Journal of Network Intelligence
Taiwan Ubiquitous Information
 ISSN 2414-8105
Volume **9**, Number **3**, August **2024**

Classification of Unbalanced Data Based on Bayesian Optimal Neural Network Model

Li-Zhu Ye*

School of Information Technology and Engineering
Guangzhou College of Commerce, Guangzhou 511363, China
School of Graduate Studies
Management and Science University, 40100 Shah Alam, Selangor, Malaysia
20180160@gcc.edu.cn

Mohammed Hazim Alkawaz

Department of Computer Science
College of Education for Pure Science
University of Mosul, Mosul 41002, Iraq
mohammed.ameen@uomosul.edu.iq

Md Gapar Md Johar

Software Engineering and Digital Innovation Center
Management and Science University, 40100 Shah Alam,Selangor, Malaysia
mdgapar@msu.edu.my

*Corresponding author: Li-Zhu Ye
Received November 3, 2023, revised March 13, 2024, accepted June 6, 2024.

Abstract. *Aiming at the unbalanced problem of big data classification, this paper proposes a classification method of unbalanced data on the ground of Bayesian optimal neural network model. The method firstly uses Optimized Graph Convolutional Network (OGCN) network to carry out message passing mechanism in data graph, designs neural network model with self-attention pooling and mean maximization pooling, and then uses Bayes optimization to update the posterior probability of optimization function by constantly updating the probability model and evaluating the objective function. Effectively solve the problem of speed and optimal solution of hyperparameter optimization in the neural network model. Secondly, the model is used for data set training to obtain the weight of the entire network. Secondly, the unbalanced data is mixed sampling to reduce the number of negative samples. Finally, the capability of the model was verified on the CiteSeer dataset. Experimental results showed that the accuracy rate, accuracy rate and recall rate of the proposed algorithm were 0.974, 0.982 and 0.973, respectively, which effectively improved the accuracy rate, accuracy rate and recall rate of unbalanced data classification compared with other algorithms.*
Keywords:unbalanced data; classification method; neural network; bayes optimization; mixed sampling

1. **Introduction.** With the maturity of artificial intelligence technology, people's lives are changing with each passing day and unprecedented changes have taken place. The era of big data is producing a huge amount of data information at all times, such as recruitment exams, emails, meituan comments, etc. All kinds of data information fill every corner of our lives. With the rapid growth of the number of data information, the ability of manual data processing is powerless, and the unstructured data information occupies

the majority of the total amount of data information. The digital information age will face a severe dilemma of massive data processing. Therefore, the rapid classification of current data information is an important topic in current social research. At present, there are many mature classification algorithms, such as Decision Tree (DT)[1], Naive Bayes (NB)[2], Artificial Neural Network (ANN)[3], etc. These classification algorithms have been successfully applied to human production, life, transportation and other fields. However, as technology developing, the application of classification algorithms has gradually expanded to other applications, such as security, medicine and other fields [4, 5, 6]. A common feature of many data in these fields is the unbalance of data. How to improve the classification performance of unbalanced data has significant practical importance and academic value [7, 8].

1.1. **Related Work.** Han et al. [9] put forward Borderline-SMOTE algorithm, which can solve the problem of further expansion of noise and improve the performance of SMOTE algorithm to some extent. However, the performance of the algorithm depends on the nearest neighbor K value, if the K value is too small, some minority class samples will be mistaken as noise, and this part of the sample can not generate new minority class samples. Later, Bunkhumpornpat et al. [10] suggested the Distance-based Synthetic Minority Oversampling Technique (DBSMOTE) algorithm. This algorithm is mainly on the ground of the concept of density clustering to generate new minority class samples on the boundary minority class samples. Ozenne et al. [11] proposed a two-stage Majority Weighted Minority Oversampling Technique (MWMOTE) algorithm, but this method could not detect important minority samples far from majority classes. Bae and Yoon [12] proposed an integration framework, which used technology and PLS feature learning method to detect polyp through data integration. Alibeigi et al. [13] proposed a method called Density Based Feature Selection (DBFS), which deals with high-dimensional unbalanced data through feature sorting. Beyan and Fisher [14] suggested a hierarchical decomposition method to deal with class imbalance, which integrated clustering and discrete point detection techniques. Zou et al. [15] suggested an innovative framework to improve classification performance in unbalanced classification problems by finding the optimal classification threshold. Yu et al. [16] suggested an innovative method based on compensation strategy and Extreme Learning Machine (ELM) method to deal with unbalanced classification problems. Cheng et al. [17] proposed the cost-sensitive Large margin Distribution Machine (CS-LDM) algorithm, and the CS-LDM method introduced the Cost Sensitive interval mean and cost sensitive penalty term. Krawczyk et al. [18] proposed an integrated approach on the ground of cost-sensitive decision trees and evolutionary algorithms. On the ground of fuzzy rough set theory, Lopez-Martin et al. [19] proposed a new classification method for multi-instance class unbalance classification, which mainly increased the weight adjustment of a few classes and reduced the weight adjustment of most classes. The results displayed that this method could be effectively applied to the unbalance classification problem. Zieba et al. [20] declared the Neighbourhood Cleaning Rule (NCL), which improved the editing nearest neighbor rule. In order to preserve the data distribution characteristics of most classes during undersampling, clustering based undersampling is a common technique. Lin et al. [21] proposed an under-sampling method based on K-means clustering. First, most classes are clustered (the number of classes is the number of samples of minority classes), and then the cluster center of each cluster or the samples closest to the cluster center are selected to form a new sample set of majority classes, so as to balance majority classes and minority classes. Zheng et al. [22] proposed a method to select unbalanced text features by adjusting the combination of positive and negative features in the data. Chen et al. [23] proposed a new

feature selection method based on class decomposition, which divides large categories into relatively small pseudo-subclasses and generates pseudo-class labels accordingly. Then the feature goodness measure is calculated on the new decomposed data, and the feature selection is performed accordingly. Dubey et al. [24] suggested a feature selection method on the ground of local importance and global importance. However, the false positive rate of this method is relatively high. For the purpose of solving this problem, Chen et al. [25] proposed feature selection of unbalanced data on the ground of neighborhood rough sets, but there was a problem that it could not be accurately classified.

