An improved measurement for the imbalanced dataset

First Author1[0000-1111-2222-3333] and Second Author2[1111-2222-3333-4444]

1 Princeton University, Princeton NJ 08544, USA

2 Springer Heidelberg, Tiergartenstr. 17, 69121 Heidelberg, Germany  
lncs@springer.com

**Abstract.** Imbalanced classification is a classification problem that violates the assumption of uniform distribution of samples. In such problems, traditional imbalanced datasets are measured in terms of the imbalance of the number of samples, but in various studies, the classification performance is not only related to the number of samples, but also the sample distribution, and the sample distribution often plays a more important role in this. In this paper, we proposed an improved measurement for imbalanced datasets, which is based on the idea that a sample surrounded more nearest neighbors with the same label is easier to classify, and it is a good indicator of the relationship between the distribution of samples and the classification results. The experimental results show that the proposed measurement has a higher correlation with the classification results and shows the difficulty of classification of data sets more clearly.

**Keywords:** Imbalanced Classification, Measurement, Imbalance Ratio.

1. Introduction

The classification problem is a very important part of machine learning. In the traditional classification problem, the model training is based on the assumption that the sample distribution is uniform, so the classification cost of each sample is consistent. However, in realistic data sets, the assumption of uniform distribution of samples is difficult to satisfy, in case of pursuing the global accuracy, the traditional classifier can easily overlook the identification of the minority samples, causing them to be hard to recognize. The imbalanced classification problem has appeared in many fields, such as bioinformatics[1,2], remote sensing image recognition[3], and privacy protection in cybersecurity [4]. The wide coverage of the imbalance problem has very important practical significance.

The number of samples has had a noticeable effect on the classification results. Therefore, the imbalance ratio (IR) of the number of samples in different classes has been popular as a measurement of imbalanced datasets, and the scholars have proposed many sampling algorithms to balance the datasets to improve the classification performance in the minority class However, the IR is not informative enough to measure a specific dataset overall, many studies have shown that when the number of samples is relatively large, it does not cause a reduction in the classification performance of the minority class, but when the number of samples is seriously insufficient, the rarity of the minority samples will cause a low recognition rate of the minority samples.

This paper considers the measurement of the dataset from the inconsistency of the sample classification difficulty, improves the traditional calculation method of the IR, and improves the correlation between the measurement and the final classification performance. This paper is arranged as follows: section 2 describes the related work in measurement of the imbalanced dataset, section 3 shows the proposed measurement IGIR, and section 4 describes the experimental results and analysis, the final section concludes the proposed method and the future work.

1. Related work

Measurement of the imbalanced datasets can be divided into two types: local measurements and global measurements. The local measurements refer to these methods which need traversing each sample in a data set, calculating a measurement usually accompanied by the k-NN algorithm for each sample, the overall measurement is defined by the mean value of measurement of all the samples in the dataset. Because this kind of measurement contains the calculation for each sample, and it can be used in the sampling algorithm to find a simpler dataset to model with enough information with the original dataset. Global measurement refers to the methods that are a result calculated for a sample in the entire data set, or a variety of indicators derived from statistical analysis. It is usually accompanied by a variety of calculations for the separate results of the positive and negative subsets. Such measurements are difficult to achieve in a single implemented on the sample, it can only be used as a measure of the dataset, and it is difficult to play a role in the sampling algorithm because the movement of a single sample can hardly affect the original measurement result.

These measurements are used in such two ways: indicate whether a dataset is easy to classify, and measure the sampled subset in sampling methods. Therefore, in order to achieve a better performance, the measurement should have a relatively high correlation with the classification results.

Given dataset X, which contains N+ positive samples (the minority class), N- negative samples (the majority class), and the total number of samples is N=N\_-+N\_+.

* 1. IR

Imbalance ratio the definition is as follows, it is defined as the size sample ratio:

()

When the different classes samples have the same distribution, the sample size is able to reflect whether the samples are easy to classify, but when the data has different possibility distribution function, for example, in the fig.1, the IR of data in (a) is 4 and in b is 1, but the two classes in (a) have a clear linear boundary while there are not in (b), so we can get 100% accuracy in (a) but cannot in (b) with a same scale model, which is contrary to the comparison result of IR, since IR is the proportion of sample size and does not contain any sample distribution information, complexity of the distribution of data cannot be represented in IR.

|  |  |
| --- | --- |
| F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\figa.png | F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\figb.png |
| (a) | (b) |

**Fig. 1.** The dilemma of IR.

* 1. F1

A classical measure of the discriminative power of the covariates, or features, is Fisher’s discriminant ratio[5]:

()

Where μ1, μ2, σ12, σ22 are the means and variances of the two classes, respectively. For multidimensional problems, it is not necessarily the case that all features contribute to class discrimination, so the maximum f over all features can be used. However, a zero maximum f does not necessarily mean that the classes are not separable, as it could just be that the separating boundary is not parallel to an axis in any of the given features.

