An improved measurement of the imbalanced dataset

Chunkai Zhang[1], Ying Zhou[2], Yingyang Chen[3], Lifeng Dong[4], and Changqing Qi[5]

[1235] Harbin Institute of Technology Shenzhen Graduate School, Shenzhen, China
[4] Hamline University, 1536 Hewitt Avenue, St. Paul, the USA
ckzhang812@gmail.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 sample size, without considering the distribution information, which has a more important impact on the classification performance, so the traditional measurements have a weak relation with the classification performance. This paper proposed an improved measurement for imbalanced datasets, it is based on the idea that a sample surrounded by more same class samples is easier to classify, for each sample of different classes, the proposed method calculates the average number of the k nearest neighbors in the same class in different subsets under the weighted k-NN, after that, the product of these average values is regarded as the measurement of this dataset, 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 order to pursue the global accuracy, the traditional classifier can easily get an unsatisfying classification performance 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]–[6]. The wide coverage of the imbalance problem has very important practical significance.

The traditional imbalanced classification problem has two features: the difference in sample size; the difference in misclassification cost for different classes. Scholars have proposed the data-level methods for feature 1 and the algorithm-level methods for feature 2. As the data-level methods can be regarded as a preprocessing before training the classifier, so they have been popular for many years. The data-level methods can be

divided into the oversampling to add the minority samples, the under-sampling to reduce the majority samples and the hybrid sampling methods, they aim to get a balanced dataset, which is defined by the imbalanced dataset measurement.

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 [7] [8], 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 [9] refers to 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.

As the number of samples has had a noticeable effect on the classification results. Therefore, the imbalance ratio [8-10] (IR) of the number of samples in different classes has been popular for many years as a measurement of imbalanced datasets. Based on IR, 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 informative enough to measure a specific dataset overall, as it is a global measurement, studies [10] 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. The local measurements develop the global ones as they take the distribution into consideration, meanwhile, with the understanding of the classifier and data, the distribution based sampling methods [12-14] are taking the distribution information into consideration, and also encourage the new measurements to contain the distribution information.

This paper proposes a measurement containing the distribution information, it is motivated that the nearer a sample is with the same labeled samples, the easier it can be classified correctly. The proposed method calculates the average number of the k nearest neighbors in the same class in different subsets under the weighted k-NN, after that, the product of these average values is regarded as the measurement of this dataset. It improves the correlation between the measurement and the final classification performance, which indicate that the proposed is more informative. This paper is arranged as follows: section 2 describes the related work in measurement of the imbalanced dataset, section 3 shows the proposed measurement improved generalized imbalanced ratio (IGIR), and section 4 describes the experimental results and analysis, the final section concludes the proposed method and the future work.

2 Related work

There are many different factors which have effects on the imbalanced classification, resulting in various kinds of measurements considering different factors. For example, the Imbalance Ratio (IR) is based on the difference in sample size, the maximum Fisher's discriminant ratio (F1) is based on the overlap in feature values from different classes, and the Complexity Measurement (CM), Generalized Imbalance Ratio (GIR), and the proposed method Improved Generalized Imbalance Ratio (IGIR) are based on the idea that data distribution plays important role in imbalanced classification. 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 $N=N_- + N_+$.

2.1 IR

Imbalance ratio [11], [12], [13] the definition is as follows, it is defined as the size sample ratio:

$$IR = \frac{N_-}{N_+} \tag{1}$$

When samples with different labels have the same distribution, the sample size is able to reflect whether the samples are easy to classify, otherwise, the IR is not so informative to indicate whether the dataset is easy to classify. 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 is not in (b), so we can get 100% accuracy in (a) but cannot in (b) with a same linear 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 data distribution cannot be represented in IR.