1.2. **Motivation and contribution.** It can be seen that unbalanced data classification algorithms still face challenges such as class overlap, lack of representation ability of a few classes and lack of supervision information. Therefore, this article suggests a classification method of unbalanced data on the ground of Bayesian optimization neural network model to solve the problems of unbalanced and low classification accuracy of traditional structured data. First, this method adds the attention unit optimization neural network model on the basis of gate graph neural network GGNN. Then, the optimal combination of model hyperparameters is obtained by means of Bayesian optimization by constantly updating the probability model and performing the objective function. At last, the unbalanced data is mixed sampled to reduce the number of negative samples and improve the performance of unbalanced data classification.

2. Relevant theoretical analysis.

2.1. **Bayesian optimization.** Bayesian optimization is an effective global optimization algorithm, which essentially belongs to a probabilistic model and is one of the advanced technologies in the area of probabilistic machine learning and artificial intelligence [26]. It is a model-based sequential optimization method, which finds the next evaluation position according to the information obtained from the unknown objective function, so as to reach the optimal solution fastest. Bayes theorem is used in the optimization process.

$$p(g|C_{1:s}) = \frac{p(C_{1:s}|g)p(g)}{p(C_{1:s})} \tag{1}$$

$$C_{1:s} = \{(x_1, y_1), (x_2, y_2), ..., (x_s, y_s)\} \tag{2}$$

$$y_s = g(x_s) + \vartheta_s \tag{3}$$

In Equations (1), (2) and (3), C_{1s} is the observed set; x_s is the decision vector; y_s is the observed value; s is the number of observation sets (the number of iterations); g_s is the observation error; $p(C_{1s} \mid g)$ is the likelihood distribution of y; $p(g)$ is the prior probability distribution of the unidentified cost function g; $p(C_{1s})$ is the marginal likelihood distribution of g, which is used to optimize hyperparameters; $p(g \mid C_{1s})$ is the posterior potentiality division of g, and the posterior potentiality division indicates the reliability of the anonymous objective division after dealing with the earlier division. Bayesian optimization includes two core parts: probabilistic agent model and acquisition function.

Probabilistic agent model. Gaussian Processes (GPs) are used as a probabilistic proxy model, and Gaussian processes are widely used in regression, classification, and many fields where black-box functions need to be inferred. A Gaussian process consists of a mean function $n(x)$ and a semi-definite covariance function $h(x, x')$:

$$g(x) \cong f(n(x), h(x, x')) \tag{4}$$

$$n(x) = E[g(x)] \tag{5}$$

$$h(x, x') = E[(g(x) - n(x))(g(x') - n(x'))] \tag{6}$$

From the prior distribution and likelihood distribution formulas, the marginal likelihood distribution is obtained:

$$p(y|X,\sigma) = \int p(y|g)p(g|X,\sigma)\mathrm{d}f = N(0,\sum +\delta^2 J) \tag{7}$$

The hyperparameter δ is usually optimized by maximizing this marginal likelihood distribution. According to the nature of the Gaussian process, there is the following joint distribution:

$$\begin{bmatrix} y \\ g* \end{bmatrix} \cong N\left\{0, \begin{bmatrix} \sum +\delta^2 JL* \\ L_*^S L** \end{bmatrix}\right\} \tag{8}$$

By explicitly specifying the prior mean function $n(x)$, the expression of prior information can be facilitated while increasing the explanatory nature of the model, when the prediction covariance does not change, and the predicted mean becomes:

$$\prec g* \succ = n(X_*) + L_*^S[\sum +\delta^2 J]^{-1}(y - n(X_*)) \tag{9}$$

2.2. **Graph Convolutional Network.** Graph convolutional neural networks (Graph Convolutional Network (GCN)[27] was officially announced in 2016 It is proposed that it has been widely used in various tasks in recent years, and the effect is very remarkable. Most deep learning models are on the ground of Euclidean space data, and the data structure is relatively regular, while irregular data structures such as chemical molecules and structures need to be processed by models such as GNN. The core idea of GCN is to use the characteristics of the aggregation node and the characteristics of the neighboring node set to update the feature vector of the node, and the aggregation operation is similar to the convolution operation in CNN. But the parameters of the convolution kernel in GCN are learned by learning. The essential feature of GCN is that the current node representation is updated by updating the information of neighboring nodes by performing convolution operations on the graph structure.

Given a set of text graph data with M nodes, each node has unique characteristics. The characteristics of the nodes in the figure form an $M \times M$-dimensional matrix X, and the node relationship constitutes an $M \times N$-dimensional adjacency matrix B, which forms the input to the model. Then, the propagation between the layers in the graph convolutional network GCN is expressed as Equation (10).

$$K^{I+1} = g(K^I, B) = \varepsilon(\tilde{C}^{-\frac{1}{2}}\tilde{B}\tilde{C}^{-\frac{1}{2}}K^I V^I) \tag{10}$$

where $\tilde{B}$ stands for $B + I$ and I is the identity matrix; g is the feature of each layer, and the feature in the input layer H is the matrix X; $\tilde{C}$ is the degree matrix of $\tilde{B}$, and the equation is $\tilde{C}_{jj} = \sum i\tilde{B}_{ji}$; ε is a nonlinear activation function.

In GCN, the representation vector of each layer can be seen as obtained by weighting the average of the representation vectors of neighboring nodes. Use the averaging function for each node in the graph to input the final calculated mean into the network model for training. The GCN model can effectively use the edge information between nodes to extract features, and the technology in text classification can help the model better capture the semantic relationship and text structure between words, so as to achieve good performance on graph data.

