Journal of Medical Systems (2019) 43: 39
https://doi.org/10.1007/s10916-018-1154-8

SYSTEMS-LEVEL QUALITY IMPROVEMENT

Distribution-Sensitive Unbalanced Data Oversampling Method for Medical Diagnosis

Weihong Han [1,2] · Zizhong Huang [3] · Shudong Li [1] · Yan Jia [3]

Received: 26 October 2018 / Accepted: 25 December 2018 / Published online: 10 January 2019

Abstract
Aiming at the problem of low accuracy of classification learning algorithm caused by serious imbalance of sample set in medical diagnostic application, this paper proposes a distribution-sensitive oversampling algorithm for imbalanced data. The algorithm accurately divides the minority samples into noise samples, unstable samples, boundary samples and stable samples according to the location of the minority samples. Different samples are processed differently to select the most suitable sample for the synthesis of new samples. In the case of sample synthesis, a distribution-sensitive sample synthesis method is adopted. Different sample synthesis methods are selected according to their different distance from the surrounding minority samples, so as to ensure that the newly synthesized samples have the same characteristics with the original minority samples. The real medical diagnostic data test shows that this algorithm improves the accuracy rate of classification learning algorithm compared with the existing sampling algorithms, especially for the accuracy rate and recall rate of minority classes.

Keywords Medical diagnosis · Imbalanced data · Data resampling · Oversampling · Undersampling · Classification learning

Introduction

With the advent of era of big data and the gradual opening and sharing of medical and health big data, the analysis and application of medical big data has become a new research hotspot and an important means to realize the wise medical treatment. The use of medical and health big data makes personal medical and health data become an important strategic resource, bringing the development and reform of the medical and health service industry, including personalized medical services, precision medical treatment, personalized medicine, and so on.

Medical diagnosis is one of the most important applications of health and medical big data. However, the imbalance of medical and health big data seriously affects the accuracy of medical diagnosis classification learning algorithm. Imbalanced data means that at least one category in a dataset contains less samples than other categories. Data imbalance exists widely in the real world, especially in medical big data. Imbalanced big data is due to the large difference in the number of different categories of data sample, for example, in the actual medical big data set, the data imbalance ratio often exceeds 1000:1. The traditional machine learning method assumes that the sample distribution of the dataset is basically uniform. Therefore, in the design of the algorithm, the overall accuracy rate is taken as the optimization objective, and there is no distinction between different categories of samples. Imbalanced data makes it difficult for traditional learning algorithms to achieve good results, so it is necessary to preprocess the data before learning with training sets, that is the resampling of imbalanced big data.

In the classification learning of imbalanced big data in medical diagnosis, unlike the general imbalanced big data classification learning algorithm, the recall rate of the minority classes in medical diagnosis is very high, which can properly sacrifice the overall algorithm and the classification accuracy of majority classes. In addition, there are often very few

This article is part of the Topical Collection on *Systems-Level Quality Improvement*

✉ Weihong Han
hanweihong@gzhu.edu.cn

[1] Institute of Advanced Technology in Cyberspace, Guangzhou University, Guangzhou 510006, Guangdong, China

[2] Institute of Electronic and Information Engineering of UESTC in Guangdong, Guangzhou, Guangdong, China

[3] School of Computer of National University of Defense Technology, Changsha 410073, Hunan, China

minority samples in medical diagnostic dataset, and the quality requirements of the newly generated minority samples is very high when over-sampling the minority classes. For example, in tumor diagnosis, newly generated tumor samples should conform to the overall characteristics of this category. Therefore, it is very important to study the over-sampling algorithm of imbalanced big data for medical diagnosis.

In this paper, an oversampling method for distribution-sensitive imbalanced big data is proposed in order to meet the classification learning needs of imbalanced big data in medical diagnosis, including the distribution-sensitive minority sample selection algorithm and the minority sample synthetic algorithm of weight adaptive adjustment, which improves the quality of newly generated minority samples. The real medical data set proves that the algorithm effectively improves the accuracy of medical diagnosis classification learning algorithm.

Related research

In the field of imbalanced data learning, Sun, Y., Wong, A.K. et al. [1] have given a systematic overview of imbalanced data problems and corresponding solutions. In 2009, Garc'ıa [2] gave an undersampling method for imbalanced large data classification. L'opez et al. [3] provided insights into the nature of the imbalanced data learning problem and analyzed the trend of related technologies. Wang and Yao [4] focused on the multi classification problem of imbalanced data, and gave the solution. Zhang et al. [5] further studied the multi-classification of imbalanced data, focusing on the performance of multi-classification algorithm. Krawczyk et al. [6] discussed the challenges faced by imbalanced data learning and the future development trend and research direction.

Oversampling methods for minority classes are an important research direction in imbalance data classification learning. The earliest oversampling method is random oversampling, which randomly selects minority samples for copying, thereby increasing the proportion of the minority classes. However, random selection of samples for copying may result in copying a large number of noise samples, which are not important samples. Therefore, various heuristic oversampling methods have emerged. Heuristic oversampling is also a copy of minority samples, and does not create new examples by itself. The difference is that which samples are copied is selective, not random. The method of copying minority samples is simple to implement, but it is easy to cause over-fitting of the classifier due to a large number of repeated samples.