* 1. CM

CM [6]focuses on the local information for each data point via the nearest neighbors, and uses this information to capture data complexity.

()

Where I(.) is the indicator function. The overall measure is

()

CM is determined by the type of its neighbors. If the neighbors of a sample contain more samples of the same type, the classification of the sample will be less difficult. On the contrary, if the samples are surrounded by samples with different labels, then the sample is difficult to classify correctly, and the average number of different classes samples contained in the k nearest neighbors is used as the measurement. The higher the CM, the more difficult the dataset is to learn.

* 1. GIR

The GIR[7] is an improvement of cm, it focuses on the differences in the difficulty of classifying the samples in different classes. A dataset with a larger GIR is more difficult to get a good performance of the minority class, as the classifier tends to be trained with the easier samples according to the Occam shaver principle. Because we tend to use a as simple as possible classifier to fit the whole dataset, while the more difficult samples need a more complex classifier, which may cause overfitting with single classifier, so this is the reason why ensemble can be effective in the imbalance classification, as they have the different classifiers corresponding to different levels of sample classification difficulty.

()

()

()

Where IR(x, X) is an indicator function. For a sample x, if its k-nearest neighbor’s label is the same as x, the result is 1, otherwise, it gets 0.

GIR considers the different measurements of the positive and negative subsets, which is an improvement of CM, GIR is defined as the difference between positive and negative sample subsets, and paper[7] successfully applies GIR to oversampling and under-sampling algorithms. The experimental results show that GIR-based resampling algorithm can effectively improve the classification performance.

1. The Proposed Method

In the GIR, the measurement of each sample is the average number of samples of the same label in its k-nearest neighbor. First, in the classification process, beside the label of k nearest neighbors, their distance from the sample will also affect the classification result. Second, the GIR of the data set is calculated by the measurement of negative class minus positive class, so GIR is a relative measurement. As shown in table 1, the final result shows that the two data sets have the same GIR, but it is clear that the dataset (b) is more difficult to classify than (a). Therefore, GIR is not so sufficient to fully interpret the complexity of the dataset distribution.

**Table 1.** The dilemma of GIR

|  |  |  |
| --- | --- | --- |
|  | (a) | (b) |
| T\_ | 0.9 | 0.5 |
| T+ | 0.7 | 0.3 |
| GIR | 0.2 | 0.2 |

The sample distribution plays an important role in the classification result. Therefore, the sample distribution is considered in the new measurement, based on this idea, an improved GIR, called IGIR, is proposed in this paper. The motivation of IGIR is that if there are many samples with the same label around the sample, the sample is easily classified, and on the contrary, the sample is hard to classify. Different distances of the k nearest neighbors have different effects on the classification results of the sample.

()

()

()

()

In the calculation of IGIR, the k-nearest neighbors of each sample in the dataset are calculated at first, and their neighboring class labels are retained. First, according to the calculation method in Formula (8), the weights of the k nearest neighbors are gradually reduced to 0. The main reason for not using distance is that the distances between different samples and its neighbors are inconsistent. This will result in the inconsistency of the weights of each sample in the calculation process, therefore, it is not possible to have a comparative standard for the overall results; second, to describe the dataset reasonably with an absolute measurement, and to avoid the relativity in the original GIR, IGIR is defined as the compound measurements of positive and negative subsets. In this case, to ensure that the order of magnitude is unchanged, it is processed by prescribing to better measure the difficulty of classification of the dataset.

|  |
| --- |
| Algorithm 1：IGIR：A new measurement of the imbalanced dataset. |
| Input:: dataset with N samples, consisting of positive samples in P and negative samples in N, Y: the corresponding labels.  Output: the measurement IGIR of X U Y.  Procedure:   1. For each x in X with label 2. Compute the k nearest neighbors of x 3. Count the number of samples with the same label and save the corresponding weight 4. calculate the I(x) for sample x 5. Calculate the average , according to the yighted-T\_, weighted\_d\_tares, and the SSE represents the error sum of squares.ult samples need a more complex classif formula(5) and (6) 6. Calculate the overall IGIR according to the formula(7) |

IGIR can be regarded as the average classification accuracy under a weighted k-NN. That is, the more neighbors of the same class in the sample, the more likely the sample is to be classified as the original classifier, then IGIR has the nature to be related to the final classification performance.

