An Ensemble Learning Imbalanced Data Classification Method Based on Sample Combination Optimization

Yuxin Wang
School of Software, Xi'an Jiaotong University, China

E-mail: 1010120806@qq.com

Abstract. Imbalanced data classification is one of the hot topics in data mining and machine learning in recent years. In practice, imbalanced data classification is very common, such as cancer detection, spam discrimination, credit card fraud detection, etc. Because of the large difference in the number of categories and imbalanced distribution, traditional classification algorithms have poor classification effect on minority classes, and correct identification of minority classes often brings greater value. Therefore, how to effectively identify minority classes in imbalanced data is of great importance. Practical significance. Aiming at the problems that the Bagging-based imbalanced data classification method cannot guarantee the validity and existence of classification boundaries by adding redundant noise information and sampling, an ensemble learning GABagging method based on sample combination optimization is proposed. Firstly, the sample combination optimization algorithm uses genetic algorithm to select a subset from most classes and construct a new data set with a few classes. Subsequently, several sample combinatorial optimization algorithms are used to train and integrate several classifiers. The experimental results show that GABagging can improve the correct recognition ability of minority classes on 19 imbalanced datasets compared with other similar methods such as TPR and AUC, without excessive loss of recognition ability of majority classes. It is proved that GABagging can compensate for the shortcomings of related Bagging-based methods such as easy loss, increasing samples and not guaranteeing the validity and existence of classification boundaries after sampling.

1. Introduction
Imbalanced data appears in every field of real life, which is very common. However, traditional classification algorithms are based on the assumption that data is balanced. When encountering imbalanced data, they have poor ability to recognize a few classes correctly. The most intuitive performance is that the predicted results are all classes of most samples [1]. Such results are obviously meaningless, and at the same time, compared with most sample classes, a few sample classes carry more important information [2]. The value of identifying a correct minority sample is much higher than that of most sample classes. When facing the problem of imbalanced data classification, ensemble learning often combines with data-level methods to construct multiple data sets through multiple sampling processes, which can alleviate the adverse effects of data imbalance on classification [3]. Common ensemble learning methods include Bagging and Boosting. In the related Bagging method, Barandela et al. proposed the UnderBagging method [4], which combines the undersampling technique with the Bagging method, retains all the minority samples at a time, and takes the playback and non-playback sampling from the majority samples, extracts the majority samples equal to the minority samples, and forms a data subset, the number of basic learners, that is, the number of data subsets. It is determined by...
the ratio of the number of samples in most classes to that in a few classes. Molinara et al. proposed the EBBag method [5]. EBBag combines a small number of classes with a random subset of most classes, so that there is little difference in the number of classes in the training data set for each model in the integration. Wang et al. proposed Over Bagging and SOMTE Bagging methods [6]. Both of them combine the over-sampling method with Bagging method. Over Bagging is different from the previous Bagging methods in retaining all the minority samples, but adopts the same sampling strategy for most and minority categories. Random over-sampling method is adopted to extract a certain proportion of samples from various categories, and its proportion is set to majority samples. The ratio of the number of samples to the number of the current category is the product of an artificial proportional parameter $a$; SOMTEBagging retains all the majority of samples and generates a small number of samples by random oversampling and SOMTE method, with the same sampling proportion as the former. Zhou et al. proposed the Easy Ensemble method [7], which combines random undersampling, Adaboost and Bagging. The random undersampling method is used to generate a number of data subsets. It retains all minority samples and generates a majority of samples with the same number of minority samples. Adaboost is used as the base learner. Yang et al. proposed HSBagging method [8]. Firstly, bagging method was used to generate in-pocket data and out-of-pocket number. Then, under-sampling and SOMT method were used to sample in-pocket data at different sampling ratios for several times to obtain several classifiers. Then, the out-of-pocket data was used to evaluate the classifier and select the best classifier until several base classifiers were generated. Rong et al. proposed LBUS method [9], which uses ITQ method [10] to generate several subsets of data and train classifiers on this basis.

The above method improves Bagging's random sampling process from the data level, which basically uses random undersampling or oversampling to generate data subset. However, the shortcomings of undersampling and oversampling bring great problems to these Bagging Imbalanced data classification methods based on sampling. Because of the need to delete part of the majority class samples, undersampling directly results in the loss of part of the majority class sample information, which contains important information that plays a decisive role in the classification boundary division, thus reducing the classification boundary pair [11]. Oversampling will inevitably increase redundancy and noise samples due to the addition of a few classes of samples, resulting in over-fitting phenomenon [12], and the classification boundary will be over-matched to the current number. At the same time, undersampling and over-sampling methods only make the number of samples of different categories no longer vary greatly, the validity and existence of classification boundaries of data sets cannot be guaranteed.