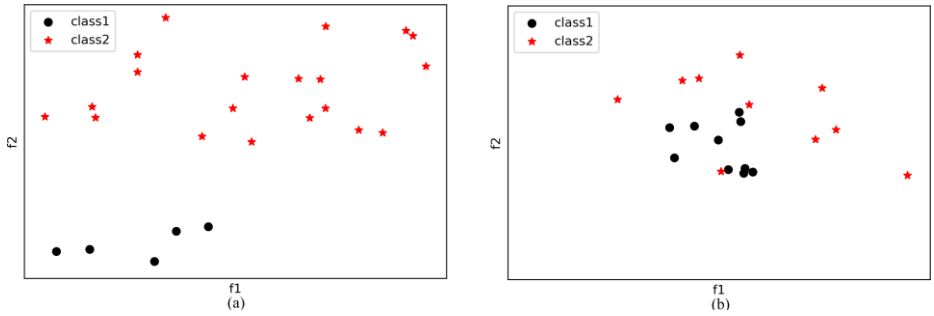

Fig. 1. The dilemma of IR.

2.2 F1

A classical measure of the discriminative power of the covariates, or features, is Fisher's discriminant ratio [9], and F1 is the maximum Fishers discriminant ratio:

$$f = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2} \tag{2}$$

Where μ_1, μ_2, σ_1^2, σ_2^2 are the means and variances of the positive and negative subsets, respectively. For multidimensional problems, the maximum f over all features can be used. However, if f gets 0, it 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.

2.3 CM

CM [7] focuses on the local information for each data point via the nearest neighbors, and uses this information to capture data complexity.

$$\mathrm{CM}_{k(j)} = I\left(\frac{^{Number\ of\ patterns}_{j' in\ N_j\ with\ y_j'=y_j}}{k} \leq 0.5\right) \tag{3}$$

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

$$\mathrm{CM_k} = \frac{1}{n}\sum_{j=1}^{n} CM_{k(j)} \tag{4}$$

CM is determined by the label of its neighbors. If the neighbors of a sample contain more samples in the same class, the sample will be easier to classify. 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.

2.4 GIR

The GIR [8] 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 most simple 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.

$$T_+ = \frac{1}{N_+}\sum_{x \in P}\frac{1}{k}\sum_{r=1}^{k} IR(x, X) = \frac{1}{N_+}\sum_{x \in P} t_k(x) \tag{5}$$

$$T_- = \frac{1}{N_-}\sum_{x \in N} t_k(x) \tag{6}$$

$$GIR = \mathrm{T_-} - \mathrm{T_+} \tag{7}$$

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, 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 [8] successfully applies GIR to oversampling and under-sampling algorithms. The experimental results show that GIR-based resampling algorithm can effectively improve the classification performance.

However, there are two problems in GIR, first, in the classification process, besides 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 the 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

3 The Proposed Method

We proposed an improved measurement called IGIR in this paper, it is based on the idea that the sample distribution plays an important role in the classification result, 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.

$$weight_r = \frac{k-r}{k}, r = 0,1,2 \cdots k-1 \tag{8}$$

$$\text{wei-}T_+ = \frac{1}{N_+} \sum_{x \in P} 1/k \sum_{r=1}^{k} weight_r * IR(x,X) = \frac{1}{N_+} \sum_{x \in P} t_k(x) \tag{9}$$

$$\text{wei-}T_- = \frac{1}{N_-} \sum_{x \in N} t_k(x) \tag{10}$$

$$\text{wei-IGIR} = \sqrt{\text{wei-}T_- * \text{wei-}T_+} \tag{11}$$

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 and spired by the definition of geometric mean, 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: Computing the Improved Generalized Imbalance Ratio IGIR

Input: A dataSet X,label Y,number of nearest neighbors k in k-NN
Output: IGIR

compute *weight* according to formula (8)
for x *in* X *with label* y_x **do**
    $M \leftarrow$ the k nearest neighbors of x
    $t_k(x) \leftarrow \frac{1}{k} \sum weight * IR(x, M)$
end
$wei - T_- \leftarrow \frac{1}{N_-} \sum t_k(x) * sgn(y_x == 0)$
$wei - T_+ \leftarrow \frac{1}{N_+} \sum t_k(x) * sgn(y_x == 1)$
IGIR $\leftarrow \sqrt{wei - T_- * wei - T_+}$
return IGIR

In the proposed method, firstly, calculate the *weight* according to formula (8), secondly, compute the k nearest neighbors of each sample and $t_k(x)$ for each sample, thirdly, compute the average $t_k(x)$ of the positive and negative subsets, finally, compute the IGIR according to the formula (11).