The new sample synthesis method is not simply to copy minority samples, but to generate new minority samples by using the existing minority samples, which can effectively avoid the over-fitting problem caused by too few samples in the classification process. The method of synthesizing new samples has been extensively studied. The SMOTE [7] (Synthetic Minority Over-sampling Technique) method proposed by Chawla et al. is used to calculate the similarity of the samples in the feature space and to select and synthesize new minority samples based on the similarity. Bunkhumpornpat et al. [8] improved the SMOTE method by introducing the concept of security level, which assigned a security level to each new minority sample before it was synthesized to ensure that all the newly synthesized minority samples were in the security zone. Han et al. [9] proposed an over-sampling algorithm Borderline-SMOTE based on classification boundary. The algorithm considers that the samples of classification boundary have the greatest influence on the classification effect, so only the minority classes near the boundary are over-sampled. Bunkhumpornpat et al. put forward a density based SMOTE sampling method DBSMOTE [10, 11]. First, the DBSCAN method is used to scan the data set and form a clustering. Then, for each minority class sample, a new minority sample is synthesized on the shortest path between each minority class cluster center and it. Finally, there are more new synthetic samples near the minority clustering centers. KE Bennin et al. have recently proposed a method for the synthesis of minority samples based on the theory of chromosomal inheritance, which interprets two different samples as parents, inheriting different characteristics from each parent and contributing to the diversity of data distribution [12].

At present, the research on over-sampling methods for minority samples in imbalanced data classification is aimed at general data sets, and the quality of minority samples is not as high as that of medical diagnostic data [13–15]. The effect of existing methods applied to medical diagnostic data sets needs to be improved.

Distribution-sensitive unbalanced data oversampling method

In this paper, a distribution-sensitive oversampling algorithm for imbalanced data is proposed. Firstly, the user is allowed to input the number of minority samples needed, and the number of newly generated samples is determined according to the user's needs. Secondly, the distribution of minority samples is analyzed, and the samples are selected to synthesize new minority samples according to the distribution of minority samples. Finally, in the process of synthesizing new samples, different new sample generation methods are adopted according to the characteristics of different minority samples. The algorithm ensures that the new generated minority samples retain all the features of the original sample as far as possible. Using the above techniques, the minority sample sets that meet user's needs are finally generated, which can effectively improve the classification accuracy of imbalanced big data.

Definition and symbolization

Training set: $D = \{(x_1, y_1)\cdots(x_n, y_n)\}$ includes n samples. Each sample has d attributes, $x_i \in \mathbb{R}^d$, $y_i \in L = \{1, -1\}$ is its corresponding category label.

Test set: $T = \{(x_1, y_1)\cdots(x_m, y_m)\}$ includes m samples. Each sample has d attributes, $x_i \in \mathbb{R}^d$, $y_i \in L = \{1, -1\}$ is its corresponding category label.

The symbols used in the algorithm and their meanings are as follows:

N_{min}: The number of minority samples in training set;

N_{min}': The number of minority samples after oversampling;

K: The numerical value of the k near neighbor input by the user;

r: 2 k/3 < =r < =k, threshold to determine if a sample is noise, which is set by the user;

p: k/2 < =p < r, threshold to determine if a sample is unstable, which is set by the user;

q: k/3 < =q < p, threshold to determine if a sample is the boundary sample, which is set by the user;

c: Replication ratio of minority samples;

s: The number of neighbor samples selected when the new sample is synthesized;

Self-adaptive sample selection algorithm

When oversampling the minority samples in a medical diagnostic data set, the first step is to select samples to synthesize new samples. We use a self-adaptive sample selection algorithm to divide the minority samples into noise samples, unstable samples, boundary samples and stable samples according to their locations (the distribution of majority samples around them). Different samples are processed differently, so as to select the most suitable samples to synthesize new samples to ensure that the newly synthesized samples have the same characteristics as the original minority samples. The algorithm thought is as follows:

For all the minority samples, the number of the majority samples and the minority samples in the k near neighbors of each sample is calculated. Different operations are selected for samples according to different situations of each sample, including deleting the sample, doing nothing, copying the sample, or synthesizing the new sample using the sample. The specific selection methods are as follows (Fig. 1):

1. If the number of majority samples is greater than or equal to n(n is the threshold to determine if the minority sample is the noise, which can be set as needed, but the condition:2 k/3 < =n < =k must be satisfied.), that is, most majority samples are around this sample, then this minority sample is judged as the noise, and this sample is deleted from the sample set.
2. If the number of majority samples is less than n, greater than or equal to p(p is the threshold to determine if the minority sample is unstable, which can be set as needed, but the condition:k/2 < =p < n should be satisfied), that is, the most majority samples are around this sample, then it is judged that the minority sample is the unstable sample. It is not duplicated or synthesized, but the sample is not deleted from the sample set.
3. For data sets after eliminating the minority noises and unstable samples that cannot be used to synthesize samples, the replication ratio c of the minority samples is calculated. c = | (Number of minority samples after oversampling-Number of unstable samples)/(Number of minority samples in the training set-Number of noise samples-Number of unstable samples)|. Assuming that the original training set contains 200 minority samples, after oversampling, 500 minority samples, 20 noise samples and 30 unstable samples are expected to be obtained, then c = |(500–30)/(200–20-30)| = |3.13| = 3. That is, the replication ratio after oversampling is 3.
4. Calculate the newly generated samples number h needed by each minority sample, h = c-1. If the replication rate after oversampling is c = 3, then h = 3–1 = 2. Therefore, for each remaining minority sample, two new minority samples need to be regenerated around it.
5. If the number of majority samples is less than p, greater than or equal to q(q is the threshold to determine if the minority sample is the boundary sample, which can be set as needed, but the condition k/3 < =q < p should be satisfied), that is, the number of majority samples around this sample is basically the same as that of the minority samples, then it can be determined that this minority sample may be the boundary sample, and it is copied, and the copy number is h.
6. If the number of majority samples is less than q, then the operation of synthesizing new samples is conducted on this minority sample, and the number of new samples synthesized is h.

