An improved measurement for the imbalanced dataset

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 results are not only related to the number of samples, the sample distribution plays a more important role in this, this paper proposed an improved measurement for imbalanced datasets 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.

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 order to pursue 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 for many years as a measurement of imbalanced datasets. IR has been used as a measurement of datasets for a long time, and based on it, scholars have proposed many sampling algorithms to balance the datasets to release effect of the imbalance in sample size on the classification performance, so the measurement plays a very important role in imbalanced classification. However, the IR is not enough to measure a specific dataset overall, 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.

Measurement of the imbalanced datasets can be divided into two types: local measurements and global measurements. The local measurements refer to these methods that 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 thus 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 data set, 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.

There are two roles of unbalanced measurements: indicates whether a dataset is easy to classify, and the usage ​​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=.

1. IR

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

|  |  |  |
| --- | --- | --- |
|  |  | (1) |

when the different classes samples are 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 delimma of IR

1. F1[5]

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

|  |  |  |
| --- | --- | --- |
|  |  | (2) |

where 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[6]

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

|  |  |  |
| --- | --- | --- |
|  |  | (3) |

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

|  |  |  |
| --- | --- | --- |
|  |  | (4) |

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 a different label, then the classifier is difficult to classify correctly, and the mean value of the number of different types of samples contained in the neighbors of the sample is used as the classification difficulty of the sample. The higher the CM, the more difficult the dataset is to learn.

1. GIR[7]

The GIR is an improvement of cm, it focuses on the difference 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | (5) |
|  | | |  | (6) | | |
| (2-1) | |  | | | (7) | |

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, if different, it gets 0.

GIR considers that the different measurements of the positive and negative subsets, it 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.

The proposed method

In the GIR, the measurement of each sample is the average number of samples with the same label in its k-nearest neighbor. Firstly, in the classification process, the distance of k-nearest neighbors will also affect the classification result. Secondly, the GIR of the data set is calculated by the measurement of negative class minas that of 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) |
|  | 0.9 | 0.5 |
|  | 0.7 | 0.3 |
|  | 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.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  |  | | | | | (8) |
|  |  | | | | | (9) |
|  | | |  | (10) | | |
|  | |  | | | (11) | |

In the calculation of IGIR, the k-nearest neighbors of each sample in the dataset are first calculated, and their neighboring class labels are retained. Firstly, 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; secondly, to describe the dataset reasonably with an absolute measurement, avoiding the relativity in the original GIR, IGIR is defined as the compound measurements of positive and negative subsets. In this case, in order 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 of the classification accuracy of the positive and negative samples of the sample 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.

Experimental results

1. Dataset

The experimental data in this paper comes from the UCI machine learning database[8]. Some of them are multi-class data sets, 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 G-mean[9].

|  |  |  |
| --- | --- | --- |
|  |  | (12) |
|  |  | (13) |
|  |  | (14) |
|  |  | (15) |

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 the 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\gir.png |
| (a) | (b) |
| F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\f1.png | F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\cm.png |
| (c) | (d) |
| F:\OneDrive\mytensorflow\paper_experiment\imbalance_ir\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. Experimental analysis

In order to quantitatively analyze the relationship between different measurements and the classification results, the results are further analyzed using 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.

|  |  |  |
| --- | --- | --- |
|  |  | (16) |

Where, SST represents the total sum of squares, SSR represents the total sum of square, 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 of 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 show the superiority of IGIR. The IGIR proposed in this paper is more able to indicate the classification results, and it has a better effect on the difficulty of classification of data sets and the indication of resampling.

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.

Acknowledgement

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

Reference

[1] Y. Wang, X. Li, and B. Tao, “Improving classification of mature microRNA by solving class imbalance problem,” *Sci. Rep.*, 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 Trans. Comput. Biol. Bioinform.*, 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,” *Int. J. Appl. Earth Obs. 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,” *Concurr. Comput. Pract. Exp.*, vol. 30, no. 4, 2018.

[5] T. K. Ho, “A Data Complexity Analysis of Comparative Advantages of Decision Forest Constructors,” *Pattern Anal. Appl.*, vol. 5, no. 2, pp. 102–112, 2002.

[6] N. Anwar, G. Jones, and S. Ganesh, “Measurement of data complexity for classification problems with unbalanced data,” *Stat. Anal. Data Min. Asa Data Sci. J.*, vol. 7, no. 3, pp. 194–211, 2014.

[7] B. Tang and H. He, “GIR-based Ensemble Sampling Approaches for Imbalanced Learning,” *Pattern Recognit.*, vol. 71, 2017.

[8] M. R. Amini, N. Usunier, and C. Goutte, “uci-dataset-url,” 2013.

[9] G. Menardi and N. Torelli, “Training and assessing classification rules with imbalanced data,” *Data Min. Knowl. Discov.*, vol. 28, no. 1, pp. 92–122, 2014.