A novel αDistance Borderline-ADASYN-SMOTE algorithm for imbalanced data and its application in Alzheimer's disease classification based on Dense Convolutional Network

Yun Feng1,3, Jingjiao Li2*

1School of Computer Science and Engineering, Northeastern University, Shenyang, Liaoning Province, 110169, China
2College of Information Science and Engineering, Northeastern University, Shenyang, Liaoning Province, 110819, China
3School of Information Engineering, Shenyang University, Shenyang, Liaoning Province, 110044, China
*Corresponding author’s e-mail: lijingjiao@ise.neu.edu.cn

Abstract. The classification problem of imbalanced data has become a very important issue in the fields of machine learning and data mining. At present, relatively effective oversampling methods for processing imbalanced data include SMOTE, Borderline-SMOTE, and ADASYN. These algorithms have their own advantages; however, they do not adequately consider the distance factor, which is an important factor for balancing data precisely and reducing the misclassification probability of a minority boundary sample. Therefore, a new algorithm, αDistance Borderline-ADASYN-SMOTE algorithm, is proposed in the paper by combining the optimized Borderline-SMOTE algorithm with the optimized ADASYN algorithm. In the new algorithm, both the amount and the distance distribution of the nearest neighbor samples are considered. A few formulas are created to realize the algorithm. After being balanced by the algorithm, the data obtained from ADNI data set is trained, verified and tested by the Dense Convolutional Network. The experimental results show that the new model improves the classification performance of the Alzheimer's disease.

1. Introduction

The current methods for solving the problem of imbalanced data classification can be divided into two categories. One is on data level, changing the data set samples by sampling methods; the other is on learning algorithm level.[1] Furthermore, the sampling methods include oversampling and undersampling methods. And some of the existing effective oversampling techniques include SMOTE (Synthetic Minority Oversampling Technique), Borderline-SMOTE, and ADASYN (Adaptive Synthetic Sampling).[2] The SMOTE is a popular oversampling technique. It adds a number of synthetic minority class samples along line segments between existed minority class samples and some of their nearest minority class neighbors.[3] Later, many scholars propose a few improved algorithms based on the SMOTE algorithm. For example, the Borderline-Smote algorithm, which synthesizes new samples only for the boundary points of a minority class,[4] avoiding to generate redundant samples; the ADASYN algorithm, whose key idea is to use density distribution as a criterion to automatically determine the number of synthetic samples for each minority sample.[5]
However, the above algorithms have their own limitations, when balancing data, they pay great attention to the amount or proportion of majority class samples, not adequately considering the distance factor. Look at this situation as follow.

Figure 1. A circle in the figure is $p_i$, which is a sample of minority class to be classified. A few of squares and triangles are its 5($m=5$) nearest neighbor samples. Respectively, the squares are minority samples, and the triangles are majority samples. Obviously, the triangles are close to $p_i$, while the squares are far from $p_i$.

Although there are more than $5/2$ ($m/2$) ($m$ is the total number of the selected nearest neighbor samples) minority class samples in the 5 nearest samples in figure 1, the distance between them and the sample $p_i$ is too far compared with the other two majority class samples, so the sample $p_i$ can be easily mistakenly classified into a majority class sample.

Therefore, in order to process the imbalanced data more precisely, a new oversampling algorithm, $\alpha$DBASMOTE ($\alpha$Distance Borderline-ADASYN-SMOTE) algorithm, is proposed in this paper. Both the amount and the distance distribution of the nearest neighbor samples are considered in the algorithm. Furthermore, for the purpose of reducing the misclassification probability of minority class and improving the total classification performance of Alzheimer's disease, the classification model based on DenseNet (Dense Convolutional Network) combined with the $\alpha$DBASMOTE algorithm is proposed in the paper.

2. **Design $\alpha$DBASMOTE algorithm to deal with imbalanced data**

The $\alpha$DBASMOTE algorithm improves the Borderline-SMOTE algorithm and the ADASYN algorithm, and merges the two improved algorithms.

