A Comprehensive Integration Method Based on Unbalanced Data Classification Problems

Yu Lian 1, *, Di Zhang1, Zhengkui Lin1, XueHao Sun1 and WenLi Yu1

1School of Computer Science and Technology, Dalian Maritime University

*525812069@qq.com

Abstract. This paper first introduces the characteristics of non-equilibrium data and the main problems brought by its classification. Based on this, it proposes a comprehensive integration method based on non-equilibrium data classification problems, and the method improvement background and strategy. The construction of base classifiers, and the selective integration methods of base classifiers are described in detail. Finally, the method proposed in this chapter is verified by experiments. The experimental results show that the proposed method is effective.

1. Introduction
In recent years, with the explosion of knowledge brought about by the development of informatization, the problem of classification of non-equilibrium data sets has attracted more and more attention from researchers. The so-called non-equilibrium data set refers to the fact that there are far more samples of certain classes in the data set than other classes[1], in which a class with many samples is called a majority class, and a class with a few samples is called a minority class. The classification of unbalanced data sets is very common in practical applications, such as fraudulent identification of fraudulent card transactions, oil well eruption of satellite image detection, failure prediction of telecommunication equipment, medical diagnosis[2], bankruptcy prediction, radar image surveillance, financial loan management, and fraud. Detection, customer loss prediction, malicious arrrears user identification, network intrusion detection[3], transaction fraud identification, biological data identification and text classification and so on. The problem of sending class has one thing in common, that is, the minority information is the focus of our attention. For example, in the monitoring of credit card transaction fraud, the number of fraud users is a minority and it is our target; in the diagnosis process of cancer patients, Cancer patients are the minority and are the objects of our concern. The problem of classification of non-equilibrium data sets exists widely in various fields. Therefore, the research on the classification of non-equilibrium data has a wide range of application prospects and practical significance.

2. related work and problems
At present, there are three classification strategies to study non-equilibrium data: First, from the perspective of the data level by changing the balance of the class distribution of the data; Second, from the perspective of improving the traditional algorithm; Third, through the data level and algorithm level Combine this with an improved perspective.

The most common method to change the balance of data distribution from the data level is re-sampling. The SMOTE (Synthetic Minority Over-sampling Technique) algorithm proposed by
Chawla et al.[4] determines the K-nearest neighbors for each minority class sample, and then randomly fetches the points on the connection line between the sample and its neighbors to generate a duplicate minority sample. The advanced under-sampling method mainly includes Kubat and Matwin [5] unilateral sampling method and Wilson’s Editing method.

The method of processing unbalanced data at the algorithm level mainly includes ensemble learning methods. Fan et al.[6] proposed that the AdaCost algorithm is one of the representative methods of cost-sensitive learning. Joshi et al.[7] showed that the AdaCost algorithm has high accuracy and recall rate for a few class samples. The most commonly used methods in ensemble learning are the Boosting algorithm and the Bagging algorithm. Considering that SMOTE algorithm is a mainstream sampling learning algorithm, and it is effective in the preprocessing of non-equilibrium data, some scholars have combined it with ensemble learning in the classification of non-equilibrium data, such as SMOTE. SMOTEBoost algorithm combined with Boosting. In 2002, Zhou et al[8]. demonstrated theoretically the validity of selective integration for the first time, and proposed a selective ensemble learning method based on real-valued coding genetic algorithm (GASEN, Genetic Algorithm based Select Selective Ensemble). GASEN method can achieve better classification accuracy with fewer neural networks.

3. Improvement of Unbalanced Data Classification Integration Method

3.1. The characteristics of unbalanced data sets and their main problems

Non-equilibrium data refers to the situation in which certain types of sample sizes are much larger than those of other types of data in some data sets. In practical problems, such classification problems are widespread. For example, in the analysis of customer churn, churn customers are often categorized as rare categories because there are usually only a few changes to service providers, and in the case of cancer diagnosis, samples of afflicted diseases also account for a very small percentage of total data.

The main characteristics of non-equilibrium data sets are that the number of samples in different categories differs greatly and sometimes they are heavily skewed; second, the distribution of different categories is non-equilibrium, resulting in the inability of the characteristics of rare samples to be detected. It is also the main cause of the low classification accuracy of a few samples. In addition, non-equilibrium data can easily cause other problems in the classification process, mainly including:

1. Data fragmentation problem
2. Summarizing deviations
3. Noise problem
4. Evaluation index problem

3.2 Improvement of Bagging Integrated Classification Algorithm Based on Unbalanced Data Sets

3.2.1. Algorithm Improvement Background and Strategy. Bagging algorithm is based on the simplest and most intuitive integrated approach to training set. Its own characteristics determine that it is suitable for designing multiple base classifiers in parallel. At the same time, multiple training sets with independent sample distributions are generated, and then learning is performed on different training sets to generate different base classifiers.