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.

4 Experimental results

4.1 Datasets

The experimental data in this paper comes from the UCI machine learning database [14]. 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, and the details are shown as Table 4.

4.2 Evaluation

In the binary imbalanced classification, the confusion matrix is often used to evaluate the performance of the classifier, which is defined in Table 2:

Table 2. 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 [15].

$$\text{sensitivity} = \frac{TP}{TP+FP} \tag{12}$$

$$\text{recall} = \frac{TP}{TP+FN} \tag{13}$$

$$\text{F-value} = \frac{(1+\beta^2)\times\text{recall}\times\text{precision}}{\beta^2\times\text{recall}+\text{precision}} \tag{14}$$

$$\text{Gmean} = \sqrt{\frac{\text{TP}}{\text{TP+FN}}\times\frac{\text{TN}}{\text{TN+FP}}} \tag{15}$$

4.3 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.

The classification results and measurements of different datasets are shown in Table 5, as we can see from the table, the IR and F1 have no value restriction, so the value can be very huge, while the CM, GIR and IGIR are limited in [0, 1], the scatter plots are used to show a more clear relation between the measurements and classification as shown in Fig. 2. 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.

4.4 Analysis

In order to quantitatively analyze the relationship between different measurements and the classification results, the results are further analyzed by the determination coefficient R^2. R^2 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.

$$R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}} \tag{16}$$

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.