2.1. **Create a few formulas for the algorithm**

2.1.1. create the weighted-distance value formula

On the Base of above view, weighted-distance value, named as $\alpha$, is created according to the distance between a minority class sample point $p_i$ to be classified and one of its nearest neighbor samples. If $p_j$ ($j \in [1,m]$) is one of the $m$ nearest neighbor samples of $p_i$, the distance from $p_j$ to $p_i$ is defined as $\text{dis}_j$. Because those samples with a small distance to $p_i$ should obtain big weight value, the weight value cannot be directly represented by distance. In the paper, the reciprocal of distance is defined as the weighted-distance value to represent the weight of a neighbor sample. For any sample point $p_j$, its weighted-distance formula is created as follow.

$$\alpha_j = \frac{1}{\text{dis}_j} \tag{1}$$

2.1.2. create the weighted-distance sum formula

Here, $\text{pnum}$ and $\text{nnum}$ are used to denote the total number of minority samples and the total number of majority samples in the $m$ nearest neighbors of $p_i$, respectively. Then, the weighted-distance sum of the minority class and the weighted-distance sum of the majority class are respectively as follows.

$$m_p = \alpha_p = \sum_{j=1}^{\text{pnum}} \alpha_j \tag{2}$$
\[ m'_n = \alpha_n = \sum_{j=1}^{\text{num}} \alpha_j \]  

(3)

Compare \( m'_p \) and \( m'_n \). If \( m'_n > m'_p \), it means that the weighted-distance sum of the majority class is greater than the weighted-distance sum of the minority class, which represents that majority class plays a decisive role in \( p_i \) classification and \( p_i \) is classified mistakenly easily. So, minority class sample \( p_i \) should be placed into a danger set. Conversely, if \( m'_p > m'_n \), \( p_i \) is safe and nothing needs to be done.

2.1.3. Create the ratio formula of weighted-distance sum

The \( \alpha \)DBASMOTE algorithm improves the ADASYN algorithm. The number of minority class samples to be generated is according to not only the proportion of majority class samples but also the distance distribution of the nearest neighbor samples of \( p_i \). Therefore, the ratio formula of weighted-distance sum, named as \( r_i \), is created for each danger minority sample \( p_i \).

\[ r'_i = \frac{m'_n}{m'_p}, \ i=1, 2, ..., \text{dnum} \]  

(4)

Here, \( \text{dnum} \) is the number of samples in danger set. The bigger \( r'_i \), the bigger the ratio of the majority weighted-distance sum and the minority weighted-distance sum. So, a big \( r'_i \) represents a big misclassification risk. Therefore, more minority samples should be synthesized in the place where \( r'_i \) is big. In the \( \alpha \)DBASMOTE algorithm, the number of minority class samples to be generated is proportional to \( r'_i \).

2.2. The Flowchart of \( \alpha \)DBASMOTE algorithm

On the base of above analyses and definitions, the flowchart of \( \alpha \)DBASMOTE algorithm is designed as in figure 2.

![figure 2. the \( \alpha \)DBASMOTE algorithm flowchart.](image-url)
2.3. Steps of the αDBASMOTE algorithm

The steps of the αDBASMOTE algorithm proposed in the paper are as follows.

Step1: For each boundary sample \( p_i \) of the minority class, calculate and find its \( m \) nearest neighbor samples.

Step2: Calculate \( m_p' \) and \( m_n' \) according to formula (1), formula (2) and formula (3). Compare the value of \( m_p' \) with \( m_n' \). If \( m_p' > m_n' \), \( p_i \) should be put into the danger set; otherwise, nothing will be done.

Step3: Calculate the total number of minority samples that need to be synthesized. The formula is shown as follow.

\[
G = (n_{num} - p_{num}) \times \beta, \quad \beta \in (0,1]
\]  

When \( \beta \) equals 1, the data is completely balanced in a sense.

Step4: For each minority class sample \( p_i \) in the danger set, calculate the ratio \( r_i' \) according to formula (4).

Step5: Regularize \( r_i' \) for each minority class sample \( p_i \) in the danger set.