Distribution-sensitive sample synthesis algorithm

In the process of sample synthesis, medical diagnostic data require high quality of new synthetic samples, and new synthetic samples are required to have the same characteristics as the original minority samples. Therefore, we propose a distributed-sensitive sample synthesis algorithm when we do the new sample synthesis. Different sample synthesis methods are proposed according to the different distance from it to the surrounding minority samples. For the sample with close distribution, a near neighbor sample is selected to synthesize a new sample with it. The more the majority samples around the sample are selected, the

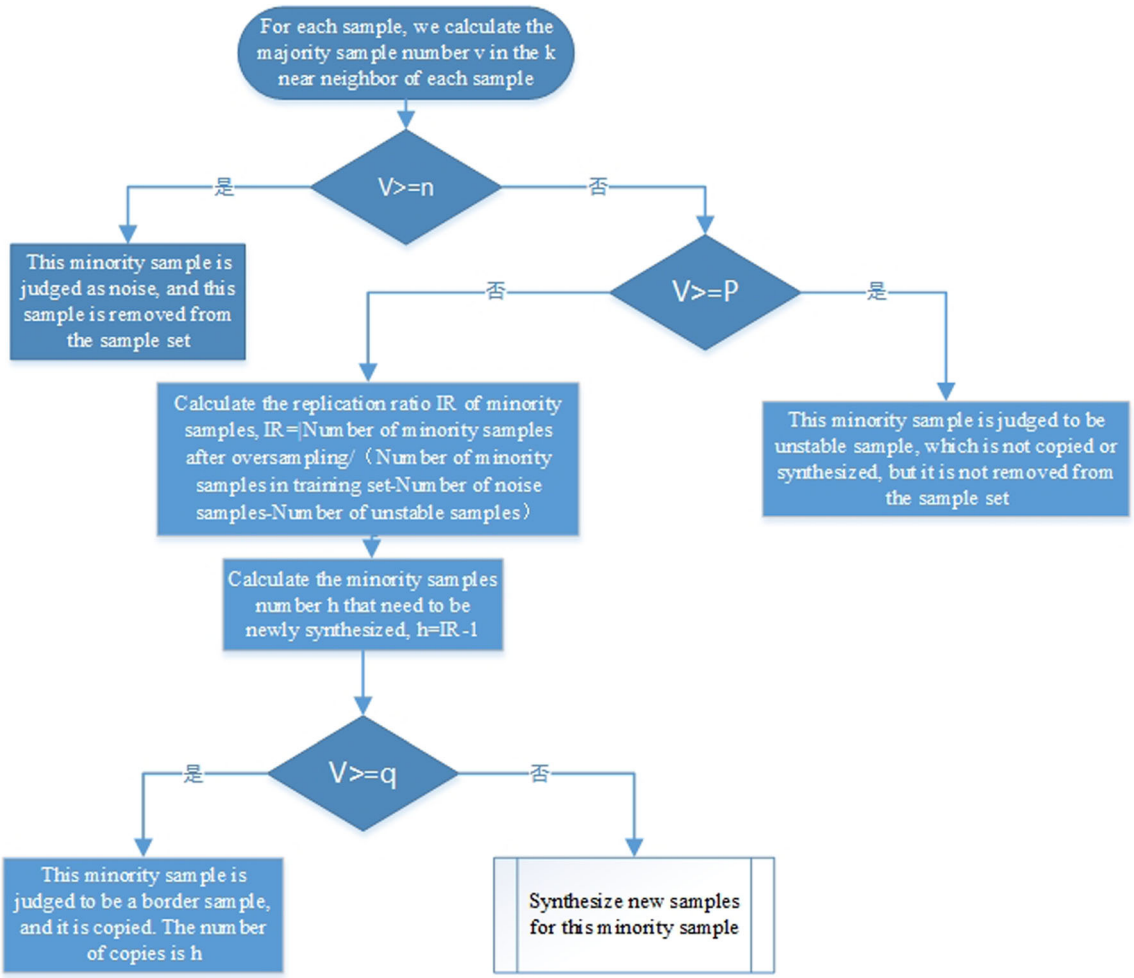


Fig. 1 Self-adaptive sample selection algorithm

lower the probability of the sample being selected is. For sparsely distributed samples, s samples are selected to synthesize new the sample with it. The situation that the sparse sample is combined with a neighboring sample that deviates from the normal value to generate the new sample should be avoided, so that the newly synthesized sample is more consistent with the distribution characteristics of the sample. The algorithm thought is as follows (Fig. 2):

1. Calculate the average distance d from all the minority samples to the surrounding k near neighbors (Select only surrounding k minority near neighbors).
2. If the average distance d_i between the minority sample i and the surrounding k near neighbor is less than or equal to d, that is, this minority sample is very close to the surrounding minority samples, then a sample is selected from the surrounding minority samples to synthesize a new sample with this minority sample. The selection and synthesis methods are as follows:

[1] Traverse the k nearest neighbor samples of i. For each minority near neighbor j, k nearest neighbors are calculated. At this time, the minority samples and majority samples around j are all considered, because you need to determine whether j is suitable for synthesizing new minority samples.