2. Sample combination optimization algorithm
Following are the specific steps of the sample combination optimization algorithm.

Algorithm: Sample combination optimization algorithm

**Input:** Data $S = \{(x_1,y_1),(x_2,y_2),(x_3,y_3), \ldots, (x_n,y_n)\}$

**Output:** Data $S^*$

1. The maj of most class samples and min of a few class samples in $S$ were extracted, and the number of samples was $N$ and $M$, respectively.
2. Initialize $T$ to the current algebra,
   Initialize $n$ as the number of individuals per generation.
   Initialize $\text{max\_gen}$ as the largest evolutionary algebra.
   Initialize last_population to record the population of the previous generation of the current generation.
   Initialize last_average = 0 to record the average fitness of the previous generation.
   Initialization stop_time represents the number of stops.
   Initialization stop_time represents the number of stops.
   Initialize final_population as the final output group
3. Initialization divides maj into $M$ parts and combines with min to form the first generation population.
4. While stop_time != 10 or $T$ != $\text{max\_gen}$
4.1: Calculate the fitness of each individual in the T generation and descend to get queue A
4.2: Calculate the difference between each individual in queue A and last average, and exclude the negative individual.
4.3: 30% of the individuals excluded from queue A
4.4: if len(A)≥2, then jump to Step 4.5
   else jump to Step 4.6
4.5: Calculate the average fitness s of generation T
   if s == last average, then stop_time←stop_time+1
   else stop_time←0
   last populaition←A
   last average←s
   T←T+1
4.6: Roulette betting chooses two individual points from last population to intersect and produces m/3 individuals
4.7: Roulette betting chooses two bodies from last population to intersect at two points and produces m/3 individuals.
4.8: Roulette betting varies from last population to m/3 individuals.
To Step 4
5: Output the data set of the best individual S* in last population for the final sampling

When the sample combination optimization algorithm performs cross-over operation, it randomly selects the same location point from the majority of the sample data of two individuals, divides the majority of the sample data into two parts, and completes single-point cross-over by exchanging data sets of different parts of two individuals. When crossing two points, similar to single point crossing, we select two points at the same position for minority and majority samples of two individuals to complete the exchange. The mutation is aimed at an individual in a population. 10% of the samples are randomly removed from most and a few samples, and the remaining samples are retained to form a new individual. To evaluate an individual's fitness, a logistic regression model is trained by using the current individual, i.e. the current data set, to classify and predict the validation set, and to obtain the AUC value. The AUC value is used as the current individual's fitness. The higher the fitness value is, the higher the AUC value is, the clearer and more effective the classification boundary of the current individual is. The convergence conditions of the whole algorithm are two. First, if the evolution times of the population reach the maximum iteration threshold, the algorithm terminates. Second, when the average fitness of the continuous N-generation population remains unchanged, the algorithm is considered to be convergent. When either of the two convergence conditions is satisfied, the algorithm terminates and the individuals with the highest fitness value in the current population are output.

To overcome the shortcomings of sampling technology that cannot effectively remove overlapping areas and guarantee the existence and validity of classification boundaries, sample combination optimization algorithm is used to remedy the shortcomings. Random under-sampling and over-sampling as well as SOMTE algorithm are used in Bagging-based Imbalanced data classification method. For random under-sampling, the number of samples to be cut is selected according to random probability, with one. For random oversampling, only a few samples have been copied, which is no different from increasing the overlapping region. At the same time, SOMTE algorithm has a certain randomness in selecting synthetic samples, and the overlapping region still exists. Sample combination optimization algorithm regards the problem as a sample combination problem. An optimal sample combination has clear and effective classification boundary and the overlapping area is eliminated. The genetic algorithm is used to find the optimal combination. The current AUC is used as fitness evaluation to guide the genetic algorithm to find the sample combination with the highest AUC value, that is, to find an effective and clear data set with classification boundary. Compared with the sampling technology, the data set obtained by this method has removed the overlapping area, which ensures the existence and validity of the classification boundary.
3. Ensemble learning classification method based on sample combination optimization

