Pattern Recognition Letters
journal homepage: www.elsevier.com

Adjusting the imbalance ratio by the dimensionality of imbalanced data

Rui Zhu[a,**], Yiwen Guo[b], Jing-Hao Xue[c]

[a] *Faculty of Actuarial Science and Insurance, Cass Business School, City, University of London, London EC1Y 8TZ, UK*
[b] *Intel Corporation, Beijing 100190, China*
[c] *Department of Statistical Science, University College London, London WC1E 6BT, UK*

Article history:

Keywords: Imbalanced data, imbalance extent, imbalanced learning, imbalance ratio, Pearson correlation test

ABSTRACT

Class-imbalance extent metrics measure how imbalanced the data are. In pattern classification, it is usually expected that the higher the imbalance extent, the worse the classification performance, and thus an appropriate imbalance extent metric should show a negative correlation with the classification performance. Existing metrics, such as the popular imbalance ratio (IR), only consider the effect of the sample sizes of different classes. However, we note that the dimensionality of imbalanced data also affects the classification performance. Datasets with the same IR can present distinct classification performances when their dimensionalities are different, making IR suboptimal to reflect the imbalance extent for classification. We also observe that the classification performance becomes better with more discriminative features. Inspired by these observations, we propose a new imbalance extent metric, the adjusted IR, by adding a penalty term of the number of discriminative features that is effectively determined by the Pearson correlation test. The adjusted IR adaptively revises the IR when the number of discriminative features varies. The empirical studies demonstrate the effectiveness of the adjusted IR, in terms of its better negative correlation with the classification performance.

1. Introduction

In real-world classification problems, we often encounter imbalanced data in which the sample sizes of different classes are largely distinct. For example, fraudulent transaction detection is an important task for credit card companies to prevent huge financial losses. The number of fraudulent transactions is usually much smaller than that of non-fraudulent ones, which makes standard classifiers tend to misclassify fraudulent transactions as non-fraudulent ones. However, this is not what we desire, because the fraudulent transactions are actually the ones that we aim to detect.

Imbalanced learning [1–5] is a valuable research area studying how to better classify imbalanced data, especially when we aim to detect a class with very few instances. The classes with large sample sizes are called majority classes, while those with small sample sizes are called minority classes.

To determine how imbalanced the data are, class-imbalance extent metrics have been developed, playing an important role in imbalanced learning. For example, such a metric can be decisive in developing a new algorithm, which re-weights the training instances to reverse the negative effect of imbalance. In addition, when we aim to demonstrate the effectiveness of a new imbalanced learning algorithm, we often show that it is superior to existing methods in classifying the data that are very imbalanced, because highly-imbalanced data are usually hard to classify. Thus we expect an appropriate class-imbalance extent measure to be negatively correlated with the classification performance of imbalanced data.

The most popular class-imbalance extent measure is the imbalance ratio (IR), which is simply calculated as the ratio of the sample size of the largest majority class and that of the smallest minority class. Thus the larger the value of IR, the larger the imbalance extent. However, IR is not an effective imbalance extent measure when we have multiple classes, because the information of the classes with sample sizes in between the two extremes is not considered, and IR is considered as a low-

[**] Corresponding author: Tel.: +44(0)1227 82 7008
e-mail: rui.zhu@city.ac.uk (Rui Zhu)

resolution metric for multi-class imbalanced data. Ortigosa-Hernández et al. [6] propose the imbalance degree (ID) to measure the class-imbalance extent of multi-class imbalanced data. ID considers the information from all classes in the data and is a high-resolution metric. However, the penalty term involving the number of minority classes makes ID not well correlated with the classification performance. This is because it is not always true that the class-imbalance extent is higher for the data with more minority classes [7]. Zhu et al. [7] propose a new metric, the likelihood-ratio imbalance degree (LRID), which is based on the likelihood ratio test. It is a high-resolution metric and is negatively correlated with the classification performance.