1. Experimental results
   1. Datasets

The experimental data in this paper comes from the UCI machine learning database [8]. Some of them are multi-class datasets, in order to obtain a harder dataset to classify, we select one of the class as the minority class, and the rest of classes are regarded as the majority.

**Table 2.** Datasets

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Datasets | Samples | Attributes | Target | Minority |
| breasttissue | 106 | 9 | Carcinoma | 21 |
| breastw | 699 | 9 | malignant | 241 |
| diabetes | 768 | 8 | 0 | 268 |
| german | 1000 | 24 | 2 | 300 |
| glass | 214 | 9 | 1 2 3 | 51 |
| haberman | 306 | 3 | 1 | 81 |
| ionosphere | 351 | 34 | G | 126 |
| movement | 360 | 90 | 1 | 24 |
| satimage | 6435 | 36 | 4 | 703 |
| segment-challenge | 1500 | 19 | brick face | 205 |
| sonar | 208 | 60 | R | 97 |
| spect | 267 | 22 | 1 | 55 |
| vehicle | 846 | 18 | van | 199 |
| vertebral | 310 | 6 | AB | 100 |
| wpbc | 198 | 33 | 1 | 47 |
| yeast0 | 1484 | 8 | 0 | 244 |
| yeast1 | 1484 | 8 | 1 | 429 |
| yeast2 | 1484 | 8 | 2 | 463 |
| yeast6 | 1484 | 8 | 6 | 163 |

* 1. Evaluation

In the binary imbalanced classification, the confusion matrix is often used to evaluate the performance of the classifier, which is defined in table 3:

**Table 3.** Confusion metrics

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

FN represents the number of positive samples that are incorrectly classified as negative, and FP is the number of samples that are incorrectly classified as positive, there have been compound evaluations, such as F-value and Gmean [9].

()

()

()

()

* 1. Experimental settings and results

Set β=1 in F-value called F1\_min, all involved k-NN are set with k=5, the classifier is C4.5, all results are the average of 10 times of 10-fold cross-validation.

**Table 4.** Measurements and classification results

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | IR | GIR | CM | F1 | IGIR | F1\_min | Gmean | Sensitivity |
| breasttissue | 4.05 | 0.30 | 0.24 | 3.33 | 0.46 | 0.78 | 0.83 | 0.83 |
| breastw | 1.90 | 0.05 | 0.05 | 3.47 | 0.57 | 0.91 | 0.92 | 0.88 |
| diabetes | 1.87 | 0.22 | 0.49 | 0.58 | 0.37 | 0.57 | 0.66 | 0.57 |
| german | 2.33 | 0.37 | 0.52 | 0.35 | 0.33 | 0.47 | 0.60 | 0.48 |
| glass | 3.20 | 0.19 | 0.12 | 3.31 | 0.53 | 0.75 | 0.80 | 0.75 |
| haberman | 2.78 | 0.43 | 0.46 | 0.18 | 0.32 | 0.24 | 0.35 | 0.30 |
| ionosphere | 1.79 | 0.39 | 0.21 | 0.61 | 0.46 | 0.84 | 0.87 | 0.82 |
| movement | 14.00 | 0.44 | 0.06 | 0.98 | 0.47 | 0.59 | 0.77 | 0.69 |
| satimage | 8.15 | 0.04 | 0.01 | 5.01 | 0.59 | 0.95 | 0.97 | 0.94 |
| Segment\* | 6.32 | 0.04 | 0.03 | 1.81 | 0.58 | 0.97 | 0.98 | 0.97 |
| sonar | 1.14 | 0.16 | 0.34 | 0.46 | 0.47 | 0.59 | 0.58 | 0.62 |
| spect | 3.85 | 0.61 | 0.33 | 0.60 | 0.32 | 0.50 | 0.65 | 0.51 |
| vehicle | 3.25 | 0.10 | 0.12 | 1.12 | 0.54 | 0.88 | 0.92 | 0.89 |
| vertebral | 2.10 | 0.14 | 0.31 | 0.75 | 0.47 | 0.67 | 0.71 | 0.66 |
| wpbc | 3.21 | 0.39 | 0.43 | 0.47 | 0.32 | 0.42 | 0.56 | 0.46 |
| yeast0 | 5.08 | 0.38 | 0.21 | 0.74 | 0.41 | 0.49 | 0.68 | 0.53 |
| yeast1 | 2.46 | 0.34 | 0.42 | 0.24 | 0.37 | 0.48 | 0.62 | 0.50 |
| yeast2 | 2.21 | 0.26 | 0.48 | 0.21 | 0.37 | 0.49 | 0.61 | 0.49 |
| yeast6 | 8.10 | 0.33 | 0.08 | 2.75 | 0.47 | 0.69 | 0.82 | 0.71 |

