classification. Based on the sample space and internal relations of the data $\in X'$, the training functor G is formed. It is advisable to define G as a functor in the slice category $G : X' \longrightarrow Z$, where X' is the input data after standardization and oversampling, and construct the slice category on Y. In addition, H': $Z \longrightarrow Y$ satisfies $H = H' \circ G$. The solving of functor H is challenging in computable time, but we can define the appropriate Z by constructing the slice category, and calculate the decision rule H' of the machine learning model based on the training functor G. The functor G can find the ideal decision rule H' more quickly on Z than on X.

Figure 5 shows a snapshot of the category Z. Indeed, a machine learning model such as logistic regression (LR) can be viewed as an object $\in$ category Z, where Z represents a set of machine learning models.

4 Experimental Results

In this section, we explicitly compute an example of implementing nine machine learning models on the adult dataset. The general flowchart of the experimental procedure is presented in Figure 6. We introduce the benchmark dataset for measuring the machine learning models. After representing the combination of machine learning by category theory, the classification performance of nine models is evaluated with accuracy.

4.1 Dataset and Architecture

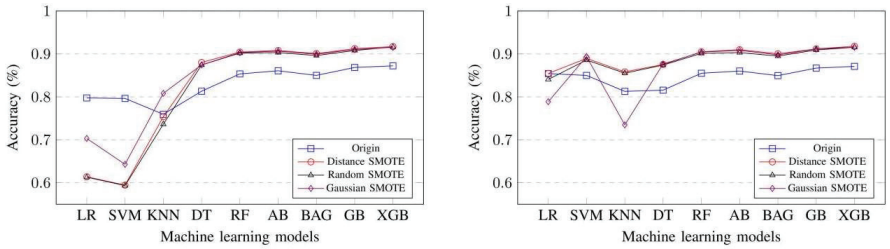
We use the adult dataset from the UCI Repository of Machine Learning Databases [30]. Table 1 displays some sample data. It contains 48,842 instances, and each instance consists of 6 continuous, 8 nominal attributes and 1 class attribute. The class attribute is recorded as a binary variable ("<=$50K": 0, ">50K": 1). 60% of the data in the dataset is used as the training set, and the remaining 40% is used to test the performance of the model.

4.2 Performance

In this section, the comparative evaluation of the logistic regression (LR), SVM, KNN, decision tree (DT), random forest (RF), AdaBoost (AB), bagging (BAG), gradient boosting (GB), and XGBoost (XGB) is presented. To test the tabular classification, we compare the nine models with category by using three oversampling techniques including distance SMOTE, random SMOTE, and Gaussian SMOTE. Figure 7(a) displays the classification evaluation result of raw data and normalized data with accuracy.

Table 1 Sample tabular dataset

Age	Workclass	fnlwgt	Education	Education-num	...	House-per-week	Native-country	Income
50	Self-emp-not-inc	83311	Bachelors	13	...	13	United-States	0
38	Private	215646	HS-grad	9	...	40	United-States	0
53	Private	234721	11th	7	...	40	United-States	0
28	Private	338409	Bachelors	13	...	40	Cuba	0
37	Private	284582	Masters	14	...	40	United-States	0



(a) Comparative evaluation on raw data.

(b) Comparative evaluation on normalized data.

Figure 7 Comparative evaluation of the models.

Table 2 Classification accuracy of different models on raw data

	Origin	Distance	Random	Gaussian
LR	0.797461	0.613881	0.612973	0.703102
SVM	0.796335	**0.594503**	**0.593393**	**0.642915**
KNN	**0.759431**	0.753701	0.736274	0.808135
DT	0.813175	0.880198	0.874209	0.874445
RF	0.853406	0.904320	0.901931	0.902940
AB	0.860675	0.907953	0.903008	0.906338
BAG	0.849977	0.900720	0.896212	0.899307
GB	0.868455	0.912192	0.907886	0.910207
XGB	**0.872242**	**0.917138**	**0.916229**	**0.915321**

Table 3 Classification accuracy of different models on normalized data

	Origin	Distance	Random	Gaussian
LR	0.854123	0.854495	0.840634	0.788723
SVM	0.849926	0.889517	0.885749	0.893756
KNN	0.812766	0.858162	0.855067	0.734861
DT	0.815632	0.875421	0.874008	0.876564
RF	0.854942	0.904858	0.901258	0.904152
AB	0.860112	0.909837	0.903075	0.908256
BAG	0.849414	0.900215	0.894967	0.897793
GB	0.866970	0.911755	0.909164	0.911351
XGB	**0.870963**	**0.917811**	**0.915422**	**0.915681**

The model obtained the highest results by using distance SMOTE in Table 2. It can be seen that the XGB model has the highest accuracy 0.917138. The XGB provides a higher overall accuracy, due to XGB using the second-order Taylor expansion, in which the prediction is closer to the true value [31]. In addition, DT, RF, AB, BAG, and GB also have significant improvements in accuracy. By contrast, the performance of KNN is unstable. LR and SVM obtained the worst result, lower than the origin.