3. Neural network model design.

3.1. **Overall structure of model.** Firstly, the pre-trained model is used to train text data, and the trained text vector representation is used as the feature vector representation of nodes in the data graph. Take the data graph as input and use the optimized neural network model (OGCN). Among them, the OGCN model is composed of attention units added to the gated graph neural network GGNN.

With the help of the OGCN network, a messaging mechanism can be made in the data graph to update the characteristics of the nodes. After training, the node representations trained by the two models are spliced and fused, and finally self-attention pooling and mean maximization pooling are adopted. The combined way to learn the graph representation is fed into the data classifier to complete label classification. The OGCN method is mainly composed of three parts, and the whole classification method can realize the label classification of the dataset, including data set pre-training and constructing text graphs, Messaging and representation and reading out of graphs. The overall structure of the OGCN method is shown in Figure 1.

3.2. **Neural network data graph construction.** To construct a data graph is to transform the data into a combination of nodes and edges to form a graph structure, so as to facilitate the algorithm and analysis on the ground of the graph. The constructed data graph is shown in Figure 2.

This paper constructs an undirected text graph for data set $D = \{g_1, g_2, ..., g_j\}$ and predicts the classification label $\bar{y}$ belonging to the data. In this paper, in the process of building the target data into a data graph, the input target data D_j builds an independent graph structure H_i, and the number of data in the database is the same as that of the constructed data graph.

In this paper, $H =< W, F >$ is used to represent the constructed data graph, where $W(|W| = m)$ represents all nodes in graph H, F represents all edges in graph H, $w_j \in W$ represents a certain node in graph H, $f_{ji} =< w_j, w_i > \in F$ represents a certain edge of nodes w_j to w_i in graph H, and $|W| = m$ represents the number of nodes in the graph. $G \in \mathbb{R}^{m \times c}$ is the node eigenmatrix in figure H, $g_j \in \mathbb{R}^c$ is the eigenvector of node j in figure H, where c is the dimension of the eigenvector g_j. $Y \in \mathbb{R}^{m \times c}$ represents the eigenmatrix of graph H, and the dimension represented by a single node vector is C-dimension. The initialization hidden state g^0 of the data graph H is set by the feature matrix Y, where P represents the updated state of each time step t during the training process of the model. $c \in \mathbb{R}^{m \times m}$ is the degree matrix in figure H, where $C_{jj} = \sum_i B_{ji}$, and $\hat{B} = C^{-1/2} B C^{-1/2}$ are the symmetric normalized adjacency matrix. $B \in \mathbb{R}^{m \times m}$ is the adjacency matrix in figure H, and the value of element B_{ji} in the adjacency matrix is shown in Equation (11).

$$B_{j,i} = \begin{cases} 1, if < w_i, w_j > \in D \\ 0, otherwise \end{cases} \tag{11}$$

In the data graph, for each data, all the following data are traversed and the co-occurrence frequency between them is calculated. The formula for calculating the PMI value between data node y and data node x is shown in Equation (12):

$$PMI(y, x) = \log \frac{p(y, x)}{p(y) \times p(x)} \tag{12}$$

Thereinto, $p(y, x)$ represents the probability of two data appearing simultaneously in the same window, $p(y)$ and $p(x)$ represents the probability of node y and node x respectively. For each co-occurrence relationship, an edge is added between the corresponding two data nodes and the weight is their co-occurrence frequency. The constructed data graph is shown in Figure 2.

