2020 19th International Symposium on Distributed Computing and Applications for Business Engineering and Science (DCABES)

Denoise-Based Over-Sampling for Imbalanced Data Classification

Wang Dan
School of Artificil Intelligence and Computer Science
Jiangnan University
Wuxi,China
e-mail: 6181914006@stu.jiangnan.edu.cn

Liu Yian
School of Artificil Intelligence and Computer Science
Jiangnan University
Wuxi,China
e-mail: lya_wx@jiangnan.edu.cn

***Abstract*—Imbalanced data classification has always been a hot topic in traditional machine learning. The usual method is oversampling. Its main idea is to randomly synthesize the new minority samples between the minority samples and their neighboring samples, to put the data in a particular state of equilibrium.. The existing improved methods have improved the classifier's performance to some extent, but most of the focus is on the minority sample. In this paper, a denoise-based over-sampling method (DNOS) is proposed, which performs different denoise processes for the majority and minority samples. Then, it is combined with ADASYN to oversampling the data. Experimental results show that DNOS has a better classification effect than ADASYN.**

Keywords-imbalanced data; over-sampling; denoise processes

I. INTRODUCTION

Imbalanced data refers to the data of majority samples much more than minority samples.Imbalanced data are widely used in equipment fault detection[1],credit card fraud, disease prediction[5],text classification[4],price prediction[10] and other social fields. However, the traditional machine learning classification method is built on a balanced data set. When the data is imbalanced, such a classifier tends to favor the majority samples, which will affect the classification accuracy of the minority samples. In view of this situation, scholars at home and abroad have put forward many methods to improve the imbalanced data classification problems[6][7]. There are two main methods in the data level — oversampling and undersampling. Undersampling is removing some samples from the majority to bring the data closer to balance. For example, Koziarski M proposed a new radial-based undersampling algorithm in 2020[8], which uses the concept of mutual potential to guide the possibility of the RBO oversampling process, but this method may delete important information. Oversampling brings the data to be balanced by generating a few new class samples. The conventional method is SMOTE[2] and ADASYN[3].In the latest research, Gu X et al. proposed to generate high-quality synthetic samples from minority samples observed empirically[9], Fan X et al. proposed a margin-based over-Sampling method[11]. These methods can effectively balance the data and improve the performance of the classifier, but the influence of noise samples still exists. Therefore, the method of noise reduction oversampling is proposed in this paper to reasonably reduce noise samples and make the synthesized samples more conducive to a classification decision. The comparison among multiple data sets shows that this method has some advantages.

II. PRELIMINARY KNOWLEDGE

A. ADASYN

In 2008, HE·H et al. proposed that ADASYN is called adaptive synthetic sampling[3]. Different from SMOTE, instead of making the same number of new samples for each minority sample, ADASYN use the specific method to calculate the number of new samples for each minority sample. First, calculate the number of samples to be synthesized G. For each minority class x_i calculate proportion r_i:$r_i = \Delta_i/k$, k is the number of nearest neighbors of x_i, Δ_i is k neighbor samples belong to a majority sample number of the class. The distribution ratio $\tilde{r}_i$ was obtained after the normalization of r_i, and then G was distributed to each minority sample x according to the distribution ratio $g_i = \tilde{r}_i \times G$. New samples were synthesized according to g_i, and finally, the data balance was achieved.

B. Classification evaluation criteria

F1-score—Among imbalanced data evaluation criteria, f1-score is a comprehensive evaluation criterion, which is the harmonic average of accuracy and recall rate.

G-mean—This indicator is designed for unbalanced data. It consists of two sub-metrics -- TPR and TNR. TPR reflects the classifier's sensitivity to the minority samples, while TNR demonstrates the majority's recognition performance. When

2473-3636/20/$31.00 ©2020 IEEE
DOI 10.1109/DCABES50732.2020.00078

the accuracy of these two values is high, the g-mean value will increase. Therefore, the g-mean value can more comprehensively evaluate the performance of the classifier. The larger the g-mean value is, the better the performance of the classifier will be.