Table 3 presents the accuracy comparison of standardized data. The results indicate that the XGB model had a better performance than the other models. The classification of the XGB model exhibits a good fit in the oversampling methods of distance SMOTE, where the value of accuracy is 0.917811. After standardizing, only a few cases achieved negative effects, and the performance of all models significantly improved, especially LR, SVM, and KNN.

5 Conclusion

In this research, category theory, a mathematical concept that manages the various parts of machine learning, is used to represent a machine learning combination. We explain the concepts of categories and functors, as well as how they can be used to represent data preprocessing and models. Furthermore, we identify the concept of slice category as the key distinguishing feature of this research; the slice category simplifies understanding and management of machine learning from a structured perspective. In a series of experiments, three oversampling methods based on SMOTE and nine machine learning models are combined to compare classification performance. The analysis reveals the compositional properties of machine learning.

The proposed framework in this paper corresponds to a prototype. In future work, the framework needs to be specific to "industrial" applications. Moreover, more category theory methods are introduced. In addition, for a representation of deep learning by category theory, choosing other concepts can provide a new, simple language for understanding and managing deep learning algorithms.

Acknowledgements

This research is supported by the Stable Support Project for Shenzhen Higher Education Institutions (SZWD2021011) and the Research Promotion Project of Key Construction Discipline in Guangdong Province (2022ZDJS112).

References

[1] M. Hofmann, R. Klinkenberg, *RapidMiner: Data Mining use Cases and Business Analytics Applications*. CRC Press, 2016.

[2] S. Ray, "A quick review of machine learning algorithms," in *2019 International conference on machine learning, big data, cloud and parallel computing (COMITCon)*. IEEE, 2019, pp. 35–39.

[3] X. Li, H. Liu, W. Wang, Y. Zheng, H. Lv, Z. Lv, "Big data analysis of the internet of things in the digital twins of smart city based on deep learning," *Future Generation Computer Systems*, vol. 128, pp. 167–177, 2022.

[4] I. Y, Ko, A. Srivastava, M. Mrissa, "Scalable and dynamic big data processing and service provision in edge cloud environments," *Journal of Web Engineering*, vol. 21, no. 1, 2022.

[5] S. Nickolas, K. Shobha, "Efficient pre-processing techniques for improving classifiers performance," *Journal of Web Engineering*, vol. 21, no. 2, 2022.

[6] N. Deepa, Q.-V. Pham, D. C. Nguyen, S. Bhattacharya, B. Prabadevi, T. R. Gadekallu, P. K. R. Maddikunta, F. Fang, P. N. Pathirana, "A survey on blockchain for big data: Approaches, opportunities, and future directions," *Future Generation Computer Systems*, Vol. 131, pp. 209–226, 2022.

[7] G. Alves, M. Amblard, F. Bernier, M. Couceiro, A. Napoli, "Reducing unintended bias of ml models on tabular and textual data," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2021, pp. 1–10.

[8] C. Zhu, Q. Zhang, L. Cao, A. Abrahamyan, "Mix2vec: Unsupervised mixed data representation," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2020, pp. 118–127.

[9] M. Hofmann, R. Klinkenberg, *RapidMiner: Data Mining Use Cases and Business Analytics Applications*. CRC Press, 2016.

[10] W. Yi, I. V. Gerasimov, S. A. Kuzmin, H. He, "A category theory approach to phytometric system conceptual modeling," in *2018 IEEE Conference of Russian Young Researchers in Electrical and Electronic Engineering (EIConRus)*. IEEE, 2018, pp. 391–393.

[11] W. Lu, X. Jiang, X. Liu, Q. Qi, P. Scott, "Modeling the integration between specifications and verification for cylindricity based on category theory," *Measurement Science and Technology*, vol. 21, no. 11, p. 115107, 2010.

[12] J. Culbertson, K. Sturtz, "Bayesian machine learning via category theory," *arXiv preprint arXiv:1312.1445*, 2013.