[2] For k nearest neighbor samples of each minority near neighbor j, the ratio of minority samples to majority samples is calculated. The larger the ratio, the more the minority samples around j. Therefore, the probability of choosing j for synthesizing new samples with i should be larger. For this reason, the ascending sort is carried out for the corresponding ratio r_j of each minority near neighbor j. The larger the ratio, the larger the ordinal number.

[3] Select a random number between 0 and 1, square it, and then multiply it by k to get the probability of selecting j.

[4] According to the number of new samples h required to be synthesized at the beginning, new minority

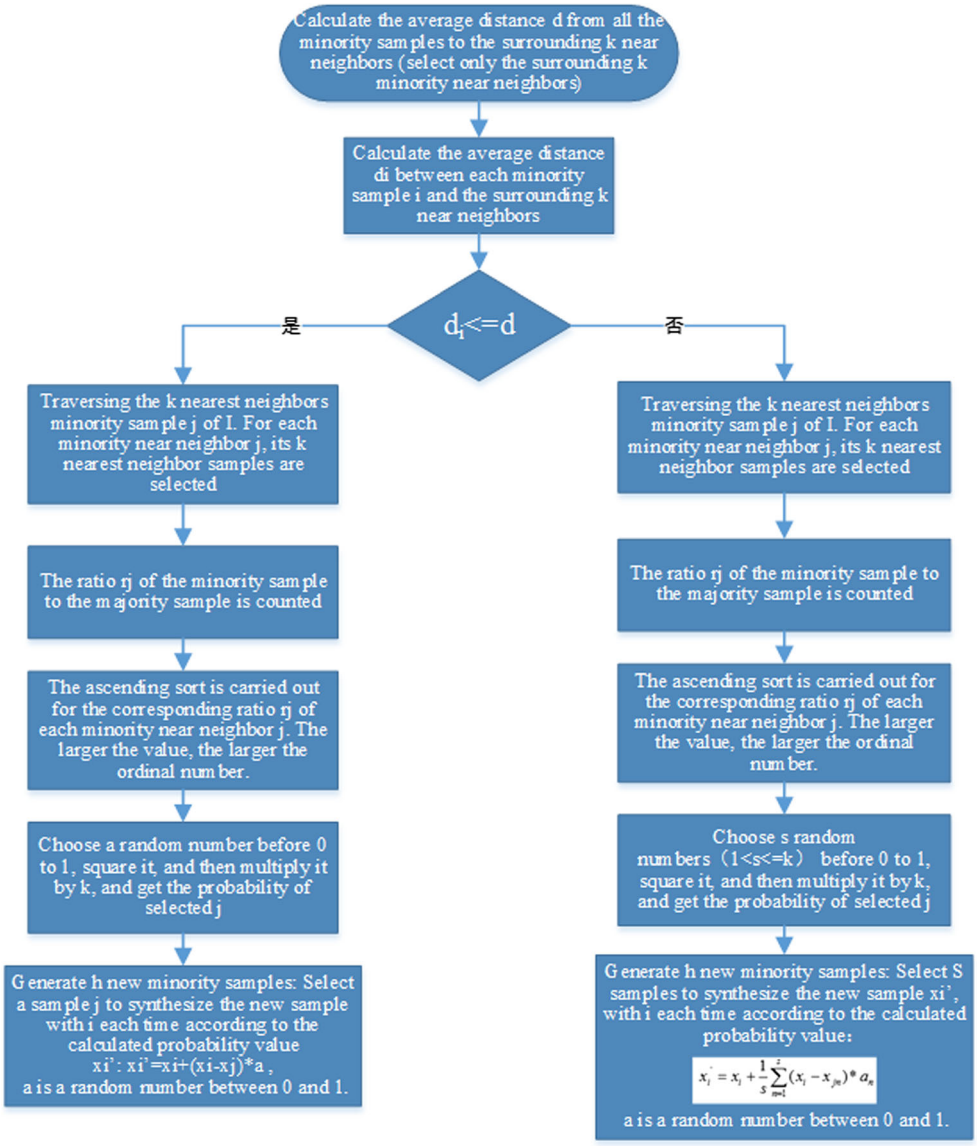


Fig. 2 Distribution-sensitive sample synthesis algorithm

samples will be synthesized iteratively. According to the probability value calculated by 3, a sample j is selected to synthesize the new sample x_i' with i. The synthetic method is: $x_i' = x_i + (x_i - x_j)*a$, in which a is the random number between 0 and 1.

3. If the average distance from the minority sample i to the surrounding k near neighbors: $d_i > d$, that is, this minority sample is very loose with the surrounding minority samples, then s samples(s can be set as required, but it needs to satisfy $1 < s < = k$) are selected from the surrounding minority samples to synthesize the new sample with this sample. That is to say, for the minority sample which is far away from the surrounding samples, we should select as many samples as possible to generate the new sample together with it, so as not to choose only one sample to synthesize the new sample, causing greater deviation, which is inconsistent with the original data. The selection and synthesis methods are as follows:

[1] Traverse the k nearest neighbor sample of i. For each minority near neighbor j, its k nearest neighbor samples are calculated. At this time, the minority samples and majority samples around j are taken into account, because it is necessary to determine whether j is suitable for the synthesis of new minority samples.