AUC—It is the area under the ROC curve, generally between 0 and 1. ROC is often used to evaluate the advantages and disadvantages of dichotomies. The higher the AUC value is, the better the effect of the corresponding model will be, and the higher the classification accuracy will be.

C. Raise issue

ADASYN calculates the distribution ratio for each minority class sample, which is related to the number of majority classes in the nearest neighbor. The more the majority classes in the nearest neighbor of the minority class sample, the larger its distribution ratio will be. The new samples will be synthesized from this sample. However, the probability that such minority class samples are noise samples is also very high. If the selected minority samples are noise, the newly synthesized samples will continue to interfere with the classifier, resulting in classification difficulties. Not only the minority samples， but also the majority have noise. Since the majority of samples' cardinality is several times that of the minority, the noise samples are not very small compared with the minority, such noise will also affect the classification performance of the model. Therefore, this paper proposes a method of denoise oversampling (DNOS) for imbalanced data.

III. DNOS

Assuming that the training set is D when the label is 0,1, it refers to a minority class samples $D_{min} = (x_{B1}, x_{B2}, \ldots, x_{Bn'})$ and the of a majority class $D_{maj} = (x_{A1}, x_{A2}, \ldots, x_{An''})$, n' is the number of the minority samples in the training set, n'' is the number of the majority samples. The method of DNOS is: denoising the minority samples, obtaining the denoising training set D_{min_new} of the minority, and denoising the majority, getting the denoising training set D_{maj_new}, combining them into a new training set, and then using ADASYN to over-sample the original training set.

A. Minority samples denoise

Judge for each minority sample x_{Bp}. If the sample is closer to the majority class sample, and there are many majority class samples around it, it is regarded as a noise sample. Specific steps are as follows:

Input: Minority sample D_{min} and majority samples D_{maj} in the training set.

Output: Minority samples D_{min_new} after denoising.

- Use k-means to find the clustering centers min_center and maj_center of the minority and majority samples, respectively.
- For each minority sample x_{Bp}, $p = 1,2,\ldots,n'$, find the distances S_min_p and S_maj_p to min_center and maj_center, respectively.
- For each sample x_{Bp} of the minority, k neighbors are calculated by Euclidean distance, Δ_p is the number of samples belonging to the majority class among the neighbors.
- For x_{Bp}, if $S\_maj_p < S\_min_p$ and $\Delta_p > k/2$, then it is a noise sample because the sample is closer to the majority cluster center, and more than half of its neighbors are majority class.
- Delete these noise samples from the minority samples to obtain new minority samples D_{min_new}.

B. Majority samples denoise

In 3.1, the clustering center maj_center of the majority class has been obtained. Let the circle be the center of maj_center and r the radius. The samples beyond the circle are noise.

Input: Majority samples D_{maj} in training set, and majority clustering center maj_center.

Output: Majority samples D_{maj_new} after denoising.

- For each majority sample $x_{Aq}, q = 1,2,\ldots,n''$, the distance dif_q to maj_center, and find the average distance, the formula is:

$$dif\_mean = \frac{\sum_{q=1}^{n''} maj\_center - Dmaj_q}{n''}$$

- Let $r = 2 \times dif\_mean$, if $dif_q > r$, indicate that x_{Aq} is not in the circle, and delete x_{Aq} from the majority samples. To obtain the new majority samples, D_{maj_new}.

IV. EVALUATION OF EXPERIMENTAL RESULTS

The experiment reflects the feasibility of the improvement in two ways: one is to compare the sampling effect in the form of visualization, the other is to analyze the impact of different methods on the classifier according to the classification evaluation index.

A. Comparison of sampling effects

Comparing the sampling effect of ADASYN with the improved method proposed in this paper, we can draw out the feasibility of the enhanced approach in this paper. To facilitate visualization, the manual data set in references[12] is used in this paper, and the sklearn package of python3.7 is used to randomly generate two sets of Gaussian data. The majority sample number is 500, and the sample center is [3.5,3.5]; the minority sample number Is 50, the sample center is [1,1]; the imbalance ratio is 10.