In Figure 1, we introduce how GABagging overcomes the shortcomings of Bagging Imbalanced data classification methods in combination with Figure 1. In Figure 1, in the last Imbalanced data DN, the minority classes correspond to most classes of D1-valued DN-1 data set, and most of them correspond to a few classes of the previous data set. This is because the sample combination optimization algorithm is essentially under-sampled, and it will reduce part of the majority of class samples. As the number of iterations increases, the number of most class samples in the data set decreases, while a few. The number of classes remains unchanged, so there will always be a time when the number of most class samples in the data set will be less than that of a few classes. At this time, the last imbalanced data set DN to be processed will be the last one. Taking the classes with fewer samples as a few classes and the classes with more samples as the majority classes, the sample combination optimization algorithm will be carried out to obtain the last balanced data set DN and join the set S, GA. The Bagging sampling process is over. In the data set S obtained by GABagging, each sample corresponds to a sample in the original Imbalanced data set without adding new samples, and the set S includes all samples in the original Imbalanced data set without reducing the original samples. At the same time, the classification boundary of each data set in the set S is clear and effective, and the number of samples in the set S is the same as that in the original Imbalanced data set. The specific samples are the same, and the information content is identical with that of the original Imbalanced data. It does not increase or decrease any sample information, which makes up for the shortcomings of the above-mentioned Bagging Imbalanced data classification method that is liable to lose important sample information and increase redundant noise sample information.

![Figure 1. Sampling process of GABagging.](image)

4. Experimental results and analysis

The experimental data of GABagging simulation are derived from KEEL Imbalanced data set in Table 1. The training set, validation set and test set of each experiment were obtained from source data D, which accounted for 50%, 30% and 20% of source data D respectively. Then the classification model was trained with these data and the classification effect was tested. In order to eliminate the influence of random factors and prove the effectiveness and promotion of GABagging method, 500 experiments were carried out, and the mean value was taken as the final experimental result. Six methods are selected to compare OverBagging, UnderBagging, ABBag, HSBagging and RBSBagging. Similarly, these five
methods are experimented with the same methods mentioned above. The experimental indexes were TNR and AUC, and the final results were the average of 500 experiments for each index.

Table 1. Imbalanced Data.

| Data    | Size  | Min/Maj | Min/Size |
|---------|-------|---------|----------|
| b1      | 683   | 239/444 | 0.35     |
| breast  | 789   | 220/569 | 0.28     |
| yeast6  | 1485  | 35/1450 | 0.02     |
| yeast4  | 1485  | 51/1434 | 0.03     |
| yeast3  | 1484  | 163/1321| 0.11     |
| yeast1  | 1485  | 429/1054| 0.29     |
| wisconsin| 684   | 239/445 | 0.35     |
| vehicle3| 847   | 212/635 | 0.25     |
| vehicle2| 846   | 218/638 | 0.26     |
| vehicle1| 847   | 217/630 | 0.26     |
| vehicle0| 846   | 199/647 | 0.24     |
| segment0| 2309  | 329/1980| 0.14     |
| pima    | 769   | 268/501 | 0.35     |
| haberman| 306   | 81/225  | 0.26     |
| glass6  | 214   | 29/185  | 0.14     |
| glass4  | 215   | 13/202  | 0.06     |
| glass2  | 214   | 17/197  | 0.08     |
| glass0  | 214   | 70/144  | 0.33     |
| spect   | 80    | 26/54   | 0.33     |

Table 2. Comparison of TPR values of different methods.

