Over-sampling Algorithm Based on in imbalanced Classification

Abstract

The problem of imbalanced classification is that in the classification problem, different classes of samples differ in the sample size, distribution and the classification cost. In the traditional classification algorithms, however, it is assumed that the samples are evenly distributed, resulting in the poor recognition of the samples of interest. In order to improve the accuracy of the minority class samples, scholars have proposed many approaches including data-level and algorithm-level methods. Data-level algorithms are used as an optional part of preprocessing and have a wider range of applications, which has led to a lot of attention. Oversampling, the algorithm uses various methods to increase the minority class samples in the training set, increases the recognition rate of the minority class samples. However, these oversampling methods are too coarse to improve the classification effect of the minority class samples, because they can’t make full use of the information in the original samples, but increase the training time because of adding extra samples. In this paper, we propose to use the distribution information of the minority class samples, use the variational auto-encoder to fit the probability distribution function of them, and reasonably expand the minority class sample set. The experimental results prove the effectiveness of the proposed algorithm.

Keywords: imbalanced classification; generative model; variational auto-encoder;

Introduction

The classification problem is a very important part of machine learning, and it is also the first step for artificial intelligence to understand human life. At present, most classifiers assume that the samples of different classes are evenly distributed, and the classification cost is the same. However, in reality, the data people are more concerned about is often scarce, such as the detection of credit card fraud and medical disease diagnosis. In the diagnosis of medical data, most of the results are normal while only a small proportion of the results are diagnosed as diseases, which indicates the even distribution in different classes samples. Second, if healthy people are misdiagnosed as diseases, they can be removed by other inspection methods, errors do not cause very serious accidents, but if the disease is diagnosed as healthy, it may cause the patient to miss the best treatment time and cause serious consequences. This is the second feature of the imbalanced classification problems: different classes of misclassification costs are inconsistent. At the same time, if samples are classified as diseases as much as possible because they are afraid to miss the disease samples, it will cause a huge waste of medical resources and intensify conflicts between doctors and patients. Therefore, it is not feasible to determine all samples as positive, and the best way is to try to separate these two results as correct as possible. Due to the scarcity of the minority class samples and perusing the global accuracy, the classifier pays less attention to the minority class, so the recognition effect of it is not good. Imbalanced classification problems arise in many fields, such as bioinformatics [1], [2], remote sensing image recognition [3], and privacy protection in cybersecurity [4]. The imbalanced problems cover widely and have a very important practical significance.

The traditional solutions to the imbalanced problems are divided into two parts: the algorithm-level algorithms and the data-level algorithms. The algorithm-level algorithms mainly focus on the different misclassification costs, and the algorithm level is mostly based on cost-sensitive method, that is, misclassification costs are different in different classes, such as improved neural network [5]: it uses the approximation of F1 value of the minority class as the cost function; the bagging algorithm[6] continues to enhance the misclassified the minority class samples, and improve the recognition rate of the minority class samples; structured SVM [7] uses the F1 value of the minority class samples as the optimization function, and thus has a better performance in the classification of the minority class samples.

Introduction

The data-level algorithms focus on the first feature of the imbalanced classification problem, which mainly adjust the distribution of data through data resampling to reduce the impact of imbalanced classification. As for the sampled objects, it can be divided into over-sampling, under-sampling and combinations of the both. Over-sampling refers to the process of increasing the of trainings times of the minority class samples that are easier to be ignored in the training process. The common methods of oversampling includes re-sampling of simple replicated samples, SMOTE[8] for Linear Interpolation of Samples, etc. Oversampling can effectively improve the classifier's attention to the minority class samples, and simply copying samples does not add additional information. However, SMOTE's method of random interpolation is too random and lacks typicality. Under-sampling [9] refers to remove the majority class samples for obfuscation classification, which can quickly reach equilibrium, but may take a risk of losing valuable samples.

The oversampling method can be divided into random sampling and information sampling.