(a) in Fig. 1 is a binary classification data set randomly generated according to the above conditions, the majority class is significantly more than the minority class data. The effect picture after ADASYN oversampling in Figure (b) shows that most of the data synthesized by the algorithm is concentrated at the boundary between the majority class and the minority class, and there are many synthesized noise samples. Figure (c) is the improved noise reduction and oversampling algorithm in this paper. The majority of data sets are more concentrated. The synthesized minority data is not only concentrated at the junction of the two types of data sets, while reducing certain noise data and avoiding synthesis of new noise data.

B. Experimental data and results

This paper selects five data sets from the imbalanced data set in the KEEL database[13] to analyze and compare the experimental results. This paper would be taken to ensure that the imbalanced ratio between the training set and the test set is consistent to verify the validity of DNOS. At the same time, this article uses python's sklearn package to randomly divide the data, dividing 70% of the data into the training set, and the remaining 30% is the test set. Table 1 shows the structure of 5 data sets, and Table 2-4 is the collation of the classification results.

From the experimental results of Table 2, Table 3, and Table 4, it can be seen that the F1-score and G-mean values of the improved method proposed in this paper are better than ADASYN. The yeast-2_vs_4 data results show that the G-mean and F1-score values are not as good as unsampled after ADASYN sampling. The improved algorithm does not have this situation, which indicates that DNOS has specific feasibility.

TABLE I. 5 KEEL data set structures

DATASET	ATTRIBUTES		
	MAJORITY	*MINORITY*	*IMBALANCE RATE*
pageblocks0	4913	559	0.1138
newthyroid2	180	35	0.1944
segment0	1979	329	0.1662
yeast2_vs_4	477	51	0.1069
yeast3	1321	163	0.1233

TABLE II. Comparison of G-mean values of classification results of different methods

DATASET	ALGORITHM		
	SVM	*ADASYN+SVM*	*DNOS+SVM*
pageblocks0	0.2438	0.6966	0.7460
newthyroid2	0.5222	0.9229	0.9622
segment0	0.7607	0.9486	0.9504
yeast2_vs_4	0.7718	0.8652	0.9451
yeast3	0.8490	0.9337	0.9340

TABLE III. Comparison of F1-Score values of classification results of different methods

DATASET	ALGORITHM		
	SVM	*ADASYN+SVM*	*DNOS+SVM*
pageblocks0	0.1104	0.2801	0.3165
newthyroid2	0.4285	0.7333	0.8461
segment0	0.7215	0.7791	0.7854
yeast2_vs_4	0.7200	0.6666	0.8000
yeast3	0.7792	0.7378	0.8131

TABLE IV. Comparison of AUC values of classification results of different methods

DATASET	ALGORITHM		
	SVM	*ADASYN+SVM*	*DNOS+SVM*
pageblocks0	0.5287	0.7297	0.7614
newthyroid2	0.6363	0.9259	0.9629
segment0	0.7882	0.9495	0.9512
yeast2_vs_4	0.7964	0.8678	0.9452
yeast3	0.8584	0.9337	0.9351

V. CONCLUSION

This paper proposes an improved noise reduction sampling algorithm based on ADASYN. To reduce the impact of noise data on the classification effect, the DNOS introduces noise reduction methods for minority samples and majority samples, and filters and removes noise samples. Compared with the traditional oversampling algorithm ADASYN Experimental classification effect on SVM classifier. However, as the number of data increases, the running time of DNOS is longer than ADASYN. Therefore, the future research direction will start by shortening the sampling time and other classification algorithms to improve the algorithm's efficiency while ensuring good classification results.