[13] G. S. Cruttwell, B. Gavranovic´, N. Ghani, P. Wilson, F. Zanasi, "Categorical foundations of gradient-based learning," in *European Symposium on Programming*. Cham: Springer, 2022, pp. 1–28.

[14] K. Kamiya, J. Welliaveetil, "A category theory framework for Bayesian learning," *arXiv preprint arXiv:2111.14293*, 2021.

[15] D. Shiebler, B. Gavranoviæ, P. Wilson, "Category theory in machine learning," *arXiv preprint arXiv:2106.07032*, 2021.

[16] D. W. Hosmer Jr, S. Lemeshow, R. X. Sturdivant, *Applied Logistic Regression*. John Wiley & Sons, 2013, vol. 398.

[17] J. Chen, Y. Yin, L. Han, F. Zhao, "Optimization approaches for parameters of svm," in *Proceedings of the 11th International Conference on Modelling, Identification and Control (ICMIC2019)*. Springer, 2020, pp. 575–583.

[18] C. Kennedy and J. Griffith, "Using markup language to differentiate between reliable and unreliable news," in *2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA)*. IEEE, 2020, pp. 619–625.

[19] H. Guan, Y. Zhang, B. Ma, J. Li, C. Wang, "A generalized optimization embedded framework of undersampling ensembles for imbalanced classification," in *2021 IEEE 8th International Conference on Data Science and Advanced Analytics (DSAA)*, 2021, pp. 1–10.

[20] R. S. Khairy, A. S. Hussein, H. ALRikabi, "The detection of counterfeit banknotes using ensemble learning techniques of adaboost and voting," *International Journal of Intelligent Engineering and Systems*, vol. 14, no. 1, pp. 326–339, 2021.

[21] M. M. Ali, R. Anwar, A. F. Yousef, B. Li, A. Luvisi, L. D. Bellis, A. Aprile, F. Chen, "Influence of bagging on the development and quality of fruits," *Plants*, vol. 10, no. 2, p. 358, 2021.

[22] T. Zhang, W. He, H. Zheng, Y. Cui, H. Song, and S. Fu, "Satellite-based ground $PM_{2.5}$ estimation u sing a gradient boosting decision tree," *Chemosphere*, vol. 268, p. 128801, 2021.

[23] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 785–794.

[24] J. A. Goguen, "A categorical manifesto," *Mathematical Structures in Computer Science*, vol. 1, no. 1, pp. 49–67, 1991.

[25] N. Tsuchiya, S. Taguchi, H. Saigo, "Using category theory to assess the relationship between consciousness and integrated information theory," *Neuroscience Research*, vol. 107, pp. 1–7, 2016.

[26] A. Fernández, S. Garcia, F. Herrera, N. V. Chawla, "SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary," *Journal Of Artificial Intelligence Research*, vol. 61, pp. 863–905, 2018.

[27] J. de la Calleja, O. Fuentes, "A distance-based over-sampling method for learning from imbalanced data sets," in *Proceedings of the Twentieth International Florida Artificial Intelligence*, vol. 3, 2007, pp. 634–635.

[28] Y. Dong, X. Wang, "A new over-sampling approach: Random-SMOTE for learning from imbalanced data sets," in *Knowledge Science, Engineering and Management*, eds H. Xiong and W. B. Lee, Berlin, Heidelberg: Springer, 2011, pp. 343–352.

[29] H. Lee, J. Kim, S. Kim, "Gaussian-based smote algorithm for solving skewed class distributions," *Int. J. Fuzzy Logic and Intelligent Systems*, vol. 17, pp. 229–234, 2017.

[30] R. Eberhart, J. Kennedy, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neural Networks*, vol. 4. Citeseer, 1995, pp. 1942–1948.

[31] C. Wang, C. Deng, S. Wang, "Imbalance-xgboost: Leveraging weighted and focal losses for binary label-imbalanced classification with xgboost," *Pattern Recognition Letters*, vol. 136, pp. 190–197, 2020.

Biographies

Heng Zhao is an assistant professor, College of Big Data and Internet, Shenzhen Technology University, and M.Sc. Supervisor, Shenzhen University. His research interests include Big Data modeling, machine learning, etc.

Yixing Chen is a graduate student in the College of Computer Science and Software, Shenzhen University, Shenzhen, China. His research direction is machine learning.

Xianghua Fu is currently a professor of the College of Big Data and Internet, and Vice Dean of the College of Big Data and Internet, Shenzhen Technology University. His research interests include natural language processing, information retrieval, machine learning and data mining, etc.