Random sampling means sampling directly using samples, which includes simple repetition [11], linear interpolation [12], nonlinear interpolation [13], etc.; SMOTE [12], as a classic oversampling algorithm, interpolates linearly in the minority class samples will increase the amount of information and rationality of synthesized samples in random oversampling, which improves the classification effect. Border-line-smote [14], on the basis of SMOTE, has filtered the minority class samples that needed to be interpolated, and in order to reduce the risk of overfitting, it mainly interpolates linearly in boundary samples. The above oversampling methods only consider the influence of the sample size on the classification effect and the local sample distribution, instead of considering the overall distribution of the sample.

Sampling with information [15] uses the distribution information in the sample to fit its probability distribution function and sample it according to the modeling content. Chen[16] proposed normal distribution based on oversampling approach and this approach use the probability of minority class to model and oversampling, which the experimental results is better than SMOTE and random oversampling. Different scholars have proposed oversampling algorithms based on various distributions, such as the Gaussian distribution[16][17], Weibull distribution [18],etc. Due to the information in sampling, these algorithms have made greater progress than random interpolation method. However, the problems are also obvious: assume the sample obeys some prior distribution such as normal distribution, we can use EM to calculate the parameters in distribution. If the distribution meets this hypothesis, it will get better results, otherwise, if the distribution disobeys the hypothesis, not only the improvement is limited, but also the true distribution of the sample is hard to get. Because of these above situations, it is inconsistent in their effect on the distribution of different datasets.

Data level processing is of great matter in imbalanced classification. In spite of the consistency in the importance of practical problems, it do will have a positive effect on the final classification results.

Since the factors that affect the datasets classification include not only the sample size, but also the sample distribution, while the current oversampling methods only use direct replication or simple linear interpolation to generate data, which do not make full use of information of distribution and not guarantee the rationality of the interpolation sample. Therefore, in this paper, we propose vae model to expand the minority class samples.

Our method aims at the difficulty in identify the minority class in imbalanced datasets, and according to the distribution relationship between samples, we use vae which generates data well in neural network to increase minority instances, and the experimental results prove the effectiveness of the algorithm.

本文所提出的算法利用vae对少数类样本分布进行建模，神经网络的输出维度不受限制，可以生成任意维度的数据，其强大的拟合能力，可以模拟任意的分布函数而不需要提前设定未知分布的概率分布函数，使用该模型对少数类样本过采样，解决分类问题中的不平衡问题。

We propose this algorithm to use vae to model the distribution of minority class sample. To our knowledge, first, the output dimension of the neural network is not limited so it can generate data of any dimension; second, the strong fitting ability of the neural network can simulate any distribution function without setting the probability distribution function of the unknown distribution in advance.

Therefore, in order to resolve the imbalanced problem, we use this model to oversampling the minority class samples.

We organize the paper as follows. Section 2 describes relevant work of this paper. Section 3 presents our algorithm and analyze it. Section 4 introduce the basic datasets and show our experimental results. Section 5 concludes the paper.

# Related work

The neural network proposed in the paper is used to fit the probability distribution function of minority class data due to the strong ability of fitting. Therefore, it is possible to avoid the inconsistency between hypothesis and original distribution, and to improve the quality of the generated samples and to elevation the effect of classification.

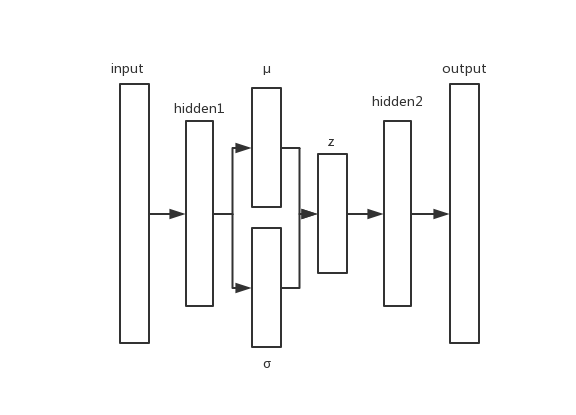
Vae(variational auto-encoder)

In 2003, KM[19] proposed vae: add variational inference to auto-encoder and use parameterization trick to make the variational inference combined with stochastic gradient descent. The overall structure of vae network is in the form of encoder-decoder, while it assume the hidden variables to be normal distribution. And these feature, which is easy to sample and the final probability distribution function is uncertain, coincides with the characteristics of distribution-based oversampling.