REFERENCES

[1] Yang Z, Tang W H, Shintemirov A, et al. Association Rule Mining-Based Dissolved Gas Analysis for Fault Diagnosis of Power Transformers[J]. IEEE Transactions on Systems Man & Cybernetics Part C Applications & Reviews, 2009, 39(6):597-610.

[2] Chawla N V, Bowyer K W, Hall L O, et al. SMOTE: Synthetic Minority Over-sampling Technique[J]. Journal of Artificial Intelligence Research, 2011, 16(1):321-357.

[3] HE H, BAI Y, GARCIA E A, et a1. ADASYN : adaptive synthetic sampling approach for imbalanced learning[C]// Proceedings of the 2008 IEEE International Joint Conference on Neural Networks (IEEE World Congress on Computational Intelligence). Piscataway, NJ : IEEE, 2008 : 1322—1328.

[4] Xu R, Chen T, Xia Y, et al. Word Embedding Composition for Data Imbalances in Sentiment and Emotion Classification[J]. Cognitive computation, 2015, 7(2):1-15.

[5] Mohammed, Khalilia, Sounak, et al. Predicting disease risks from highly imbalanced data using random forest[J]. Bmc Medical Informatics & Decision Making, 2011.

[6] Mullick S S, Datta S, Dhekane S G, et al. Appropriateness of Performance Indices for Imbalanced Data Classification: An Analysis[J]. Pattern Recognition, 2020, 102:107197.

[7] Guolong, Chen, Yong. Imbalanced data classification based on scaling kernel-based support vector machine[J]. Neural Computing & Applications, 2014.

[8] Koziarski M. Radial-Based Undersampling for Imbalanced Data Classification[J]. Pattern Recognition, 2020, 102:107262.

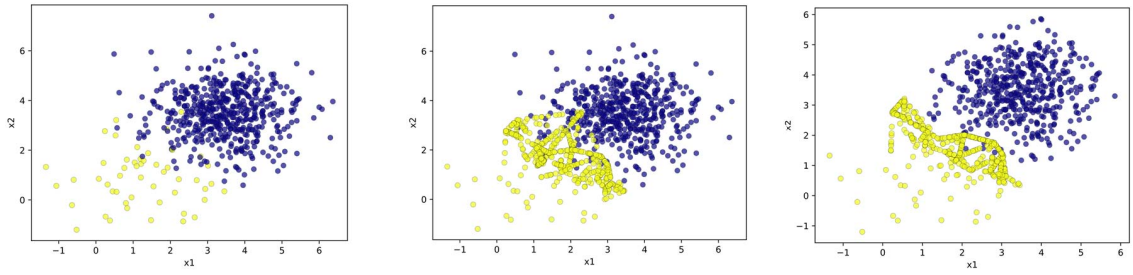
[9] Gu X, Angelov P P, Soares E A. A self‐adaptive synthetic over‐sampling technique for imbalanced classification[J]. International Journal of Intelligent Systems, 2020, 35(6).

[10] Hasmita S, Nhita F, Saepudin D, et al. Chili Commodity Price Forecasting in Bandung Regency using the Adaptive Synthetic Sampling (ADASYN) and K-Nearest Neighbor (KNN) Algorithms[C]// 2019 International Conference on Information and Communications Technology (ICOIACT). 2019.

[11] Fan X, Tang K7, Weise T. Margin-Based Over-Sampling Method for Learning from Imbalanced Datasets[C]// Advances in Knowledge Discovery & Data Mining-pacific-asia Conference. Springer, Berlin, Heidelberg, 2011.

[12] Barabási A L , Albert R.Emergence of scaling in random networks[J].Science, 1999, 286（5439）: 509-512.

[13] Alcalá-Fdez J, Fernández A, Luengo J, et al.KEEL datamining software tool : data set repository, integration of algorithms and experimental analysis framework[J].Journal of Multiple-Valued Logic & Soft Computing, 2011, 17 : 255-287.



(a) (b) (c)

Figure 1. Sampling comparison figure: (a) raw data; (b)Sampled data with ADASYN; (c) Sampled data with DNOS.