[2] For k nearest neighbor samples of each minority near neighbor j, the ratio of minority samples to majority samples is calculated. The larger the ratio, the more the minority samples around j. Therefore, the probability of choosing j for synthesizing new samples with i will be larger. For this reason, the ascending

Table 1 Basic information of dataset

ID	Data set name	Total number of samples	Attribute number	Number of minority samples
bre	breast-cancer	35,305	9	329
lun	lung-cancer	21,783	57	197
hea	heart disease	50,391	24	289
der	Dermatology	15,610	32	213

order is carried out for the corresponding ratio r_j of each minority near neighbor j, and the ordinal number is x. The larger the value, the larger the ordinal number.

[3] Select a random number between 0 and 1, square it, and multiply it by x to get the probability of choosing j.

[4] According to the number of new samples h required to be synthesized at the beginning, new minority samples will be synthesized iteratively. According to the probability value calculated by 3, s samples j1–js are selected to synthesize the new sample x_i' with i, and the synthetic method is:

$$x_i' = x_i + \frac{1}{s}\sum_{n=1}^{s}\left(x_i - x_{jn}\right) * a_n$$

a_n is a random number between 0 and 1.

Experiment analysis

Experiment data

Our Experiment data was based on real medical data sets, from more than 100,000 real medical data collected by Xiangya Hospital. In order to compare with other algorithms for testing, the data set format we extracted was the same as the four data sets on disease detection in the UCI public data set [12]: breast-cancer, lung-cancer, heart disease, and Dermatology. The basic information of the data set is shown in the table below (Table 1). From the distribution of data sets, it can be seen that medical data sets are seriously imbalanced data sets, and majority class samples are much larger than minority class sample. Through the test, it was found that the imbalanced ratio was too large, the replication ratio of minority samples was too high, which would lead to the reduce of the performance of the learning algorithm, because too many replication of minority samples would lead to a large number of invalid samples. Since this paper focuses on the over-sampling technique of minority samples and doesn't involve the under-sampling of the majority samples, we randomly select the majority samples with the imbalance ratio of about 1:10. For example, there are 329 minority samples of breast-cancer, and we keep 3290 majority samples for follow-up.

During the experiment, the medical data set was divided into four parts, 4/5 was used as training set to train the classification algorithm, and 1/5 was used to test the performance of the classification algorithm.

Results and analysis

The parameter settings in oversampling algorithms of minority class have an important impact on the performance of the algorithm. We tested the performance of the algorithm under different parameter settings in the algorithm. These parameters included the nearest neighbor number k, and the replication ratio c of minority examples. The number of samples selected when synthesizing a new sample was s, and the reference value of parameter setting for actual algorithm was given by testing. Then we compared the algorithm proposed in this paper with other data oversampling algorithms to verify the effect of the proposed algorithm.