在vae中，假设变量是由隐层的压缩编码z决定的，跟在其后的encoder可以将z映射成最终的可见形式，而该压缩编码z服从某种特定的分布（如高斯分布等），在已知该特定分布的情况下，可以根据其CDF进行采样并通过encoder部分，理论上可以产生无限样本。因此，本文中采用vae结构对少数类样本分布进行拟合，并对模型进行采样，以解决少数类的过采样问题。

In vae, we assume the variables is determined by the hidden compression code z, and encoder can map z to visible form. It make z obey a particular distribution (such as Gaussian distribution, etc.),

The structure of vae as shown below:



Assume z is a latent variable, and its distribution function is, we can use Bayesian conditional probability formula to calculate:

However, in z's prior distribution, most of z cannot generate reliable samples, that is tends to 0, so is tends to 0. To simplify the calculation, only need to be calculated. We directly calculate and sample the larger value z of  我们采用直接对较大值的z进行计算和采样,. In vae’s encoder, will be mapped and this part of z can be directly mapped to X by decoder, but only consider this part of z can not generate samples that is not in original data. So we need to assume the distribution of and complete the error through decoder.

由自编码器的编码部分可得, assume generates the distribution , we use KL divergence to calculate the distribution fitting error:

Definition of KL divergence:

The above equation shows that if pq is closer, KL divergence will tends to 0. And the objective function of vae model is

Apply Bayes rule to, we can get both and

Apply the into it

Note that X is fixed, and Q can be any distribution, not just a distribution which does a good job at mapping X to the z’s to produce X. since we’re interested in inferring P(X), it makes sense to construct a Q which does depend on X, and in particular, one which makes small:

Because is fixed, the minimum will transform to maximize value of right side of the equation, and is the probability of X decoded by z. 计算为同原始样本的交叉熵或者均方误差，后者可以理解为的假设先验分布同encoder中映射出的z的分布的误差。

# Method

In this paper, based on the distribution of oversampling for the imbalanced classification problem, a generation model using vae is proposed: in order to improve the classification effect, we use vae to sample the hidden space z and generate the synthetic minority class samples. The network structure is shown in Figure 1. Since the original vae is applied to image generation, the synthesis results can be naturally visualized. However, the sequence data is used directly in this article and it requires some improvement:

1. There might have discrete features in the sequence data, while the features generated by the stochastic gradient descent must be continuously differentiable, so this part of feature must be filtered before vae training. Use 1NN to classify the generated samples and combine the continuous features with the discrete features of the nearest original sample into a new composite sample;
2. Due to the sample size is too small to determine reliably whether the feature is discrete, we assume that it is discrete feature if it appears less than 2. In fact, it is useless in classification if the feature has only one value in each sample.

The mathematical description of the method is as follows:

Given training dataset ， is the sample of d dimension, is the labels represent negative and positive. We use P and N to represents a positive class sample subset and a negative class sample subset, where P contains positive samples, N contains negative samples, and .

During the training of the vae model, feature appears in Jth dimension in training set needs statistics first to exclude discrete features, the formula shown as bellow:

，

If , the feature in column j is discrete, otherwise, the feature is continuous. Divide the features in the dataset into continuous and discrete features in order and take out the continuous features as the final training set.

Train a vae model with and randomly sample it, assume is synthetic sample:

is the final synthetic sample, and is the final training set.

算法架构Algorithm Architecture

|  |
| --- |
| Algorithm 1：VAEOS：VAE-based oversampling approach for imbalanced learning. |
| Input: : dataset with N samples, consisting of positive samples in P and negative samples in N  Output: Classifier H trained with dataset after the oversampling algorithm  Procedure:   1. Divide X into training dataset and testing dataset 2. Data preprocess according to formula () 3. Compute for each feature in , decide each discrete feature and continuous feature 4. Decide according to formula 5. Use to train a model and randomly sample with the corresponding model 6. Synthesize the according to formula 7. Train a classifier H with |

# Experiment

Dataset description

In this paper, all datasets are select from UCI Machine Learning Repository, and some of them are multi-label datasets, so we select one class as the minority class and the remaining samples as majority class.