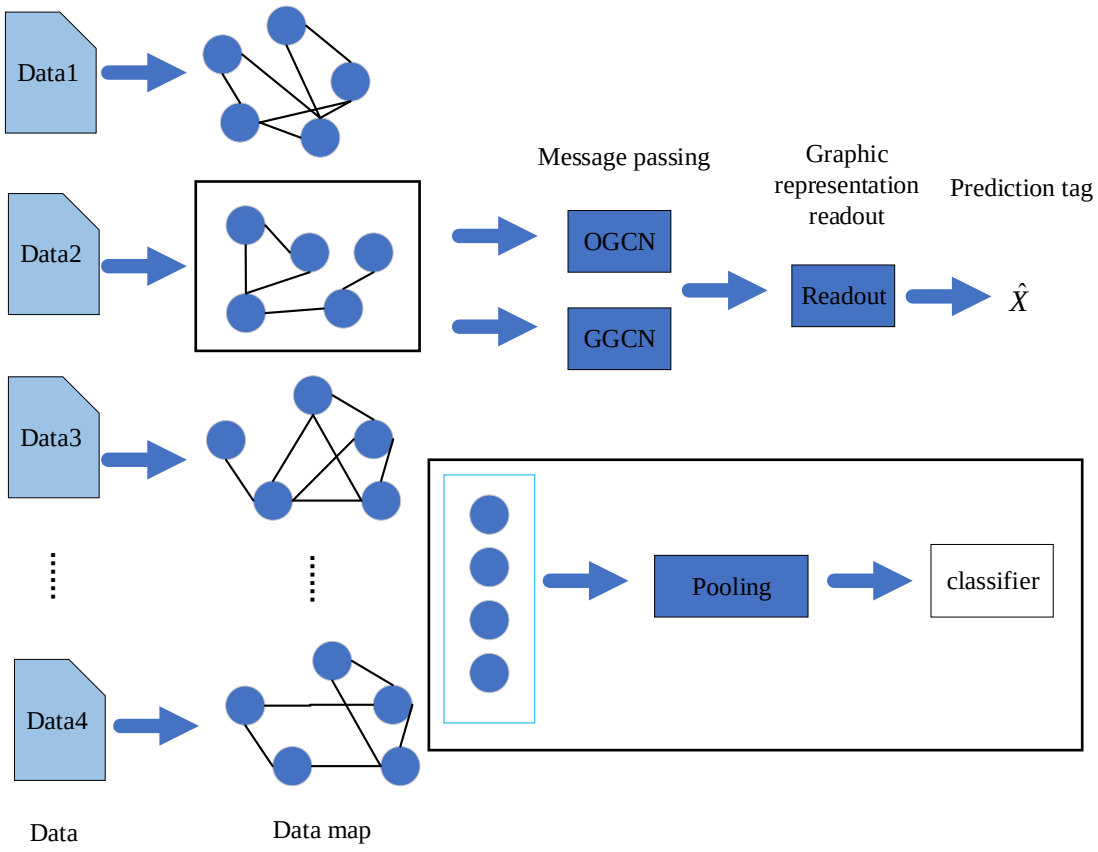

Figure 1. Structure diagram of neural network model

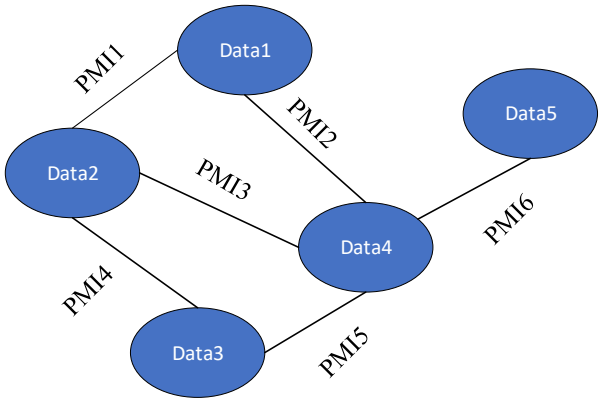

Figure 2. Data graph

3.3. **Data transfer and readout.** In the message passing stage, when the GCN model is used to train the data, it ends when the number of training iterations reaches 100, or when the loss value of the training verification set does not decrease for 20 consecutive iterations, and the node feature representation after training is obtained. Finally, the node vector representation learned by the GCN model is fused and spliced, and then represented by the pooling operation learning graph. After the training of the two models, the node vectors are respectively expressed as: $G^T = [g_1, g_2, \ldots, g_m]$, $G^H = [m_1, m_2, \ldots, m_m]$, and the two vector representations are fused and spliced to obtain the node feature vector representing G. The splicing process is shown in Equation (13).

$$G = [G^T; G^H] = [g_1, g_2, \ldots, g_m] \tag{13}$$

Then the learned graph representation vector G_H is input into the classifier to predict the classification label $\hat{X}$, the process is shown in Equation (14). Where C identifies the bias vector and W represents the linear transformation matrix.

$$\hat{X} = softmax(WG_H + C) \tag{14}$$

Finally, the loss of the real label is minimized by the cross-entropy function, which is shown in Equation (15).

$$G = -\sum_j x_j H_j \log(H_j \hat{X}) \tag{15}$$

4. **Classification of unbalanced data based on Bayesian optimal neural network model.**

4.1. **Data preprocessing and Bayesian optimization parameters.** On the ground of the above neural network graph construction, to solve the problem of imbalance between data and avoid the influence of most class samples on classification in overlap area, this paper proposes an unbalanced data classification algorithm based on Bayesian optimized neural network model. Firstly, the standardized method is used to preprocess the data, and then Bayesian optimization is used to update the posterior potentiality of the optimization division by continuously updating the potentiality model and assessing the objective division, so as to obtain the optimal combination of model hyperparameters, which can effectively solve the algorithm hyperparameters secondly, the neural network model designed in the previous section is used for dataset training to obtain the weights of the whole network, and finally the unbalanced data is mixed sampling to dwindle the number of negative samples and refine the precision of classification.

Assuming that the original sample data has m variables, the training data set can be represented as: $D = \{(y_1, x_1), (y_2, x_2), ..., (y_m, x_m)\}$, where $(y_j, x_j), j = 1, 2, ..., m$ is called the sample or sample point. $y_j \in \gamma$ represents the feature of the j-th sample, $\gamma \subset R^m$ is the m-dimensional feature space, $x_j \in \chi = \{1, 2, ..., D\}$ represents the category of the j-th sample. A classification when $D \geq 3$ is called a multiclassification, and $D = 2$ is called a binary classification when $\chi = \{0, 1\}$ is called a class set, where 0 represents the majority class and 1 represents the minority class. In view of the unbalanced classification problem, the sample set of a few classes in D is defined as D_{min}, the sample set of most classes is D_{max}, and the sample number of a few classes is much smaller than that of most classes $|D_{min}| \leq D_{max}$. The imbalance ratio is defined as: $JS = \frac{D_{max}}{D_{min}}$.

After normalization, most of the sample data is clustered at the origin, and although the data is also scaled, the covariance of the sample data remains constant, as shown in Equation (16). Covariance is of great significance in data processing, as it can measure the statistical relationship between two random variables, and standardized data can keep the covariance constant, making the data more fidelity.

$$\mathrm{cov}(\gamma, \chi) = \mathrm{cov}(\gamma D, \chi D) \tag{16}$$

Then the hyperparameter combination of the Bayesian optimization model is $y^* = \arg\min_{\gamma \in Y} g(\gamma)$, where $g(\gamma)$ is the objective function, $\min g(\gamma)$ is the minimum value of the function, and y^* is the optimal hyperparameter combination required. The updated posterior probability is optimized to $P(F|C) \propto P(C|F)P(F)$, where $P(F)$ is a Gaussian distribution and

$P(C|F)$ is a Gaussian regression process. This can be determined by the nuclear matrix Σ of Equation (17).

$$\sum = \begin{bmatrix} l(y_1, y_1) & \ldots & l(y_1, y_m) \\ \ldots & \ldots & \ldots \\ l(y_m, y_1) & \ldots & l(y_m, y_m) \end{bmatrix} \tag{17}$$

The value obtained from the pre-processed data $y_j(s)$ through feature extraction and feature fusion is used as the network input.

4.2. **Classification of unbalanced data based on Bayesian optimal neural network model.** On the ground of the above neural network graph construction, to solve the imbalance between data and avoid the influence of most class samples on classification in overlap area, this paper proposes an unbalanced data classification algorithm on the ground of Bayesian optimized neural network model. Firstly, the standardized method is used to preprocess the data, and then Bayesian optimization is used to update the posterior potentiality of the optimization division by continuously updating the probability model and evaluating the objective division, so as to obtain the optimal combination of model hyperparameters, which can effectively solve the algorithm hyperparameters secondly, the neural network model designed in the previous section is used for dataset training to obtain the weights of the whole network, and finally the unbalanced data is mixed sampling to refine the number of negative samples and enhance the precision of classification.