| Data    | OverBagging | UnderBagging | ABBag | HSBagging | RBSBagging | GABagging |
|---------|--------------|--------------|-------|-----------|------------|-----------|
| b1      | 0.55         | 0.76         | 0.692 | 0.6911    | 0.45       | 0.8       |
| breast  | 0.75         | 0.7          | 0.755 | 0.5424    | 0.6523     | 0.95      |
| yeast6  | 0.59         | 0.83         | 0.59  | 0.7986    | 0.59       | 0.88      |
| yeast4  | 0.7          | 0.85         | 0.7011| 0.7585    | 0.76       | 0.9       |
| yeast3  | 0.59         | 0.79         | 0.743 | 0.6993    | 0.81       | 0.91      |
| yeast1  | 1.0          | 0.9          | 0.9915| 0.8991    | 0.8        | 1.0       |
| wisconsin| 0.5667     | **0.9333**   | 0.8   | 0.84      | 0.9        | 0.9333    |
| vehicle3| 0.8267       | 0.9          | 0.8633| 0.7887    | 0.89       | 0.9293    |
| vehicle2| 0.3308       | 0.585        | 0.5551| 0.5686    | 0.5212     | 0.6842    |
| vehicle1| 0.8571       | 0.9714       | 0.8771| 0.9818    | 0.8        | 1.0       |
| vehicle0| 0.8694       | 0.9196       | 0.8259| 0.8979    | 0.9        | 0.995     |
| segment0| 0.5391       | 0.7792       | 0.4682| 0.7535    | 0.7952     | 0.8526    |
| pima    | 0.9131       | 0.9586       | 0.9487| 0.9662    | 0.9127     | 0.9909    |
| haberman| 0.2666       | 0.7          | 0.3788| 0.7897    | 0.45       | 0.9222    |
| glass6  | 0.6855       | 0.96         | 0.7988| 0.8647    | 0.81       | 0.98      |
| glass4  | 0.4612       | 0.7012       | 0.4883| 0.7977    | 0.6537     | 0.8482    |
| glass2  | 0.7339       | 0.9131       | 0.7555| 0.8953    | 0.6874     | 0.9443    |
| glass0  | 0.1945       | 0.8236       | 0.4112| 0.8786    | 0.79       | 0.9       |
| spect   | 0.4          | 0.8          | 0.8275| **0.87**  | 0.7539     | 0.8571    |
In Table 2, the bold font of each line is the maximum value of the line, and the accuracy of TPR value for minority categories represents the correct recognition ability of the model for minority categories. The larger the TPR value, the stronger the correct recognition ability of the model for minority categories. It can be seen from the table that the TPR value of GABagging method is the highest on other data sets except SPECT data set, which is slightly lower than the maximum value of 0.0129. Through the above data, it can be proved that GABagging can be effectively and correctly recognized. Compared with other Bagging-based Imbalanced data classification methods, a few classes have improved significantly. Next, in order to prove that GABagging improves the correct recognition ability of a few classes without excessive loss of the overall classification ability, relevant experiments are carried out.

Table 3. Comparison of AUC values of different methods.

| Data     | OverBagging | UnderBagging | ABBag | HSBagging | RBSBagging | GABagging |
|----------|-------------|--------------|-------|-----------|------------|-----------|
| b1       | 0.1696      | 0.1566       | 0.3385| 0.4578    | 0.9165     | 0.9041    |
| breast   | 0.1138      | 0.1285       | 0.3543| 0.4977    | 0.9592     | 0.9834    |
| yeast6   | 0.1826      | 0.0928       | 0.2425| 0.4224    | 0.9158     | 0.9286    |
| yeast4   | 0.1233      | 0.1207       | 0.3587| 0.2958    | 0.9666     | 0.9793    |
| yeast3   | 0.2066      | 0.1696       | 0.4225| 0.2156    | 0.9126     | 0.9085    |
| yeast1   | 0.0062      | 0.0206       | 0.3854| 0.2664    | 0.9802     | 0.9725    |
| wisconsin| 0.1812      | 0.1329       | 0.2454| 0.4345    | 0.9275     | 0.8458    |
| vehicle3 | 0.0615      | 0.0635       | 0.3578| 0.4556    | 0.981      | 0.9623    |
| vehicle2 | 0.3688      | 0.4107       | 0.2657| 0.3587    | 0.564      | 0.6853    |
| vehicle1 | 0.0488      | 0.0107       | 0.2479| 0.3556    | 0.9687     | 0.9984    |
| vehicle0 | 0.0256      | 0.0276       | 0.3457| 0.5355    | 0.9855     | 0.9931    |
| segment0 | 0.1934      | 0.1926       | 0.2147| 0.3335    | 0.8142     | 0.8803    |
| pima     | 0.0103      | 0.0168       | 0.2523| 0.4125    | 0.9972     | 0.9978    |
| haberman | 0.3142      | 0.2271       | 0.3541| 0.2578    | 0.7568     | 0.779     |
| glass6   | 0.0806      | 0.0416       | 0.2147| 0.3688    | 0.9723     | 0.9842    |
| glass4   | 0.2902      | 0.2716       | 0.4525| 0.2456    | 0.7059     | 0.7836    |
| glass2   | 0.081       | 0.0568       | 0.2146| 0.3357    | 0.9491     | 0.975     |
| glass0   | 0.3178      | 0.1253       | 0.2567| 0.3558    | 0.8848     | 0.9348    |
| spect    | 0.2119      | 0.1331       | 0.2227| 0.1488    | 0.8897     | 0.931     |