Table 1 dataset description

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Index | Dataset | 样本总数 | 属性数 | 少数类 | 多数类 | 不平衡率 |
| 1 | breast-w | 699 | 9 | 241 | 485 | 1.90 |
| 2 | vehicle | 846 | 18 | 199 | 647 | 3.25 |
| 3 | segment-challenge | 1500 | 19 | 205 | 1295 | 6.32 |
| 4 | Diabetes | 768 | 8 | 268 | 500 | 1.87 |
| 5 | Ionosphere | 351 | 34 | 126 | 225 | 1.79 |
| 6 | Sonar | 208 | 60 | 97 | 111 | 1.14 |

Data preprocess

There are missing values in the datasets, in order to ensure the integrity of the datasets, we use the most frequent value as a supplement. For datasets whose attribute values are not within a certain range, we use normalization to scale it, the formula is shown bellow:

Then

Evaluation indicators base on confusion matrix

在传统的分类算法中由于目标函数倾向于最小化整体的错误率，但在不平衡样本分类中，少数类样本的错误代价更大，

In traditional classification method, overall accuracy is used as the evaluation indicator, while in imbalanced problem, using this indicator will make the classifier insensitive to the minority class for the number of positive class is less. In extreme conditions, assume the dataset only contain 1% minority class, if the classifier decides all samples as majority class, the accuracy still can reach 99%, however, it is但是这对我们关心的少数类是非常不利的.

Therefore, the traditional classification algorithm will make the minority class easily divided into majority class, which resulting in low recognition rate.

In binary classification, the confusion matrix is often used to evaluate the performance of the classifier, which is defined as follows:

表2 二分类问题的混淆矩阵

|  |  |  |
| --- | --- | --- |
|  | Positive prediction | Negative prediction |
| Positive class | True positive(TP) | False negative(FN) |
| Negative class | False positive(FP) | True negative(TN) |

Among them, TP is the number of correct predictions that an instance is positive. TN is the number of incorrect predictions that an instance is positive, FN is the number of incorrect of predictions that an instance negative, and FP is the number of correct predictions that an instance is negative. There are some new evaluation metrics based on confusion matrix to calculate the accuracy and recall of imbalance data such as F-value, G-mean and AUC[20].

In paper [21], it explained AUC is not reliable in the extreme conditions, so it only use F-value and gmean to analyze the classification in experiment.

F-value id the evaluation matric of accuracy and recall, and it biased to the performance of minority class, the definition is as follows:

Then and , the value of is [0,+∞].

And in this experiment, we choose for F-value is the average weight between recall and accuracy. 召回率和准确率之间权重的平均

Gmean is the geometric mean of the classification accuracy of minority class and majority class, then:

Only when the precision of minority class and precision of majority are high at the same time, gmean will be maximum.

**Algorithm and parameters cooperation**

In this paper, we compare different oversampling algorithm such as NDO-sampling[7] and random interpolation algorithm SMOTE. The classifier is unified with naïve Bayes to reduce the impact on classification. In order to ensure the data distribution is more consistent with the original one, minority class and majority class should be fragmented simultaneous in cross validation. To reduce the impact of the randomness on the results of comparison, each algorithm calculates the average of 10 times with 10-fold cross validation.

|  |  |  |  |
| --- | --- | --- | --- |
| 算法 | 本文方法 | NDO | SMOTE |
| 参数 | Vae：5层  Hidden1:[]  Hidden2:[]  Hidden3:[]  Encoder和decoder为对称结构 | Same with paper [7] | k=5 |