Use the ratio of \( r_i' \) and the sum of \( r_i' \) of all dangerous samples to regularize \( r_i' \). So, the sum of \( r_i' \) will be 1. This is the formula.

\[
r_i'^* = \frac{r_i'}{\sum_{i=1}^{d_{num}} r_i'}, \quad \sum_{i=1}^{d_{num}} r_i'^* = 1
\]

Step6: According to formula (5) and (6), the number of minority samples to be synthesized for each minority class sample \( p_i \) in the danger set can be calculated as the following formula.

\[
g_i = r_i'^* \times G
\]

Step7: Synthesize \( g_i \) new minority samples for each minority sample \( p_i \) in the danger set.

For each minority sample \( p_i \) among the \( m \) nearest neighbor samples of \( p_i \), calculate the distance, \( d_{ij} \) \((j=1,2, \ldots, g_i)\), then multiply it by the random number \( r_{ij} \) \((j=1,2, \ldots, g_i)\) which is between 0 and 1. Finally, generate \( g_i \) new minority samples between \( p_i \) and some nearest neighbor samples.

\[
synthetic_j = p_i + r_{ij} \times d_{ij}, \quad j=1, 2, \ldots, g_i
\]

3. Experimental validation

3.1. Data collection

Alzheimer's Disease (AD) is a central nervous system degenerative disease that gradually reduces the patients' memory and cognitive function. It has become most common cause of dementia, accounting for about 60-80 percent of patients with dementia.\(^{[6]}\) It has become the sixth leading cause of death on the world.\(^{[7]}\) The data in this study was obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The MR imaging (MRI) samples of ADNI database were collected. They included AD patients, Mild Cognitive Impairment (MCI) patients and Normal Controls (NC) data. In order to verify the effectiveness of the algorithm proposed in the paper, the collected data was unbalanced, as shown in Table 1.

| Group | AD number | NC number | MCI number | Unbalanced ratio |
|-------|-----------|-----------|------------|-----------------|
| Group1 | 65        | 150       | 0          | 0.43            |
| Group2 | 33        | 0         | 80         | 0.41            |
| Group3 | 0         | 110       | 52         | 0.47            |

3.2. Experiment procedure

At first, the collected data of three groups was balanced by the SMOTE algorithm, the Borderline-
SMOTE algorithm, and the αDBASMOTE algorithm respectively. Next, every balanced group data was divided into three parts, training set, verification set and test set.

Then, the DenseNet network architecture of CNN (Convolution Neural Network) was built in the study. Between input and output, there are three densely connected modules (Dense Blocks), transition Layers were built between any two blocks.

These models got the parameters in training process, using the verification set to monitor training process, optimizing the parameters, and finally got the optimal model for every algorithm and for every classification group. Next, the models had been trained were used to classify corresponding test sets. In order to objectively compare the effects of these algorithms, these experimental data were all the results of 10-fold cross-validation.

Finally, the classification results of these models were shown.

3.3. Evaluation Index
In addition to the Accuracy, the Recall rate and the Precision, evaluation criteria for imbalanced data classification include the F-measure and the AUC.

Table 2. the Confusion Matrix.

| Reality       | Prediction |        |        |
|---------------|------------|--------|--------|
| Positive      | Positive   | TP     | FN     |
| Negative      | FP         | TN     |        |

Accuracy = \[
\frac{TP+TN}{TP+TN+FN+FP}\]

Recall rate = \[
\frac{TP}{TP+FN}\]

In this experiment, the Accuracy, the Recall rate and the AUC were used as the evaluation indexes of the classifier performance. Relevant concepts and formulas of the Accuracy and the Recall rate were shown in table 2, formula (9) and (10). The AUC referred to the area under the ROC curve. The larger the AUC value, the better the classification performance of the classifier.

3.4. Experimental results and discussion
At last, the classification results of models were compared. The comparison tables were given as follows.