Use the parameter η to represent the target number of mixed samples, it is almost equivalent to the amount of data in the middle class. The data of the middle class represents the grade of balance. In a detailed and exact way, delimit η as shown in Equation (18).

$$\eta = \left\lceil \frac{\max\{|\upsilon_{Kj}|\} + \min\{|\upsilon_{Kj}|\}}{2} \right\rceil \tag{18}$$

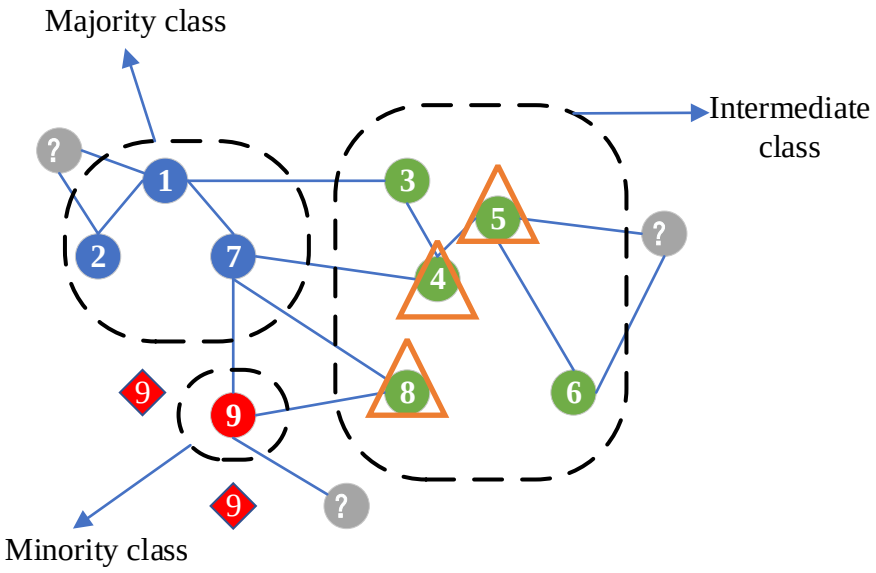

Figure 3. Mixed sampling graph

where $\lceil y \rceil$ represents the integer up operation. For each ν_{Kj}, it is decided which sampling method to use based on the relationship between its base $|\nu_{Kj}|$ and η. The mixed sampling scheme is as follows.

If $|\nu_{Kj}| > \eta$, ν_{Kj} is considered the majority class, and is denoted as ν_{Kjnb1}. For each ν_{Kjnb1}, choose one of the η nodes such as the green node with triangle in Figure 3. The set of selected η nodes is denoted ν'_{Kjnb1}.

If $|\nu_{Kj}| < \eta$, ν_{Kj} are considered to be minority classes and are denoted as ν'_{Kjnb1}. In view of the above analysis of oversampling, this paper uses the information perturbation (i.e., information enhancement) means to advance the amount of minority class nodes. Node-specific level enhancement is a common method of graph data enhancement, which adds noise to node attributes by using 0 random shielding part in node features. For $\nu_j \in \nu'_{Kjnb1}$, a part of the dimension mask of its feature vector s_j is randomly selected to generate some new nodes as its subsidiary nodes. Formally, a mask vector $\tilde{n} \in \{0,1\}^c$ is first generated. Then, the features $\tilde{s}_j$ of the newly generated node are computed with $\tilde{s}_j = s_j \times \tilde{n}$. The set ν_{Kjnb1} is made up of a small number of class nodes which are red in Figure 3 and their subsidiary nodes are the upsampled node set, denoted as ν'_{Kjnb1}.

If $|\nu_{Kj}| = \eta$, it is considered an intermediate class and remains unchanged. After a balanced set is obtained from the above mixed sampling, the loss function can be expressed as the Equation (19).

$$L = \frac{1}{n}\sum_{j=1}^{n}\frac{-1}{|P(j)|}\sum_{p\in P(j)}\log\frac{\exp(sim(S_j,S_P)/\tau)}{\sum_{b\in Ba(j)}\exp(sim(S_j,S_P)/\tau)} \tag{19}$$

where $Ba(J)$ is the balanced set $a(j)$ obtained by mixed sampling.

The loss function includes nodes to obtain discriminative representations through encoding. In addition to this, we also have the conventional cross-entropy loss of data classification.

$$L_{DF} = -\sum_{j=1}^{M}\sum_{i=1}^{n} x_{ji} In\widetilde{x_{ji}} \tag{20}$$

5. **Algorithm performance testing and analysis.**

5.1. **Performance comparison.** For the purpose of verifying the performance of the algorithm designed in this paper, this paper experiments are carried out on the algorithm and other existing algorithms, and all experiments are completed on the Python platform in personal computers. The computer configuration is as follows: Windows 10 operating system, 8GB RAM and Intel I5 processor. The experimental big dataset is an unbalanced version of the well-known CiteSeer dataset [28], which consists of more than 2million datasets 20 training and test samples for prediction attributes. For ease of analysis, literature [29] is denoted as LUA, literature [30] is denoted as CSG, and literature [31] is denoted as SCI, and the algorithm in this article is denoted as OUR.

The model performance is first verified on the CiteSeer dataset. figure1 shows the accuracy rate, accuracy rate, recall rate and F1 values obtained in the final test set using different types of models on CiteSeer data set, where F1 value is shown in Equation (21).

$$F1 = \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \times 2 \tag{21}$$

According to the analysis, compared with LUA, CSG and SCI models, the performance of OUR model has improved in the four evaluation indexes of accuracy, accuracy, recall and F1 value. Accuracy has been improved by 0.153, 0.059 and 0.081 respectively; Precision

by 0.125, 0.088 and 0.066 respectively; Recall by 0.139, 0.072 and 0.055 respectively; F1 by 0.132, 0.08, 0.06 respectively. This model performs better than the twin neural network-based news text classification method on the four indexes. The LUA and CSG models have in common that they are variants of the RNN model. The difference is that SCI cannot encode information from back to front, whereas LUA can capture both front to back and back to front information. OUR model has better performance than LUA, CSG and SCI.