Experiment result

不同算法的对比

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 100% | | | 200% | | | 300% | | |
|  | VAE | NDO | SMOTE | VAE | NDO | SMOTE | VAE | NDO | SMOTE |
| 1 | 94.22 | **94.38** | **94.38** | **94.99** | 94.38 | 94.38 | **94.99** | 94.38 | 94.38 |
| 2 | 58.07 | 55.66 | 56.26 | 58.75 | 56.63 | 56.45 | 59.36 | 56.27 | 56.45 |
| 3 | 66.49 | 65.47 | 62.44 | 69.56 | 66.55 | 61.15 | 70.86 | 61.90 | 61.15 |
| 4 | 65.88 | 65.93 | 66.27 | 65.50 | 66.74 | 66.33 | 64.98 | 65.59 | 66.33 |
| 5 | 87.02 | 82.34 | 80.54 | 87.62 | 82.63 | 82.71 | 86.20 | 81.44 | 82.71 |
| 6 | 58.34 | 71.85 | 70.40 | 43.66 | 73.90 | 70.41 | 32.21 | 73.89 | 70.43 |

不同算法的对比

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 100% | | | 200% | | | 300% | | |
|  | VAE | NDO | SMOTE | VAE | NDO | SMOTE | VAE | NDO | SMOTE |
| 1 | 96.77 | **96.89** | **96.89** | 97.22 | 96.89 | 96.89 | 97.21 | 96.89 | 96.89 |
| 2 | 74.40 | 72.06 | 72.35 | 76.43 | 71.98 | 71.87 | 78.22 | 71.73 | 71.90 |
| 3 | 91.54 | 91.79 | 89.69 | 92.81 | 92.22 | 89.41 | 93.30 | 92.47 | 89.03 |
| 4 | 79.85 | 79.73 | 80.25 | 78.19 | 77.78 | 76.71 | 77.86 | 74.95 | 74.48 |
| 5 | 93.53 | 89.62 | 87.79 | 93.98 | 89.84 | 88.56 | 93.49 | 90.07 | 89.45 |
| 6 | 61.33 | 63.77 | 64.67 | 62.05 | 65.13 | 65.03 | 66.13 | 65.02 | 65.39 |

不同算法的对比

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 100% | | | 200% | | | 300% | | |
|  | VAE | NDO | SMOTE | VAE | NDO | SMOTE | VAE | NDO | SMOTE |
| 1 | 96.04 | 96.35 | 96.35 | 96.47 | 96.35 | 96.35 | 96.45 | 96.35 | 96.35 |
| 2 | 75.00 | 72.71 | 73.27 | 75.79 | 73.50 | 73.18 | 76.31 | 73.30 | 73.34 |
| 3 | 90.60 | 89.55 | 88.91 | 91.47 | 89.23 | 89.36 | 91.78 | 89.41 | 89.01 |
| 4 | 73.43 | 73.57 | 73.84 | 72.92 | 74.07 | 73.01 | 72.48 | 73.24 | 73.03 |
| 5 | 88.83 | 86.47 | 85.32 | 89.08 | 86.66 | 85.99 | 87.81 | 86.86 | 86.98 |
| 6 | 58.48 | 67.31 | 67.37 | 49.97 | 68.81 | 67.59 | 40.20 | 68.73 | 67.81 |

过采样率同结果的关系

The result shown in table indicate that vae performs better in generating samples than NDO and SMOTE when the number of oversampling is the same. In the meanwhile, compared with the traditional oversampling algorithms which sacrifice some majority class to ensure the classification effect of minority, vae can guarantee the rational distribution of synthetic samples and improve the overall classification effect。

？？？

从实验数据中可以看到，在过采样数量相同的情况下，vae产生的样本比NDO和SMOTE生成的样本更能提高分类器的分类效果，同时，比起传统的过采样算法中，为了保证少数类的分类效果而牺牲掉部分多数类，本文中所提出利用vae进行过采样，保证了合成样本的分布，并且能够提高整体的分类效果。

从不同的过采样算法的对比结果来看，本文中所提出的过采样算法能够产生更加合理的样本，和以往的过采样算法会牺牲多数类的识别率不同，本文所提出的算法能够帮助分类器产生更合理的分类结果，提高整体的识别效果。

# Conclusion

In this paper, we use a generation model to replace the traditional oversampling mechanism in order to make full use of the information in the dataset. The experiment results prove the effectiveness of this idea, but in the procedure of sampling, it is still the rough random oversampling. And our next expansion work is to add a screening mechanism for generating samples

在这篇论文中，我们利用生成模型来代替传统的过采样机制，以便充分利用数据集中的信息， 实验结果证明了该种思路的有效性，但是在采样中，仍然是粗糙的随机过采样的过程而缺少对生成样本的筛选机制，这是接下来的拓展工作。

# Reference

[1] Y. Wang, X. Li, and B. Tao, “Improving classification of mature microRNA by solving class imbalance problem,” *Scientific Reports*, vol. 6, p. 25941, May 2016.