The traditional Bagging algorithm in some classifiers such as C4.5 decision tree, while classifying non-equilibrium data sets, can achieve high classification accuracy, but the classification accuracy of a few of them is not high, and In the classification of non-equilibrium data, people tend to pay more attention to the classification results of a few classes. Therefore, it is necessary to improve the traditional Bagging algorithm in order to improve its classification accuracy for a few classes of data on an unbalanced data set.

Combined with the implementation principle of the traditional Bagging algorithm, refer to the main strategies currently adopted to deal with the classification problem of unbalanced data sets. Improvement strategy:
In order to ensure that the distribution of the data samples in the training data set is consistent with the overall distribution, the classification accuracy of the base classifier for a few classes of data on an unbalanced data set is improved. According to the literature [1], firstly, the data set is changed based on the SMOTE algorithm idea and the undersampling method, and the unbalanced data is changed into relatively equalized data. Then the traditional Bagging algorithm is substituted into the base classifier to weaken the minority and majority classes.

In order to change the traditional Bagging algorithm using simple voting or averaging method to integrate the classification results of the base classifier, due to the failure to fully consider the problem of the classification accuracy caused by the contribution of the base classifier in the classifier integrated system, we proposed a A Classifier Selective Integration Strategy Based on Hybrid Genetic Algorithms.

3.2.2. Classifier Construction was Improved Based on Bagging Algorithm. Definition: 

\[ X = \{x_1, x_2, \ldots, x_n\} \]

is the set of total sample instances, 

\[ Y = \{y_1, y_2, \ldots, y_n\} \]

is the attribute collection, and Type is the sample category array. 

\[ X_{\min} \] is a collection of a few class samples in \( X \), 

\[ X_{\max} \] is a collection of most class samples in \( X \).

(1) the few classes was identified

Traversing \( X \), \( N \) records the number of samples included in each category, and stores the sample data contained in each category during the traversal; computes the mean value of \( N_{\text{avg}} = \text{AVG}(N) \), and the variance \( \text{square} = \text{Square}(N) \); For the second classification problem, concave for each category \( \text{type}[j] \) if the number of samples \( N_j < \text{avg} \) and \( |N_j - \text{avg}| > \text{square} \) then this class is a minority class, recorded as a positive class; otherwise it is a majority class, a negative class. If it is a multi-classification problem, the few classes that are most concerned are recorded as positive classes, and the rest are negative categories.

(2) Using a mixed sampling method to change the data set, changing the unbalanced data to relatively balanced data

\[ \text{oversample was fixed: for a few class samples } X_{\min}[i]; \text{ within each sample, } \]

\[ L = \left( \text{avg} - X_{\min} \right) / X_{\min} \] samples need to be reconstructed; set the parameter \( K \) value in the K-adjacent algorithm, that is, select the nearest neighbor, using the K nearest neighbor algorithm Find each sample \( X \), the nearest K of the same sample, \( j_1, j_2, \ldots, j_k \) calculate the attribute difference value \( \text{attr} \) of each \( X \) and the K samples, set \( U \) to be a random number between \( (0,1) \), and the new minority data was reconstructed \( X_{\text{new}} = X_i + u * \text{attr} \).

\[ \text{undersampling was Defined: Reduce } T = \left( X_{\max} - \text{avg} \right) / X_{\max} \] samples for each sample in most classes. In order to ensure that the data characteristics of the original majority samples can still be preserved after the undersampling, the method proposed in the reference literature sets \( \text{Avg}_{\max} \) as the mean value point of the majority of the samples; the distance between each sample of the majority class and \( \text{Avg}_{\max} \) is calculated, and The distance is arranged from the largest to the smallest; Set the counter \( \text{time} = 0 \), traverse the sorted data sample set, each time a data sample, Time plus 1, if and only if Time is an integer multiple of \( 1/T \), deleting this Class data.

(3) Substituting the new sample set into the traditional Bagging algorithm trains several base classifiers.