However, the class-imbalance extent measures proposed in literature all focus on describing the class distribution, which only considers the sample sizes of different classes. Besides the sample size, the dimensionality (i.e. the number of variables or features) is another important property of a dataset, which can affect the classification performance. Dimension reduction methods to select or extract discriminative features can improve classification performances [8–10]. We note that the dimensionality of an imbalanced dataset can also affect the classification performance. Given datasets with the same class distribution, it is possible that their classification performances are distinct when their dimensionalities are different. For example, suppose we have two datasets that the distributions of their sample sizes of different classes are the same. Then we will obtain the same class-imbalance extent for the two datasets by using the existing metrics, because these metrics only consider the class distribution. Now if one dataset adds more features that contain discriminative information than the other dataset, we can expect that the classification performance of the first dataset will be better than the other, because of the valuable discriminative information introduced by the additional discriminative features. In this case, the existing metrics do not work because they produce the same imbalance extent for both datasets and cannot show a negative correlation with classification performance. We need a smaller imbalance extent for the first dataset while a larger one for the second dataset, even though the two datasets have exactly the same class distribution.

This motivates our proposal to adjust the IR by penalising it with the dimensionality of imbalanced data, as we aim to establish a new metric that presents a better negative correlation with classification performance. In statistics and machine learning, adjusting the metrics by the dimensionality has been often used in model selection to select a parsimonious model for the ultimate aim, in line with the Occam's razor principle. One famous example is the Akaike information criterion (AIC), which is a metric to assess how a statistical model, e.g. a linear regression model, fits a given dataset [11]. We usually expect a good statistical model to have high goodness-of-fit, i.e. can fit the data as well as possible, and to be as a simple model as possible. This is because a model with a large number of variables can have high goodness-of-fit, but results in overfitting to the data which is harmful to predict future cases. AIC achieves this expectation by adjusting the commonly used goodness-of-fit measure, the log likelihood, by a penalty term that penalises the number of variables in the model. Inspired by the metrics such as AIC, we aim to improve the IR, the most widely used imbalance extent metric, by penalizing it with the dimensionality of the data.

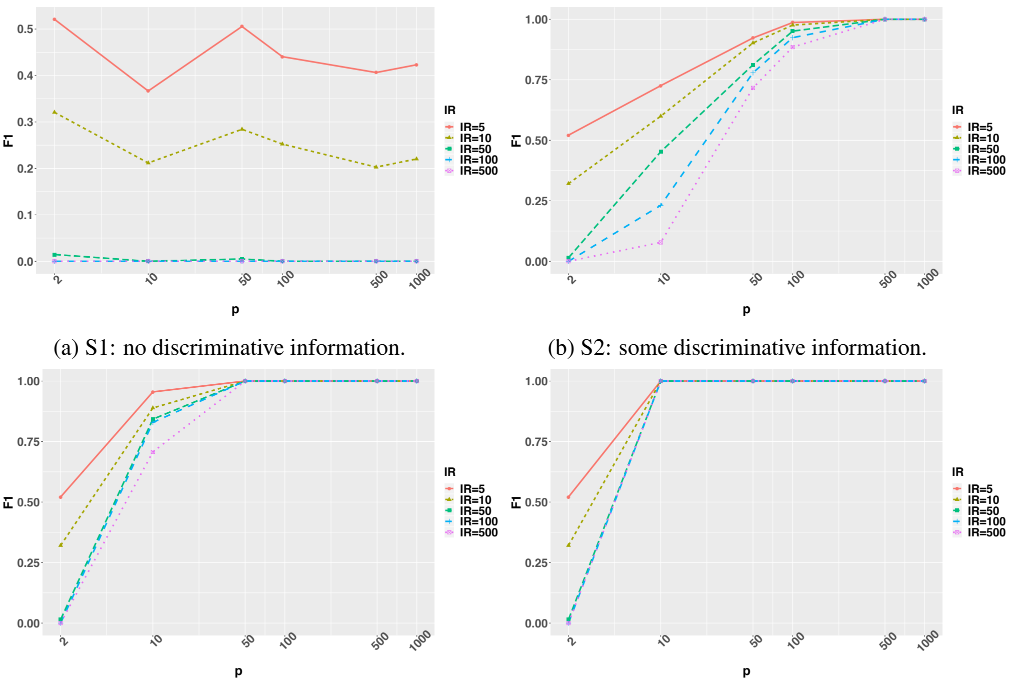
Therefore, in this letter we propose a new class-imbalance extent measure, the adjusted imbalance ratio (adjusted IR), by penalising the original IR with a penalty term involving the number of *discriminative* features in a dataset. We first present some empirical evidence to support our argument that, for datasets with the same class distribution, their classification performances can be different when they have different dimensionalities. We also show that the classification performance will get better if the number of discriminative features increases, given that the total number of features keeps the same. A discriminative feature is able to distinguish different classes, and we can expect a dataset with more discriminative features to have better classification performance. Thus instead of using the total number of features of a dataset, we design a penalty term that is based on the number of discriminative features that are effectively determined by the Pearson correlation test. The adjusted IR can provide a smaller value for data with more discriminative features while a larger value for data with fewer discriminative features, when the two datasets have the same class distribution. Hence in this case, the adjusted IR is negatively correlated with the classification performance while the IR cannot show this essential property.