|  |  |
| --- | --- |
| F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\imbalance-ratio.png | F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\f1.png |
| (a) | (b) |
| F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\cm.png | F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\gir.png |
| (c) | (d) |
| C:\Users\zhouying\AppData\Local\Microsoft\Windows\INetCache\Content.Word\wei-igir.png | |
| (e) | |

**Fig. 2.** Measurements and sensitivity

Taking the sensitivity of minority class as an example, the Fig.2 shows the relationship between different measurements and classification results. It can be seen that the correlation between CM and IGIR have a stronger linear relation with sensitivity as the measurement while there is no obvious trend in the rest measurements. In addition, the points in CM are more dispersed and the ones in IGIR are more concentrated, which means datasets with the same IGIR are more likely to have the same degree of classification difficulty than those with the same CM.

* 1. Analysis

In order to quantitatively analyze the relationship between different measurements and the classification results, the results are further analyzed by the determination coefficient R2. R2 reflects how many percentages of the fluctuation of Y can be described by the fluctuation of X. That is to say, what percentage of the variance of the representation variable Y can be explained by the controlled variable X.

()

Where SST=SSR+SSE, SST represents the total sum of squares, SSR represents the regression sum of squares, and the SSE represents the error sum of squares.

**Table 5.** R2 of measurements and classification results.

|  |  |  |  |
| --- | --- | --- | --- |
|  | F1\_min | Gmean | sensitivity |
| IGIR | 0.92 | 0.88 | 0.93 |
| IR | 0.18 | 0.34 | 0.28 |
| GIR | -0.70 | -0.58 | -0.67 |
| CM | -0.80 | -0.85 | -0.84 |
| F1 | 0.70 | 0.70 | 0.71 |

In IGIR, we calculate the number of samples of the average k-nearest neighbors by each sample, so the calculated value can be considered as the probability that the sample is classified as its own class. To a certain extent, this measurement can be regarded as Gmean under the k-NN classifier, and it is reasonable to indicate the classification performance of other classifiers.

The R2 in table 5 also shows the superiority of IGIR. The IGIR proposed in this paper is more capable to indicate the classification results, and it has a stronger relevance with the final classification performance and can be a better indicator of the sampled subset in resampling methods.

1. Conclusion

In this paper, an improved measurement for imbalanced datasets is proposed. This measurement has a higher correlation with the classification results and can be used in the sampling algorithm. The future work will be sampling algorithms based on this measurement to improve the classification results.

Acknowledge

This work is supported by the Foundation Item: Shenzhen Foundation Research Project (No.JCYJ20170307151518535).

References

1. Wang Y, Li X, Tao B: Improving classification of mature microRNA by solving class imbalance problem. Scientific Reports 6, 25941 (2016).

2. Stegmayer G, Yones C, Kamenetzky L, Milone DH: High Class-Imbalance in pre-miRNA Prediction: A Novel Approach Based on deepSOM. IEEE/ACM Transactions on Computational Biology and Bioinformatics 14(6), 1316–26 (2017).

3. Leichtle T, Geiß C, Lakes T, Taubenböck H: Class imbalance in unsupervised change detection – A diagnostic analysis from urban remote sensing. International Journal of Applied Earth Observation and Geoinformation 60, 83–98 (2017).

4. Li C, Liu S, Shigang Liu C: A comparative study of the class imbalance problem in Twitter spam detection. Concurrency & Computation Practice & Experience 30 (4), (2018).

5. Ho T: A Data Comxity Analysis of Comparative Advantages of Decision Forest Constructors. Pattern Analysis & Applications 5(2), 102–12 (2002).

6. Anwar N, Jones G, Ganesh S: Measurement of data complexity for classification problems with unbalanced data. Statistical Analysis and Data Mining 7(3), 194–211 (2014).

7. Tang B, He H: GIR-based ensemble sampling approaches for imbalanced learning. Pattern Recognition 71, 306–19 (2017).

8. Amini MR, Usunier N, Goutte C, http://archive.ics.uci.edu/ml/datasets. html, last accessed 2018/03/22.