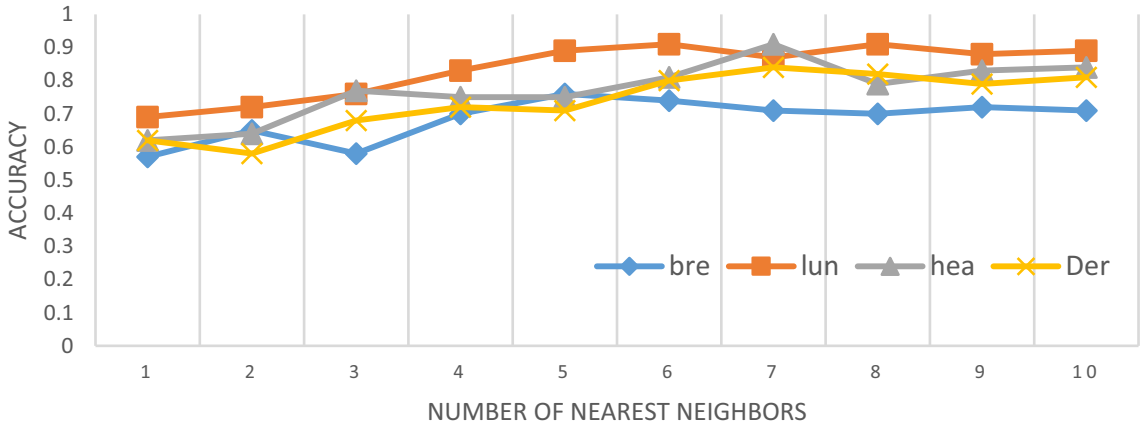


Fig. 3 Influence of nearest neighbor number k on algorithm accuracy

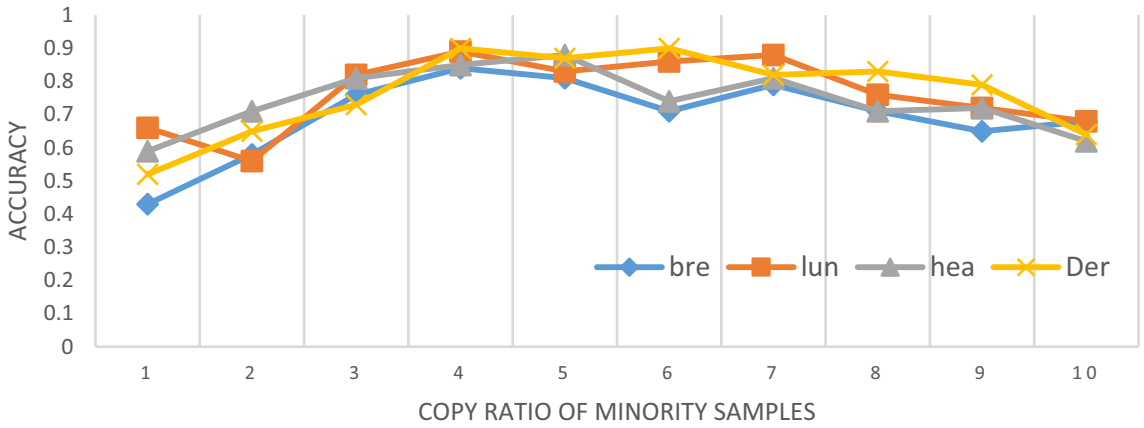


Fig. 4 Effect of replication ratio c of minority samples on accuracy of algorithm

The influence of the number of neighbors k

One of the main parameters in the oversampling algorithm is the number of near neighbors k, that is how many nearest neighbors are selected each time for subsequent calculations. The optimal values of k values for different data sets are also different. In the SMOTE algorithm, the default is k = 5. The impact of changes in the medical dataset k on the classification effect was tested. In the case where the other settings are unchanged, the classification accuracy change on each data set when k is gradually increased from 1 to 10 is as shown in Fig. 3. It can be seen from Fig. 3 that when k takes about 6 and 7, the accuracy of the algorithm reaches a good value. After that, the change of k value does not have much influence, and the overall stability is stable. The k values that achieve the best results for different data sets are different. In addition, the value of the sample number s selected when synthesizing a new sample is also related to the k value. As the k value increases, the performance of the algorithm decreases, and the best k value can be determined by testing in a specific application. In the subsequent experiments, in order to facilitate comparison with other algorithms, we uniformly set k = 6.

The effect of the replication ratio c of minority samples

The replication ratio of minority samples determines how many new samples need to be generated for each minority sample. If the replication ratio of minority class samples is too high, the performance of the learning algorithm will be degraded. Therefore, in the resampling algorithm, the over sampling of minority classes and the undersampling of majority classes are adopted. On the one hand, some minority samples are synthesized, and on the other hand, majority samples are selectively deleted, so that the final generated sample set can improve the performance of the classification learning algorithm. Therefore, it is necessary to test the effect of replication of minority samples on classification performance. Using medical dataset, we tested the changes in classification accuracy on each dataset when the replication ratio was from 1 to 10. The results are shown in Fig. 4. As can be seen from Fig. 4, in the set of medical diagnostic samples, the accuracy of the algorithm is better when c is 4 or 5, and then the increase of c has little effect on the accuracy of the algorithm. But when c is greater than 7, the accuracy of the algorithm begins to decline. The reason may be that too many duplicate minority samples will lead to a large number of invalid samples, and the