In Table 3, the bold font of each line is the maximum value of the line. AUC represents the overall classification ability of the model, including the correct recognition ability of a few and most categories. The larger the AUC value, the stronger the overall classification ability of the model. It can be seen from the table that except for b1, yeast1, Wisconsin and vehicle3 data sets, the AUC value of GABagging is the highest in other data sets, and the TPR value of GABagging is also the highest in the corresponding data of Joint Table 2, which indicates that GABagging has a high recognition rate of a few classes, and the overall classification ability of GABagging is higher than other methods.

5. Conclusion
Aiming at the problem that the Bagging-based Imbalanced data classification method cannot guarantee the validity and existence of classification boundary by adding redundant noise information and sampling, an integrated learning Imbalanced data classification method GABagging based on sample combination optimization is proposed. In GABagging, sample combination optimization algorithm is used to sample source data, which can remove overlapping regions of different categories of data subsets after sampling and ensure an effective and clear classification boundary. Secondly, by using the sample combination optimization algorithm many times and eliminating most of the samples that have been sampled before each use, until all samples are sampled, to preserve all data sample information. However, GABagging method has good prediction effect, but its time complexity is too high and
running time is too long. Considering introducing the idea of parallelization, the parallelization of genetic algorithm is studied to further improve the performance of GABagging method and shorten the running time.

References
[1] S. Barua, M. M. Islam, X. Yao and K. Murase, "MWMOTE--Majority Weighted Minority Oversampling Technique for Imbalanced Data Set Learning," in IEEE Transactions on Knowledge and Data Engineering, vol. 26, no. 2, pp. 405-425, Feb. 2014.
[2] B. Wang and J. Pineau, "Online Bagging and Boosting for Imbalanced Data Streams," in IEEE Transactions on Knowledge and Data Engineering, vol. 28, no. 12, pp. 3353-3366, 1 Dec. 2016.
[3] M. R. Hasan, H. Gholamhosseini and N. I. Sarkar, "A new ensemble classifier for multivariate medical data," 2017 27th International Telecommunication Networks and Applications Conference (ITNAC), Melbourne, VIC, 2017, pp. 1-6.
[4] H. R. Sanabila and W. Jatmiko, "Ensemble Learning on Large Scale Financial Imbalanced Data," 2018 International Workshop on Big Data and Information Security (IWBIS), Jakarta, 2018, pp. 93-98.
[5] M. Molinara, M. T. Ricamato and F. Tortorella, "Facing Imbalanced Classes through Aggregation of Classifiers," 14th International Conference on Image Analysis and Processing (ICIAP 2007), Modena, 2007, pp. 43-48.
[6] R. Nicole, “Title of paper with only first word capitalized,” J. Name Stand. Abbrev., in press.
[7] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," 2009 IEEE Symposium on Computational Intelligence and Data Mining, Nashville, TN, 2009, pp. 324-331.
[8] X. Liu, J. Wu and Z. Zhou, "Exploratory Undersampling for Class-Imbalance Learning," in IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics), vol. 39, no. 2, pp. 539-550, April 2009.
[9] M. H. Popel, K. M. Hasib, S. Ahsan Habib and F. Muhammad Shah, "A Hybrid Under-Sampling Method (HUSBoost) to Classify Imbalanced Data," 2018 21st International Conference of Computer and Information Technology (ICCIT), Dhaka, Bangladesh, 2018, pp. 1-7.
[10] T. Rong, X. Tian and W. W. Y. Ng, "Location bagging-based undersampling for imbalanced classification problems," 2016 International Conference on Wavelet Analysis and Pattern Recognition (ICWAPR), Jeju, 2016, pp. 72-77.
[11] Qianqian Huang and Xiaolong Zhang, "An improved ensemble learning method with SMOTE for protein interaction hot spots prediction," 2016 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), Shenzhen, 2016, pp. 1584-1589.
[12] H. R. Sanabila and W. Jatmiko, "Ensemble Learning on Large Scale Financial Imbalanced Data," 2018 International Workshop on Big Data and Information Security (IWBIS), Jakarta, 2018, pp. 93-98.