Table 1. Comparison of different models

Model	Accuracy	Precision	Recall	F1
LUA	0.821	0.857	0.834	0.845
CSG	0.915	0.894	0.901	0.897
SCI	0.893	0.916	0.918	0.917
OUR	0.974	0.982	0.973	0.977

To further verify the optimization effect of Bayesian optimization method on the model, the performance of Bayes + models was tested. OUR Accuracy reached 0.978, Precision reached 0.987, Recall reached 0.976 and F1 reached 0.981. Compared with the original model, it is improved by 0.004, 0.005, 0.003 and 0.004 respectively. This model has the best performance compared with other models in classification results. On the one hand, thanks to Bayesian optimization parameters, the effective information of data category is further increased, and the model obtains excellent classification performance through training; On the other hand, in the structure of the model, the attention mechanism and information interaction are closely combined to further explore the deep relationship between data categories and improve the accuracy of data classification. On this basis, the neural network system is further used to improve the generalization ability of the model, avoid the occurrence of overfitting problems, and further enhance the performance of data classification. The specific experimental outcome is indicated in Figure 4.

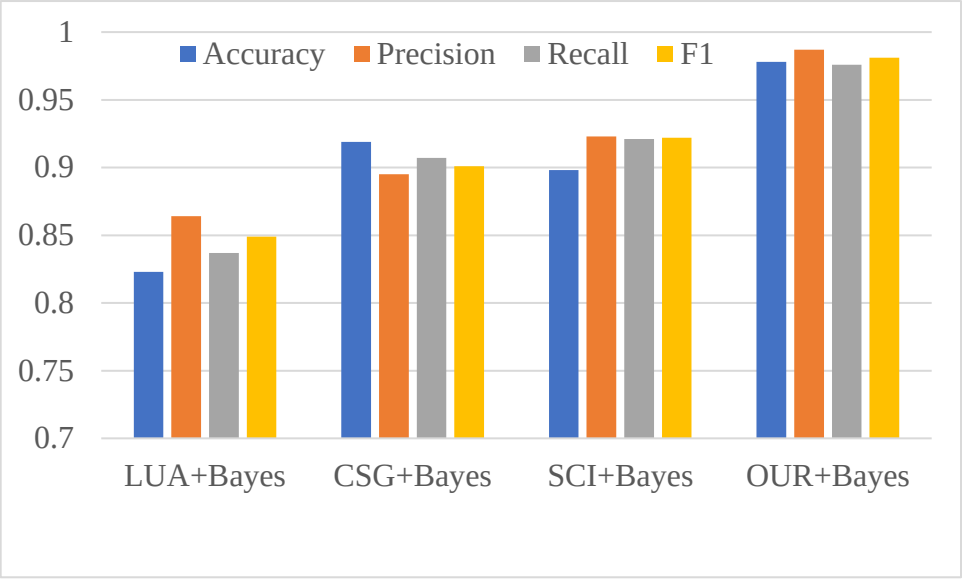

Figure 4. The performance comparison of algorithms after adding Bayesian optimization

5.2. **Ablation experiment.** For the purpose of proving the role of mixed sampling in OUR model, an ablation experiment was performed. Table 2 shows the Accuracy, Precision and Recall values of the ablation experiment conducted after LUA, CSG, SCI and OUR were integrated into the Bayesian optimization method on the CiteSeer dataset. It can be seen from the results in the table that OUR method proposed in this chapter has better results under the four evaluation indicators of CiteSeer data set. This shows that mixed sampling is necessary in OUR method and is effective for solving the unbalanced node classification problem. In addition, the outcome also indicate that OUR is markedly better than LUA, CSG and SCI. Because the accidental undersampling greatly diminishes the all amount of negative samples, which is harmful to contrast studying. A small number of class nodes achieved by accidental oversampling will cause overfitting, resulting in unsatisfactory classification results.

Table 2. Comparison of indicators of different models

Model	**Accuracy**	**Precision**	**Recall**
LUA+Bayes+XR	0.859	0.872	0.851
CSG+Bayes+XR	0.947	0.916	0.925
SCI+Bayes+XR	0.924	0.935	0.937
OUR+Bayes+XR	0.981	0.992	0.988

5.3. **Experiment on the influence of different imbalance ratios.** For the purpose of exploring the influence of unbalance ratio on OUR method, we conducted validation experiments on different algorithms with different unbalance ratio on CiteSeer dataset. The experiment was performed as follows: We fixed the amount of educating data for most classes of datasets to 100, set the imbalance ratio to 20, 16, 12, 8, 4, and the degree of imbalance ranged from severe to slight. Each experiment was repeated 10 times, and the average AUC (area under ROC) and AUPRC (area under accurate recall curve) values were shown in Figure 5 and Figure 6, respectively. As can be seen from the figure, AUC and AUPRC scores reached the tallest value when sampling ratio was among 0.9 to 1.1. When the sampling ratio surpasses 1.1, the educating set presents a reverse imbalance, that is, the original majority class becomes a minority class, and the original minority class becomes a majority class, so the AUC and AUPRC gradually decrease. On the ground of this, the sample scale is set to 1.0 in this chapter. The AUC values of LUA, CSG and SCI were lower than those of this method when the unbalance ratio of CiteSeer data set was 2 to 20, indicating that the proposed method was effective at different unbalance ratios.