3.3. Base classifier selective integration based on hybrid algorithm

According to the literature[9], a hybrid algorithm based on genetic algorithm and particle swarm optimization for multi-subsidiary stratification is proposed. Combined with 1.2.2, the base classifier
selective integration algorithm based on hybrid genetic algorithm (Hybrid Genetic Algorithm Based) is proposed. Selective Ensemble (HGASEN) is described as follows:

1. The training set is generated through the bootstrap method and the training subset $D_i$ is obtained. After training with $D_i$, the base classifier $L_i$ is obtained. Using a hybrid genetic algorithm to select the base classifier $L$:

   1. First initialize $N$ subgroups randomly, $GA_i, i = 1,2,..., N$. Each subgroup independently runs its own genetic algorithm;
   2. For the best individuals in each population formed in the selection, the DPSO was used to evolve elite groups. When initializing the velocity of the elite group particle, the direction of the particle's velocity in each dimension is limited to be positive or negative at the same time, that is, the velocity interval is used at the same time, $[0, v_{j\text{max}}]$ or $[v_{j\text{max}}, 0]$.
   3. After a certain amount of algebra, it is judged whether the stop criterion is satisfied. The difference between this algorithm and other genetic and particle swarm algorithms is that the adoption of genetic algorithm as the main body ensures the superior global search ability of the algorithm; secondly, the use of multiple population structures can better maintain the diversity of the population and can complete the acceleration of evolution; and using dissipative particle swarm optimization algorithm to perform mutation and local search on elite individuals, the convergence speed is fast and the precision is high; finally, a hierarchical structure is used to separate the individuals responsible for global search at the bottom layer from the individuals responsible for local search at the top layer. Not only can speed up the search speed, but also can avoid premature convergence and fall into a local optimal solution.

4. Experiment and result analysis

4.1 Initial data set class balance characteristics

In this paper, we use the lower three groups of UCI imbalance datasets for comparison experiments. In the experiment, a few classes are defined as positive classes, and most classes are defined as negative classes.

The data set car_evaluation, the data has 6 features per sample, a total of 4 categories, the test selected 'Vgood' as a positive class, and the rest as a unified class; The data set balance-scale, the data has 4 characteristics per sample, a total of 3 categories, the test selected "B" as a positive class, the rest as a unified class; The data set breast_cancer, the data has 9 features per sample, and there are 2 categories. In the experiment, ‘recurrence-events’ is selected as the positive category, and the rest are unified as the negative category.

| data set        | Characteristic dimension | Positive sample number | Negative sample size | Positive and negative sample ratio |
|-----------------|--------------------------|------------------------|----------------------|-----------------------------------|
| Car_evaluation  | 6                        | 65                     | 1663                 | 0.05                              |
| Balance_scale   | 4                        | 49                     | 576                  | 0.17                              |
| Breast_cancer   | 9                        | 85                     | 201                  | 0.42                              |
Table 2. Data set Class Balance Characteristics After Processing

| data set        | Characteristic dimension | Positive sample number | Negative sample size | Positive negative sample ratio |
|-----------------|--------------------------|------------------------|----------------------|-----------------------------|
| Car_evaluation  | 6                        | 896                    | 831                  | 1.08                        |
| Balance_scale   | 4                        | 294                    | 288                  | 1.02                        |
| Breast_cancer   | 9                        | 170                    | 201                  | 0.85                        |

As shown in Table 1, Table 2, the data processed according to the method in 3.2.2 of this chapter has basically achieved class balance.

4.2 Classifier classification performance comparison

The experimental uses the C4.5 decision tree was used as a Bagging base classifier. The accuracy rate and recall rate were selected as evaluation indicators. The classification performances of the traditional Bagging classifier integration algorithm, GASEN algorithm and the HGASEN algorithm proposed in this chapter were experimentally compared.

In the experiment, the initial weights of each base classifier are the same. The hybrid particle swarm optimization algorithm uses real-number coding. The integrated AUC value is used as the fitness function value of the genetic algorithm and the particle swarm algorithm. The crossover probability pc is 0.6 and the mutation probability pm is 0.2. There are 30 subgroups and 10 individuals in each subgroup. The optimal particles in each subgroup constitute the upper elite subgroup with 30 particles. Iterated for 50 rounds of iterations. Each round of genetic algorithm and particle swarm algorithm runs for 20 generations. The number of individuals in each substitution substratum group is set to 2. The experimental results are shown in Table 3 and Table 4 below.