The contributions of this letter are two-fold. First, we propose a new class-imbalance extent metric, the adjusted IR, to consider the effects of both the dimensionality and the class sizes of imbalanced data on the classification performance. To the best of our knowledge, it is the first time that the dimensionality is involved in designing an imbalance extent metric. Second, we also show extensive simulation studies on the effect of the dimensionality on the classification performance of an imbalanced dataset, which supports our proposal for using the data dimensionality to equip IR with better correlation with classification performance.

The rest of this letter is organised as follows. We first show some empirical results of the effect of the dimensionality of imbalanced data on classification performance, and propose the adjusted IR in Section 2. We then in Section 3 demonstrate that the proposed adjusted IR has good negative correlations with the classification performances. In Section 4, we discuss the effectiveness of using the number of *discriminative* features in the penalty term, comparing it with simply using the number of features. Finally, we draw the conclusions in Section 5.

2. Methodology

In this section, we propose a new class-imbalance extent metric, the adjusted IR, that considers both the class distribution and the dimensionality of an imbalanced dataset. Before introducing the new metric, we show some empirical results on how the dimensionality affects the classification performance of an imbalanced dataset, which motivates our proposal.



(a) S1: no discriminative information.

(b) S2: some discriminative information.

(c) S3: large discriminative information.

(d) S4: very large discriminative information.

Fig. 1: The F1 scores for data with different degrees of discriminative information.

2.1. How does the dimensionality affect the classification performance of an imbalanced dataset?

The imbalance ratio (IR) is the most commonly used measure to describe the imbalance extent of a dataset. IR is defined as

$$\mathrm{IR} = \frac{N_{\mathrm{maj}}}{N_{\mathrm{min}}}, \tag{1}$$

where N_{maj} is the sample size of the majority class and N_{min} is the sample size of the minority class. When there are multi-classes, i.e. the number of classes is larger than 2, N_{maj} is the sample size of the largest majority class and N_{min} is the sample size of the smallest minority class. It is clear that when $\mathrm{IR} = 1$, we have an exactly balanced dataset. When $\mathrm{IR} > 1$, the larger the IR, the larger the imbalance extent of the dataset.

We usually expect that a dataset with higher IR is more difficult to classify, such that, if a new imbalanced classifier has better classification performance on this dataset than other existing classifiers, we can verify the superiority of the new classifier. However, this is often not true for IR.