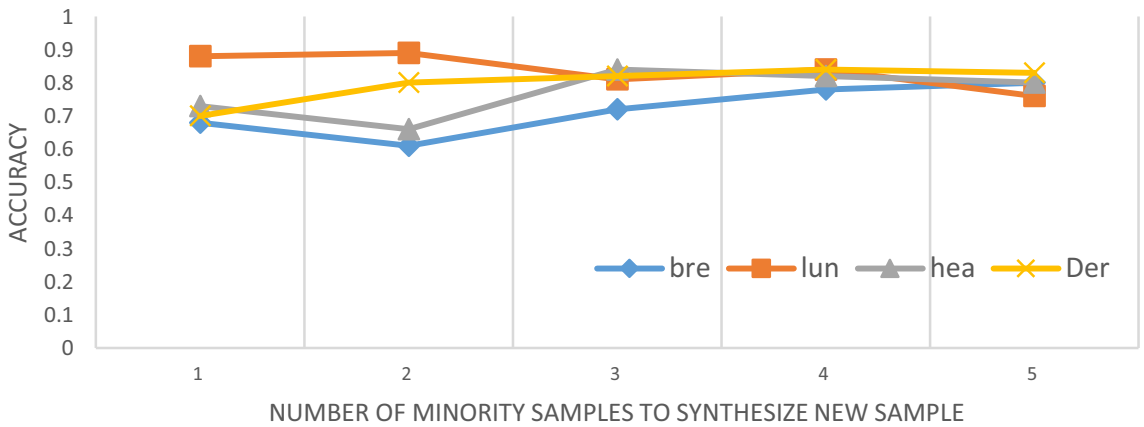
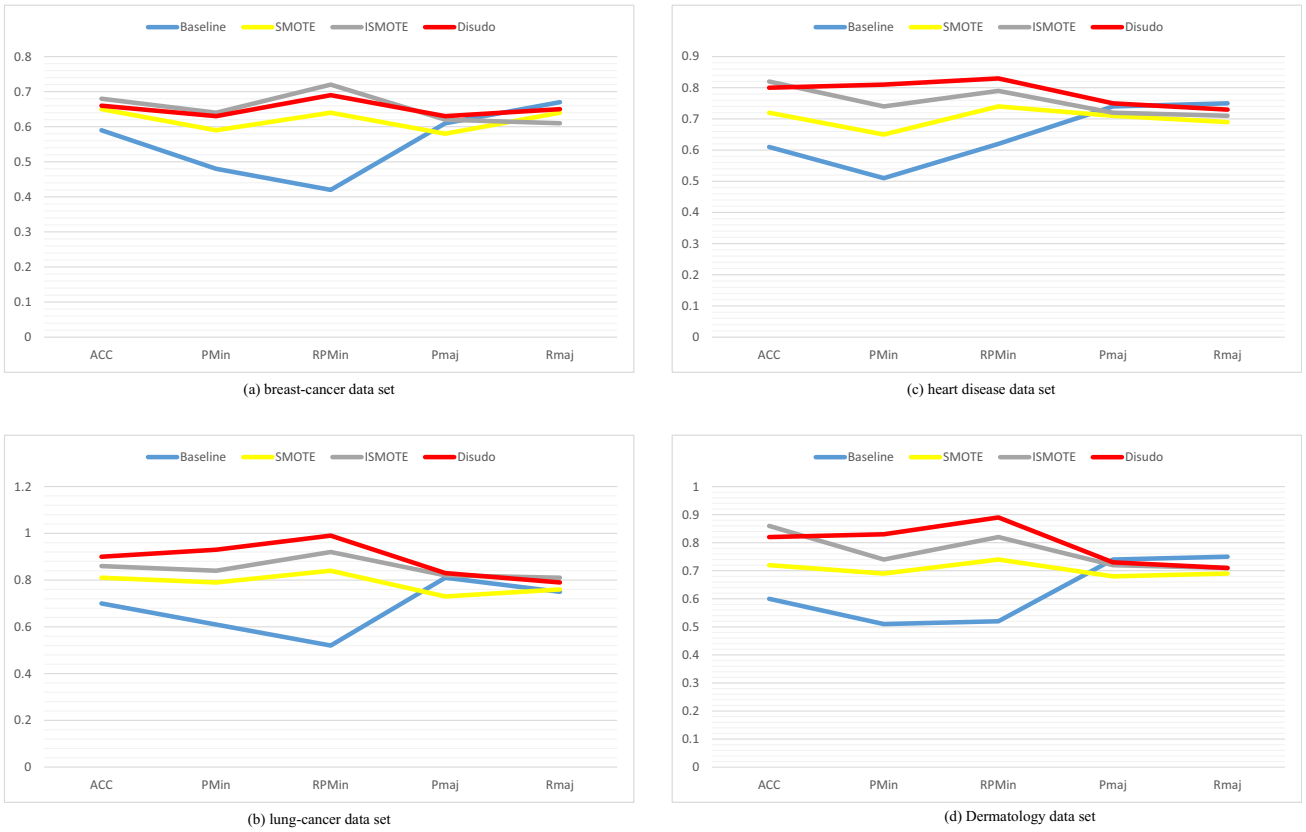


Fig. 5 The effect of the number of minority samples s selected in the synthesis of new samples on the accuracy of the algorithm



(a) breast-cancer data set

(c) heart disease data set

(b) lung-cancer data set

(d) Dermatology data set

Fig. 6 Comparison test of different algorithms

performance of classification learning algorithm will be reduced. The c values for achieving the best results of different data sets are different. The best c value can be determined by testing in specific application. In subsequent experiments, we uniformly set c = 4.

Influence of the number of minority samples s selected in the synthesis of new samples

In the synthesis of the new sample, if a sample is far away from the surrounding minority samples, s surrounding samples are selected to generate the new sample with it, so as to avoid selecting only one sample to synthesize a new sample, causing a large deviation, which is in consistent with the original data. Based on the medical diagnostic data set, we tested the classification accuracy of s from 1 to 6 data sets. Because when the number of synthetic samples exceeds 6, the performance of the algorithm is not only reduced, but the accuracy is also reduced. Therefore, our tests only performed tests of 1–6. The results are shown in Fig. 5. As can be seen from Fig. 5, the optimal values of s are different on different data sets. On the breast-cancer data set, the performance of the algorithm is best when s takes 5, and on the heart disease and Dermatology data sets, the accuracy is the best when s is 3–4 For the breast-cancer data set, the classification accuracy is the best when s takes 2. From the test results, for the data set, the fewer the attributes, the larger the number of s needed to synthesize new samples. This may be because data sets with fewer attributes are more prone to bias when synthesizing new samples.

Comparison with other existing algorithms

In order to test the validity of the proposed method, Distribution-sensitive unbalanced data oversampling (Disudo) is compared with Baseline, SMOTE and ISMOTE. Fig. 6 lists five index values for four different algorithms on four datasets, corresponding to Acc: Accuracy, PMin: Precision of minority, RMin: Recall of minority, PMaj: Precision of majority, and RMaj: Recall of majority.

It can be seen from the test results that the Disidu method performs better on all data sets, especially in the accuracy and recall rate of minority classes. The Disidu method has obvious advantages. In medical diagnostic applications, more emphasis is placed on the recall rate of minority types, that is, as many minority samples as possible (ie, sick cases) should be found, even if there are cases that are not ill, they can be excluded through subsequent examinations. But you can't miss the sick case. For majority classes, the accuracy of each algorithm is relatively close. Only oversampling of minority class is performed in this test, the undersampling of majority

classes is not performed, so the classification performance of majority classes is not greatly affected.