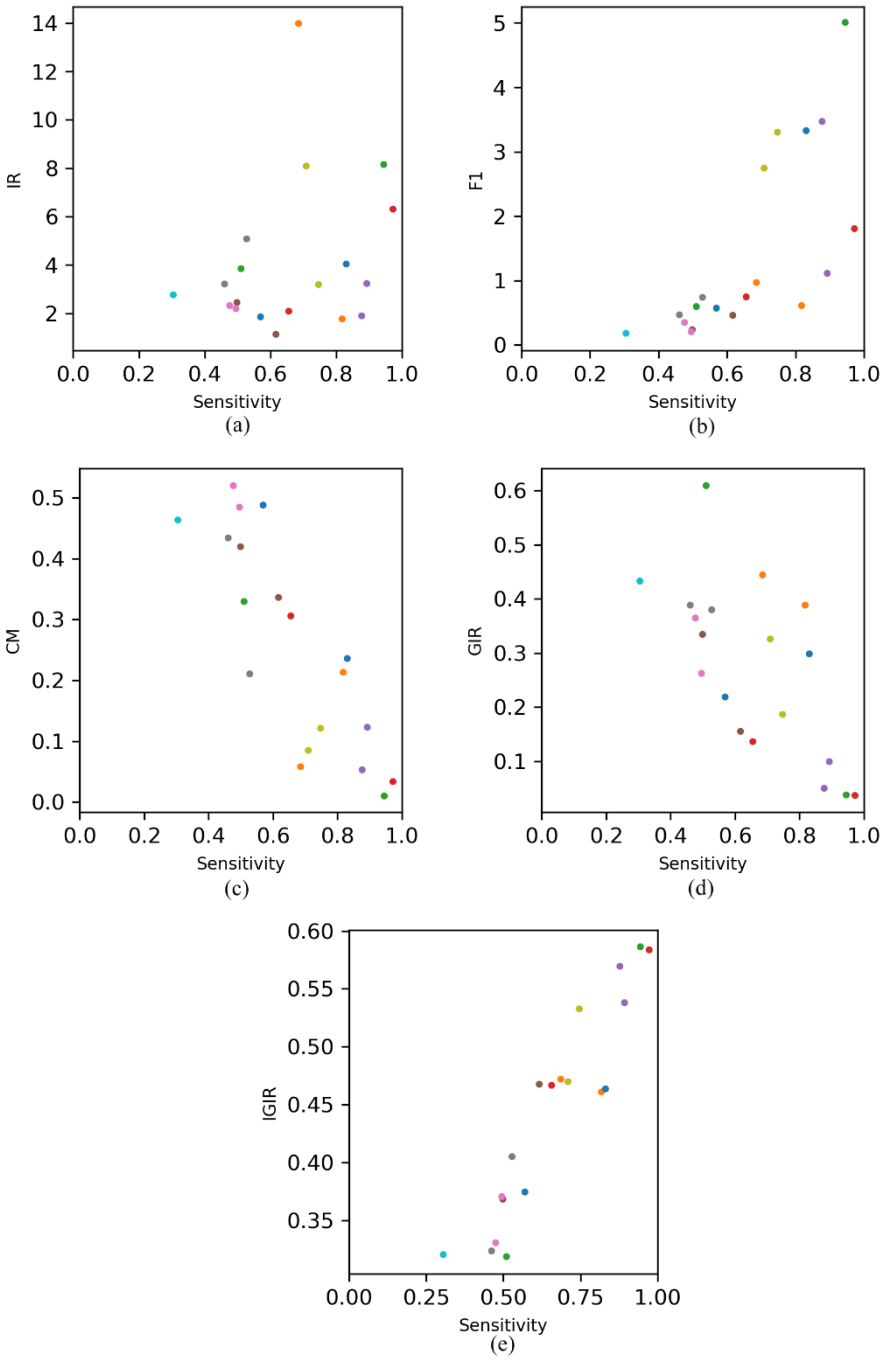

Fig. 2. Measurements and sensitivity

Table 3. R^2 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

The R^2 in Table 3 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.

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.

Table 4. 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

Table 5. 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

5 Conclusion

In this paper, an improved measurement for imbalanced datasets is proposed, it takes the distribution information into consideration and it is based on the idea that a sample surrounded by more same class samples is easier to classify, for each sample of different classes, the proposed method calculates the average number of the k nearest neighbors in the same class in different subsets under the weighted k-NN, after that, the product of these average values is regarded as the measurement of this dataset. The experimental results show that the proposed 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. Singh, S., Liu, Y., Ding, W., Li, Z.: Empirical Evaluation of Big Data Analytics using Design of Experiment: Case Studies on Telecommunication Data. (2016).
6. Hale, M.L., Walter, C., Lin, J., Gamble, R.F.: A Priori Prediction of Phishing Victimization Based on Structural Content Factors. (2017).
7. 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).
8. Tang B, He H: GIR-based ensemble sampling approaches for imbalanced learning. Pattern Recognition 71, 306–19 (2017).
9. Ho T: A Data Comxity Analysis of Comparative Advantages of Decision Forest Constructors. Pattern Analysis & Applications 5(2), 102–12 (2002).
10. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. Journal of Artificial Intelligence Research. 16 (1), 321–357 (2002).
11. Han, H., Wang, W.Y., Mao, B.H.: Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning. Lecture Notes in Computer Science. 3644 (5), 878–887 (2005).
12. Zhang, M.: Foundations of modern analysis. Academic Press (1960)
13. Weiss, G.M.: Learning with rare cases and small disjuncts. In: Twelfth International Conference on International Conference on Machine Learning. pp. 558–565. (1995).
14. Zhang, H., Wang, Z.: A normal distribution-based over-sampling approach to imbalanced data classification. In: International Conference on Advanced Data Mining and Applications. pp. 83–96. (2011)
15. Li, D.C., Hu, S.C., Lin, L.S., Yeh, C.W.: Detecting representative data and generating synthetic samples to improve learning accuracy with imbalanced data sets. Plos One. 12 (8), e0181853 (2017).
16. Moreo, A., Esuli, A., Sebastiani, F.: Distributional Random Oversampling for Imbalanced Text Classification. 805–808 (2016).
17. Amini MR, Usunier N, Goutte C, http://archive.ics.uci.edu/ml/datasets.html, last accessed 2018/03/22.
18. Menardi, G., Torelli, N.: Training and assessing classification rules with imbalanced data. Data Mining & Knowledge Discovery 28(1), 92–122 (2014).