For classification tasks, we also expect better classification performance if we reasonably expand the original feature space, e.g. create more new features from the original features, which is also one of the motivations for applying the kernel tricks [12]. Hence, it is possible that two datasets with the same IR but different dimensionalities p can have very different classification performances. That is, it is not suitable to claim that these two datasets have the same imbalance extent, although they have the same IR, because we may expect the dataset with more discriminative features has better classification performance.

To support the above argument, we show some empirical results from the following two simulations. Firstly, in simulation 1, we aim to study how the classification performance is changed when we add additional features with different discriminative abilities to the imbalanced data. Secondly, in simulation 2, we aim to study how the classification performance is changed when the features are a mixture of discriminative and non-discriminative features with different mixing proportions.

2.1.1. Simulation 1: additional features with different discriminative abilities

We simulate datasets with six values of p: 2, 10, 50, 100, 500 and 1000. For each p, we simulate two classes with N_{min} and N_{maj} instances, respectively. We test five pairs of $(N_{\mathrm{min}}, N_{\mathrm{maj}})$: (10, 50), (10, 100), (10, 500), (10, 1000) and (10, 5000), corresponding to $\mathrm{IR} = 5$, $\mathrm{IR} = 10$, $\mathrm{IR} = 50$, $\mathrm{IR} = 100$ and $\mathrm{IR} = 500$, respectively. That is, given a fixed IR, we study the classification performances of datasets with six different p's.

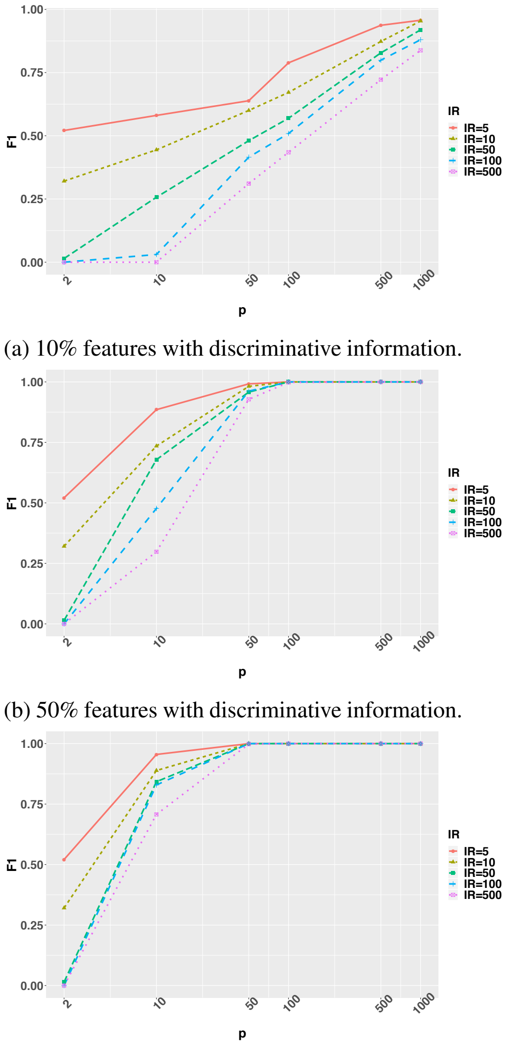
The two classes are simulated from two multivariate normal distributions, $N(\boldsymbol{\mu}_{\mathrm{min}}, \boldsymbol{\Sigma}_{\mathrm{min}})$ and $N(\boldsymbol{\mu}_{\mathrm{maj}}, \boldsymbol{\Sigma}_{\mathrm{maj}})$, respectively, where the subscripts 'min' and 'maj' denote the minority class and the majority class, respectively. For $p = 2$, we set $\boldsymbol{\mu}_{\mathrm{min}} = (1, 1)^T$, $\boldsymbol{\mu}_{\mathrm{maj}} = (2, 2)^T$ and $\boldsymbol{\Sigma}_{\mathrm{min}} = \boldsymbol{\Sigma}_{\mathrm{maj}} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$. We use $p = 2$ as the base line and add features with different degrees of discriminative ability for $p > 2$. Here we believe that given fixed variances, the larger the difference between $\boldsymbol{\mu}_{\mathrm{min}}$ and $\boldsymbol{\mu}_{\mathrm{maj}}$, the higher the discriminative abilities of the features. We simulate the following four situations for $p > 2$.