Table 3. Comparison of Classification Performance of Classifier

| Data set         | Classification algorithm | Precision rate | Recall rate |
|------------------|--------------------------|----------------|-------------|
| Balance_scale    | Bagging                  | 0.695          | 0.731       |
|                  | GASEN                    | 0.738          | 0.750       |
|                  | HGASEN                   | 0.751          | 0.762       |
| Car_evaluation   | Bagging                  | 0.933          | 0.931       |
|                  | GASEN                    | 0.942          | 0.946       |
|                  | HGASEN                   | 0.948          | 0.951       |
| Breast_cancer    | Bagging                  | 0.763          | 0.766       |
|                  | GASEN                    | 0.789          | 0.782       |
|                  | HGASEN                   | 0.795          | 0.789       |

Table 4. Comparison of Classification Performance of Minority Samples

| Data set         | Classification algorithm | Precision rate | Recall rate |
|------------------|--------------------------|----------------|-------------|
| Balance_scale    | Bagging                  | 0.200          | 0.020       |
|                  | GASEN                    | 0.400          | 0.167       |
|                  | HGASEN                   | 0.475          | 0.364       |
| Car_evaluation   | Bagging                  | 0.860          | 0.880       |
|                  | GASEN                    | 0.878          | 0.889       |
|                  | HGASEN                   | 0.879          | 0.901       |
| Breast_cancer    | Bagging                  | 0.750          | 0.318       |
|                  | GASEN                    | 0.783          | 0.354       |
|                  | HGASEN                   | 0.779          | 0.458       |
4.3 The result analysis
As can be seen from Table 3 and Table 4, the improved algorithm proposed in this chapter is superior to the GASEN and Bagging classifiers in classifier recall and precision, and it is lower than the minority class recall rate. The GASEN and Bagging classifier accuracy rate on the breast_cancer data set is lower than the GASEN difference. This chapter focuses on the classification accuracy of a few classes. The recall rate is the focus of attention, and the improved algorithm has better classification effect in a few classes.

5. Conclusions
The improved algorithm presented in this paper is superior to the GASEN and Bagging classifiers in the classifier recall and precision, and the GASEN and Bagging classifiers breast_cancer data set is better than GASEN in the minority class recall. The difference is that this chapter focuses on the classification accuracy of a few categories. The recall rate is the focus of attention, so the improved algorithm in this paper has a better classification effect in a few categories.

6. Summary and Prospect
This paper proposes that the integration algorithm for non-equilibrium data sets is based on the hybrid genetic algorithm, although to a certain extent by raising the difference between the classifiers participating in the integration, the final classification effect is raised. However, it still has some limitations. Further studies are mainly considered from the following three perspectives:

(1) Base classifier selection.
(2) Choose a better base classifier fusion method.
(3) Integrate based on heterogeneous algorithms.

References
[1] Haque M N,Noman N,Berretta R,et al. Heterogeneous Ensemble Combination Search Using Genetic Algorithm for Class Imbalanced Data Classification.[J].Plos One. 2016, 11(1):e0146116
[2] Zung M, Zou B, Wei F, et al. Effective prediction of three common diseases by combining SMOTE with Tomek links technique for imbalanced medical data[C]/Online Prediction and Computing Science (ICOACS) IEEE International Conference of. 2016:225-228
[3] Rodda S,Erothi U S R.Class imbalance problem in the Network Intrusion Detection Systems[C]/Electrical Electronics and Optimization Techniques (ICEEOT) International Conference on.IEEE, 2016:2685-2688.
[4] Chawla NV,Lazarevic A,Hall LO.SMOTEBest:improving prediction of the minority class in boosting[C]//The 7th European Conf on Principles and Practice of Knowledge Discovery in Databases. 2003:107-119
[5] Kubat M,Matwin S. Addressing the Curse of Imbalanced Data sets:One-Sided Sampling.[C]//Proc of the 14th International Conference on Machine Learning. 997:179186
[6] Fan W, Stolfo S J, Zhang J, et al. Adacost: Misclassification Cost-Sensitive Boosting[C]//Proceedings of the Sixteenth International Conference on Machine Learning. 1999:97105
[7] Joshi M,Kumar V,Agarwal R.Evaluating Boosting Algorithms to Classify Rare Classes: Comparison and Improvements[A]. First IEEE International Conference on Data Mining[C]. San Jose CA, 2001
[8] Zhou Z H, Wu J X, Tang W. Ensembling several networks: Many could be better than all. Artificial Intelligence. 2002, 137(1-2):239-263
[9] Jin M, Lu H. A multi-subgroup hierarchical hybrid algorithm based on genetic algorithm and particle swarm optimization [J]. Control Theory and Application.2013,30(10):1231-1238