Conclusions

In this paper, for the poor performance of the learning algorithm caused by the serious shortage of minority samples in learning algorithm for medical diagnosis, the distribution-sensitive imbalanced data oversampling algorithm is proposed. The algorithm allows the user to input the required number of minority samples, and determines the number of newly generated samples, according to the user, so that the method can adapt to different classification learning algorithms. The algorithm has the following characteristics:

1. When oversampling minority samples in medical diagnostic data set, the self-adaptive sample selection algorithm is used to classify the minority samples according to the location of the minority samples (the distribution of the majority class samples around them). The minority samples are divided into noise samples, unstable samples, boundary samples and stable samples. In order to ensure that the newly synthesized samples have the same characteristics as the original few samples, different processing methods are used for different samples to select the most suitable samples to synthesize new samples.
2. In the process of sample synthesis, the distributed-sensitive sample synthesis algorithm is adopted, and different sample synthesis methods are selected according to the distance between the sample and surrounding minority samples. For the samples with tight distribution, a near neighbor sample is selected to synthesize a new sample with it, and the more the surrounding majority samples of the samples, the lower the probability of being selected. For the samples with sparse distribution, s samples are selected to synthesize new samples with it, so as to avoid synthesizing a new sample by the sample with a sparse distribution and a neighbor sample deviating from the normal value, which will make the newly synthesized samples more consistent with the sample distribution characteristics.

The test of real medical diagnosis data shows that compared with the existing sampling algorithm, the accuracy of classification learning algorithm is improved, especially the precision and recall rate of minority classes.

Our future works is as follow:

1. It is very important of parameters selection for the algorithm proposed in this paper. We have discussed the impact of replication ratio of minority samples with real world experimental data. However, the process of parameters selection should be adaptive and can be automatically adjusted according to the distribution of input sample set, which makes the algorithm proposed in this paper more flexible for different sample sets.
2. The algorithm in the paper mainly aims at the preprocessing of imbalanced data for two class. In reality, many big data are multi-class. One method is to decompose the multi-class problem into two class problem, and the other is to study the unbalanced data re-sampling algorithm for multi-class problem. The multi-class big data resampling algorithm still requires further study.

Funding Funded by NSFC (No. 61672020), the national key research and development program[2016YFB0800303], Supported by DongGuan Innovative Research Team Program.

Compliance with Ethical Standards

Declaration of Conflict of Interest Weihong Han, Zizhong Huang, Shudong Li and Yan Jia declare no conflict of interest directly related to the submitted work.

Ethical Approval This article does not contain any studies with human participants performed by any of the authors.

References

1. Sun, Y., Wong, A. K., and Kamel, M. S., Classification of imbalanced data: A review. *Int. J. Pattern Recogn. Artif. Intell.* 23(04): 687–719, 2009.
2. Garcia, S., and Herrera, F., Evolutionary undersampling for classification with imbalanced datasets: Proposals and taxonomy. *Evol. Comput.* 17(3):275–306, 2009.
3. Lopez, V., Fernandez, A., Garcia, S., Palade, V., and Herrera, F., An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Inform. Sci.* 250:113–141, 2013.
4. Wang, S., and Yao, X., Multiclass imbalance problems: Analysis and potential solutions. *IEEE Trans. Syst. Man Cybernet. B: Cybernet.* 42(4):1119–1130, 2012.
5. Zhang, Z., Krawczyk, B., Garcia, S., Rosales-Perez, A., and Herrera, F., Empowering one-vs-one decomposition with ensemble learning for multi-class imbalanced data. *Knowl.-Based Syst.* 106: 251–263, 2016.
6. Krawczyk, B., Learning from imbalanced data: Open challenges and future directions. *Progress Artif. Intell.* 5(4):221–232, 2016.
7. Chawla, N. V., Bowyer, K. W., Hall, L. O. et al., SMOTE: Synthetic minority over-sampling technique [J]. *J. Artif. Intell. Res.* 16(1): 321–357, 2002.
8. Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap C., Safe-level-SMOTE: Safe-level-synthetic minority over-sampling TEchnique for handling the class imbalanced problem[C]// Pacific-Asia conference on advances in knowledge discovery and data mining. Springer-Verlag, :475–482, 2009.
9. Han, H., Wang, W. Y., and Mao, B. H., Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning[A]. *Int. Conf. Intell. Comput.* 3644(5):878–887, 2005.

10. Bunkhumpornpat, C., Sinapiromsaran, K., and Lursinsap, C., DBSMOTE: Density-based synthetic minority over-sampling TEchnique[J]. *Appl. Intell.* 36(3):664–684, 2012.
11. Bunkhumpornpat, C., and Sinapiromsaran, K. CORE: core-based synthetic minority over-sampling and borderline majority under-sampling technique.[M]. Inderscience Publishers, 2015.
12. Bennin, K.E. and Keung, J. et al., MAHAKIL: Diversity based Oversampling Approach to Alleviate the Class Imbalance Issue in Software Defect Prediction[J]. IEEE Transactions on Software Engineering, (99) :1–1, 2017.
13. Mathew, J., Pang, C. K., Luo, M. et al., Classification of imbalanced data by oversampling in kernel space of support vector machines[J]. *IEEE Trans. Neural Netw. Learn. Syst.* 29(9):4065–4076, 2018.
14. Douzas, G., Bacao, F., and Last, F., Improving imbalanced learning through a heuristic oversampling method based on K-means and SMOTE[J]. Information Sciences, 2018.
15. Jin, S., and Pedersen, T., Duluth UROP at SemEval-2018 task 2: Multilingual emoji prediction with ensemble learning and oversampling[J]. 2018.