Table 3. The results of AD vs. NC classification.

| Model          | Accuracy (%) | Recall rate (%) | AUC  |
|----------------|--------------|-----------------|------|
| SMOTE          | 90.70        | 81.54           | 0.88 |
| Borderline- SMOTE | 92.09     | 84.62           | 0.90 |
| αDBASMOTE      | 94.42        | 89.23           | 0.93 |

Table 4. The results of AD vs. MCI classification.

| Model          | Accuracy (%) | Recall rate (%) | AUC  |
|----------------|--------------|-----------------|------|
| SMOTE          | 77.48        | 64.52           | 0.68 |
| Borderline- SMOTE | 81.98     | 67.74           | 0.78 |
| αDBASMOTE      | 86.49        | 74.19           | 0.82 |
Table 5. The results of MCI vs. NC classification.

|                | Accuracy (%) | Recall rate (%) | AUC   |
|----------------|--------------|-----------------|-------|
| SMOTE          | 74.07        | 69.23           | 0.73  |
| Borderline- SMOTE | 79.63        | 73.08           | 0.79  |
| αDBASMOTE      | 83.95        | 82.69           | 0.84  |

It could be seen from the experimental results that the αDBASMOTE algorithm improved the accuracy performance of the classifier, especially, better results of Recall rate and AUC were obtained, which were sensitive to classification performance of minority class. That was to say, the misclassification probability of the minority class was reduced by using the αDBASMOTE algorithm.

4. Conclusion

A combined model of a deep learning network based on DenseNet and a novel oversampling algorithm, αDBASMOTE algorithm, is proposed in the paper. It optimizes the Borderline-SMOTE method and the ADASYN method, merges them, and makes the data distribution more balanced. It is shown from the results of AD disease data classification that the model proposed in the paper improves the accuracy of classification, especially the accuracy of minority classification.

Hope that this algorithm and classification model can offer a way of thinking and a feasible method for studies that the prediction result of minority samples dominates the total result, such as clinical practice, fraud detection and disaster prediction etc.

As the current popular methods for processing imbalanced data, the algorithms on data level and the algorithms on learning algorithm level have their own advantages. Therefore, a combined algorithm of the αDBASMOTE algorithm and a cost-sensitive algorithm is encouraged in the following study.

Reference

[1] Li, A. J., Zhang, P. (2020) Research on Unbalanced Data Processing Algorithm Base Tomeklinks-Smote. In: 2020 3rd International Conference on Artificial Intelligence and Pattern Recognition. Xiamen, China. p. 13.

[2] Das, R., Kr, S., Biswas, Devi, D., Sarma, B. (2020) An Oversampling Technique by Integrating Reverse-Nearest Neighbor in SMOTE: Reverse-SMOTE. In: Proceedings of the International Conference on Smart Electronics and Communication. Chennai, India. p.123.

[3] Srinilta, C., Kanharattanachai, S. (2021) Application of Natural Neighbor-based Algorithm on Oversampling SMOTE Algorithms. In: 2021 7th International Conference on Engineering, Applied Sciences and Technology. Online. pp. 217-220.

[4] Chen, Y., Chang, R., Guo, J. (2021) Effects of data augmentation method borderline-smote on emotion recognition of eeg signals based on convolutional neural network. IEEE Access, 9: 47495-47496.

[5] Hasmita, S., Nhita, F., Saepudin, D., Aditsania, A. (2019) Chili Commodity Price Forecasting in Bandung Regency using the Adaptive Synthetic Sampling (ADASYN) and K-Nearest Neighbor (KNN) Algorithms. In: 2019 International Conference on Information and Communications Technology. Kuala Lumpur, Malaysia. pp.434-435

[6] Amin-Naji, M., Mahdavinataj, H., Aghagolzadeh, A. (2019) Alzheimer's disease diagnosis from structural MRI using Siamese convolutional neural network. 2019 4th International Conference on Pattern Recognition and Image Analysis. Tehran, Iran. p. 75.

[7] Haque R.U., Levey A. I. (2019) Alzheimer's disease: A clinical perspective and future nonhuman primate research opportunities. Proceedings of the National Academy of Sciences, 116: 26224–26229.