[2] G. Stegmayer, C. Yones, L. Kamenetzky, and D. H. Milone, “High class-imbalance in pre-miRNA prediction: a novel approach based on deepSOM,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, pp. 1–1, 2016.

[3] T. Leichtle, C. Geiß, T. Lakes, and H. Taubenböck, “Class imbalance in unsupervised change detection – A diagnostic analysis from urban remote sensing,” *International Journal of Applied Earth Observation & Geoinformation*, vol. 60, pp. 83–98, 2017.

[4] C. Li and S. Liu, “A comparative study of the class imbalance problem in Twitter spam detection,” *Concurrency & Computation Practice & Experience*, vol. 30, no. 4, 2018.

[5] C. Zhang, G. Wang, Y. Zhou, and J. Jiang, “A New Approach for Imbalanced Data Classification Based on Minimize Loss Learning,” in *IEEE Second International Conference on Data Science in Cyberspace*, 2017, pp. 82–87.

[6] F. Provost, “Machine Learning from Imbalanced Data Sets 101 (Extended Abstract),” in *Soft Computing and Pattern Recognition (SoCPaR), 2011 International Conference of*, 2008, pp. 435–439.

[7] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun, “Support vector machine learning for interdependent and structured output spaces,” in *International Conference on Machine Learning*, 2004, p. 104.

[8] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2011.

[9] D. L. Donoho and J. Tanner, “Precise Undersampling Theorems,” *Proceedings of the IEEE*, vol. 98, no. 6, pp. 913–924, 2010.

[10] I. Bilinskis, *Hybrid Sampling*. John Wiley & Sons, Ltd, 2007.

[11] F. Olken and D. Rotem, “Random sampling from databases: a survey,” *Statistics & Computing*, vol. 5, no. 1, pp. 25–42, 1995.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: synthetic minority over-sampling technique,” *Journal of Artificial Intelligence Research*, vol. 16, no. 1, pp. 321–357, 2002.

[13] C. Zhang, J. Guo, and J. Lu, “Research on Classification Method of High-Dimensional Class-Imbalanced Data Sets Based on SVM,” in *IEEE Second International Conference on Data Science in Cyberspace*, 2017, pp. 60–67.

[14] H. Han, W. Y. Wang, and B. H. Mao, “Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning,” *Lecture Notes in Computer Science*, vol. 3644, no. 5, pp. 878–887, 2005.

[15] M. Gao, X. Hong, S. Chen, and C. J. Harris, “Probability density function estimation based over-sampling for imbalanced two-class problems,” in *International Joint Conference on Neural Networks*, 2012, pp. 1–8.

[16] S. Chen, “A generalized Gaussian distribution based uncertainty sampling approach and its application in actual evapotranspiration assimilation,” *Journal of Hydrology*, vol. 552, 2017.

[17] H. Zhang and Z. Wang, “A normal distribution-based over-sampling approach to imbalanced data classification,” in *International Conference on Advanced Data Mining and Applications*, 2011, pp. 83–96.

[18] D. C. Li, S. C. Hu, L. S. Lin, and C. W. Yeh, “Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets,” *Plos One*, vol. 12, no. 8, p. e0181853, 2017.

[19] K. Diederik P and W. Max, “Auto-Encoding Variational Bayes.”

[20] G. Menardi and N. Torelli, “Training and assessing classification rules with imbalanced data,” *Data Mining & Knowledge Discovery*, vol. 28, no. 1, pp. 92–122, 2014.

[21] Q. Zou, S. Xie, Z. Lin, M. Wu, and Y. Ju, “Finding the Best Classification Threshold in Imbalanced Classification ☆,” *Big Data Research*, vol. 5, pp. 2–8, 2016.