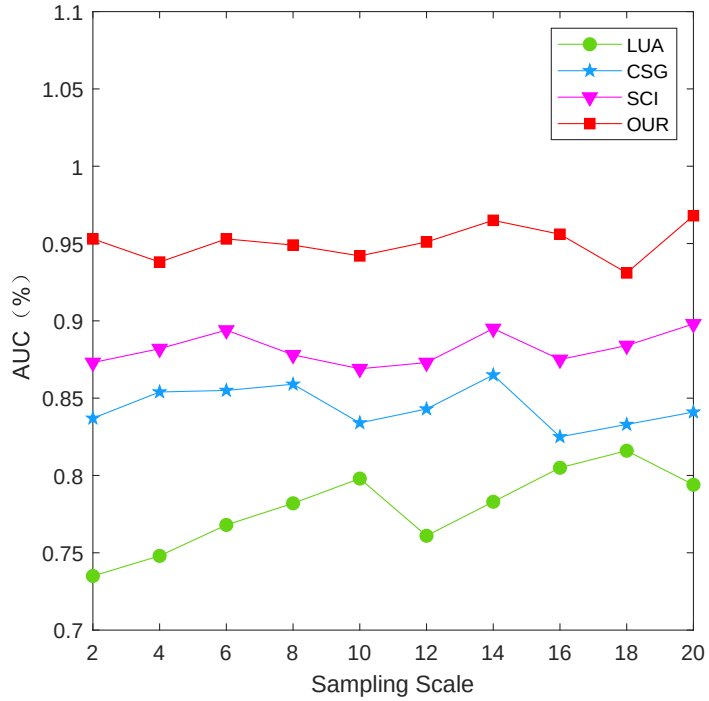

Figure 5. Comparison of AUC

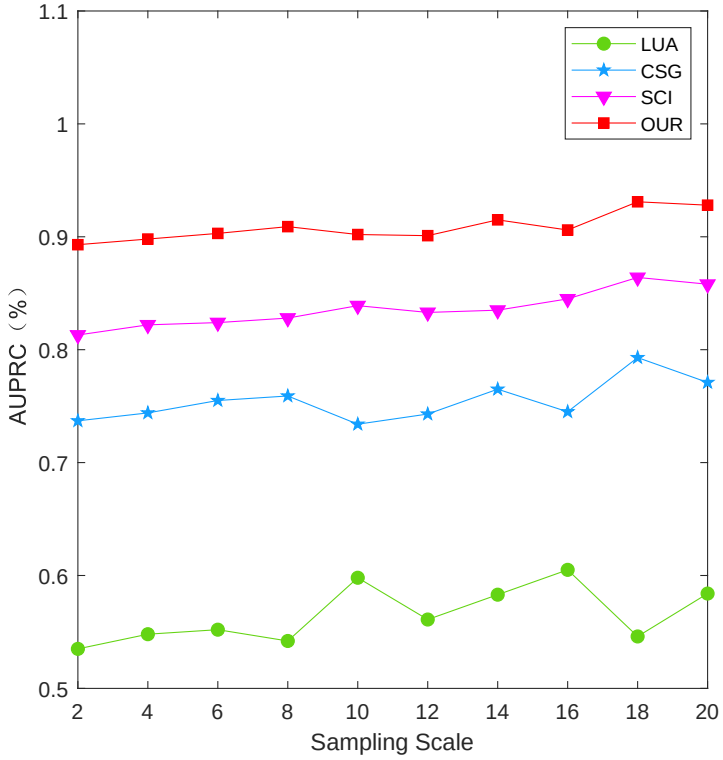

Figure 6. Comparison of AUPRC

6. **Conclusion.** In response to the problem of low accuracy in existing imbalanced data classification methods, this paper proposes a Bayesian optimized neural network model based imbalanced data classification method. This method first pre trains the model to train text data, and the trained text vector is represented as the feature vector of nodes

in the data graph to design a neural network model. Then, standardized methods are used to preprocess the data, and Bayesian optimization is used to update the posterior probability of the optimization function. Secondly, the model is trained on the dataset to obtain the weight of the entire network. Secondly, imbalanced data is mixed sampled, to reduce the number of negative samples. Finally, the experimental results indicate that the proposed method can effectively improve the accuracy, precision, and recall of the system.

Acknowledgment. The authors are grateful to the Management and Science University in Malaysia, the Guangzhou College of Commerce in China, and the University of Mosul in Iraq for their support. This work is funded by the Quality Engineering Project of Guangdong Provincial Department of Education (No.2022SJJXGG992), Quality Engineering Project of Guangzhou Business School (No. 2020XJJG01, 2021GZJG05), and Scientific Research Project of Guangzhou Business School (No. 2021XJYB07).

REFERENCES

[1] A. Izsto, G. Ru, M. Crescio, F. Ingravalle, C. Maurella, UBESP, D. Gregori, C. Lanera, D. Azzolina, and G. Lorenzoni, "Machine Learning Techniques applied in risk assessment related to food safety," *EFSA Supporting Publications*, vol. 14, no. 7, 1254E, 2017.

[2] J. Chen, H. Huang, S. Tian, and Y. Qu, "Feature selection for text classification with Naïve Bayes," *Expert Systems with Applications*, vol. 36, no. 3, pp. 5432-5435, 2009.

[3] J. Khan, J. S. Wei, M. Ringner, L. H. Saal, M. Ladanyi, F. Westermann, F. Berthold, M. Schwab, C. R. Antonescu, and C. Peterson, "Classification and diagnostic prediction of cancers using gene expression profiling and artificial neural networks," *Nature Medicine*, vol. 7, no. 6, pp. 673-679, 2001.

[4] T.-Y. Wu, A. Shao, and J.-S. Pan, "CTOA: Toward a Chaotic-Based Tumbleweed Optimization Algorithm," *Mathematics*, vol. 11, no. 10, 2339, 2023.

[5] T.-Y. Wu, H. Li, and S.-C. Chu, "CPPE: An Improved Phasmatodea Population Evolution Algorithm with Chaotic Maps," *Mathematics*, vol. 11, no. 9, 1977, 2023.

[6] Y. Ma, Y. Peng, and T.-Y. Wu, "Transfer learning model for false positive reduction in lymph node detection via sparse coding and deep learning," *Journal of Intelligent & Fuzzy Systems*, vol. 43, no. 2, pp. 2121-2133, 2022.

[7] S. P. Xu, K. Wang, M. R. Hassan, M. M. Hassan, and C.-M. Chen, "An Interpretive Perspective: Adversarial Trojaning Attack on Neural-Architecture-Search Enabled Edge AI Systems," *IEEE Transactions on Industrial Informatics*, vol. 19, no. 1, pp. 503-510, 2023.

[8] C.-M. Chen, Y. Gong, and J. M.-T. Wu, "Impact of Technical Indicators and Leading Indicators on Stock Trends on the Internet of Things," *Wireless Communications and Mobile Computing*, vol. 2022, pp. 1-15, 2022.

[9] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: a new over-sampling method in imbalanced data sets learning," *Lecture Notes in Computer Science*, vol. 3644, no. 5, pp. 878-887, 2005.

[10] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "DBSMOTE: density-based synthetic minority over-sampling technique," *Applied Intelligence*, vol. 36, pp. 664-684, 2012.

[11] B. Ozenne, F. Subtil, and D. Maucort-Boulch, "The precision–recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases," *Journal of Clinical Epidemiology*, vol. 68, no. 8, pp. 855-859, 2015.

[12] S.-H. Bae, and K.-J. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE Transactions on Medical Imaging*, vol. 34, no. 11, pp. 2379-2393, 2015.

[13] M. Alibeigi, S. Hashemi, and A. Hamzeh, "DBFS: An effective Density Based Feature Selection scheme for small sample size and high dimensional imbalanced data sets," *Data & Knowledge Engineering*, vol. 81, pp. 67-103, 2012.

[14] C. Beyan, and R. Fisher, "Classifying imbalanced data sets using similarity based hierarchical decomposition," *Pattern Recognition*, vol. 48, no. 5, pp. 1653-1672, 2015.

[15] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, "Finding the best classification threshold in imbalanced classification," *Big Data Research*, vol. 5, pp. 2-8, 2016.

[16] H. Yu, C. Sun, X. Yang, W. Yang, J. Shen, and Y. Qi, "ODOC-ELM: Optimal decision outputs compensation-based extreme learning machine for classifying imbalanced data," *Knowledge-Based Systems*, vol. 92, pp. 55-70, 2016.
[17] F. Cheng, J. Zhang, and C. Wen, "Cost-sensitive large margin distribution machine for classification of imbalanced data," *Pattern Recognition Letters*, vol. 80, pp. 107-112, 2016.
[18] B. Krawczyk, M. Woźniak, and G. Schaefer, "Cost-sensitive decision tree ensembles for effective imbalanced classification," *Applied Soft Computing*, vol. 14, pp. 554-562, 2014.
[19] M. Lopez-Martin, A. Sanchez-Esguevillas, J. I. Arribas, and B. Carro, "Supervised contrastive learning over prototype-label embeddings for network intrusion detection," *Information Fusion*, vol. 79, pp. 200-228, 2022.
[20] M. Zieba, J. M. Tomczak, M. Lubicz, and J. Swiatek, "Boosted SVM for extracting rules from imbalanced data in application to prediction of the post-operative life expectancy in the lung cancer patients," *Applied soft computing*, vol. 14, no. 1, pp. 99-108, 2014.
[21] W.-C. Lin, C.-F. Tsai, Y.-H. Hu, and J.-S. Jhang, "Clustering-based undersampling in class-imbalanced data," *Information Sciences*, vol. 409, pp. 17-26, 2017.
[22] Z. Zheng, X. Wu, and R. Srihari, "Feature selection for text categorization on imbalanced data," *ACM Sigkdd Explorations Newsletter*, vol. 6, no. 1, pp. 80-89, 2004.
[23] C.-W. Chen, Y.-H. Tsai, F.-R. Chang, and W.-C. Lin, "Ensemble feature selection in medical datasets: Combining filter, wrapper, and embedded feature selection results," *Expert Systems*, vol. 37, no. 5, e12553, 2020.
[24] R. Dubey, J. Zhou, Y. Wang, P. M. Thompson, J. Ye, and A. s. D. N. Initiative, "Analysis of sampling techniques for imbalanced data: An n= 648 ADNI study," *NeuroImage*, vol. 87, no. 3, pp. 220-241, 2014.
[25] H. Chen, T. Li, X. Fan, and C. Luo, "Feature selection for imbalanced data based on neighborhood rough sets," *Information sciences*, vol. 483, pp. 1-20, 2019.
[26] Z. Ghahramani, "Probabilistic machine learning and artificial intelligence," *Nature*, vol. 521, no. 7553, pp. 452-459, 2015.
[27] A. Qin, Z. Shang, J. Tian, Y. Wang, T. Zhang, and Y. Y. Tang, "Spectral–spatial graph convolutional networks for semisupervised hyperspectral image classification," *IEEE Geoscience and Remote Sensing Letters*, vol. 16, no. 2, pp. 241-245, 2018.
[28] P. Vuttipittayamongkol, and E. Elyan, "Neighbourhood-based undersampling approach for handling imbalanced and overlapped data," *Information Sciences*, vol. 509, pp. 47-70, 2020.
[29] L. Zhang, D.-Y. Lin, H. Wang, R. Car, and E. Weinan, "Active learning of uniformly accurate interatomic potentials for materials simulation," *Physical Review Materials*, vol. 3, no. 2, 023804, 2019.
[30] Z. Wu, S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip, "A comprehensive survey on graph neural networks," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 1, pp. 4-24, 2020.
[31] M. R. Islam, L. F. Abdulrazak, M. Nahiduzzaman, M. O. F. Goni, M. S. Anower, M. Ahsan, J. Haider, and M. Kowalski, "Applying supervised contrastive learning for the detection of diabetic retinopathy and its severity levels from fundus images," *Computers in Biology and Medicine*, vol. 146, 105602, 2022.