S1 When non-discriminative features are added: we set $\boldsymbol{\mu}_{\mathrm{min}} = (1, 1, \underbrace{1, \ldots, 1}_{p-2})^T$, $\boldsymbol{\mu}_{\mathrm{maj}} = (2, 2, \underbrace{1, \ldots, 1}_{p-2})^T$ and $\boldsymbol{\Sigma}_{\mathrm{min}} = \boldsymbol{\Sigma}_{\mathrm{maj}} = \mathrm{diag}(0.5, 0.5, \underbrace{0, \ldots, 0}_{p-2})$, where $\mathrm{diag}(\cdot)$ denotes a diagonal matrix. Thus under this setting, we simply add features with the same values for both classes with zero-variances, which does not have discriminative ability.

S2 When features with modest discriminative information are added: we set $\boldsymbol{\mu}_{\text{min}} = (1, 1, \underbrace{1, \dots, 1}_{p-2})^T$, $\boldsymbol{\mu}_{\text{maj}} = (2, 2, \underbrace{1.5, \dots, 1.5}_{p-2})^T$ and $\boldsymbol{\Sigma}_{\text{min}} = \boldsymbol{\Sigma}_{\text{maj}} = \text{diag}(\underbrace{0.5, \dots, 0.5}_{p})$.

S3 When features with large discriminative information are added: we set $\boldsymbol{\mu}_{\text{min}} = (1, 1, \underbrace{1, \dots, 1}_{p-2})^T$, $\boldsymbol{\mu}_{\text{maj}} = (2, 2, \underbrace{2, \dots, 2}_{p-2})^T$ and $\boldsymbol{\Sigma}_{\text{min}} = \boldsymbol{\Sigma}_{\text{maj}} = \text{diag}(\underbrace{0.5, \dots, 0.5}_{p})$.

S4 When features with very large discriminative information are added: we set $\boldsymbol{\mu}_{\text{min}} = (1, 1, \underbrace{1, \dots, 1}_{p-2})^T$, $\boldsymbol{\mu}_{\text{maj}} = (2, 2, \underbrace{3, \dots, 3}_{p-2})^T$ and $\boldsymbol{\Sigma}_{\text{min}} = \boldsymbol{\Sigma}_{\text{maj}} = \text{diag}(\underbrace{0.5, \dots, 0.5}_{p})$.



(a) 10% features with discriminative information.

(b) 50% features with discriminative information.

(c) 90% features with discriminative information.

Fig. 2: The F1 scores for data with different k.

For each situation above with fixed p and IR, i.e. for a fixed parameter setting, we simulate 10 datasets with 10 different seeds. We apply the support vector machine (SVM) to classify the data and record the F1 scores which are widely used in measuring the classification performance of imbalanced data [2]. The average of the 10 F1 scores is used as the classification result for a situation with fixed p and IR. For SVM, we use the svm function in R's 'e1071' package [13] with linear kernel.

The F1 scores for the four situations are shown in Fig. 1. To achieve better visualisation, we use the log scale on the horizontal axis for p. We have the following observations from Fig. 1. First and most importantly, for S2, S3 and S4, given a fixed IR, the classification performance becomes better when p increases, which supports our argument that it is not suitable to use the same IR to describe the imbalance extent of the data with different dimensionality. We can also observe that the more discriminative the additional features, the smaller the p we need to achieve F1 scores of ones. Secondly, when non-discriminative features are added as in S1, the classification performance varies with different p's to a small extent for all IRs, but there is no clear pattern. This indicates that by adding non-discriminative features, we cannot obtain better classification performance. Thirdly, for each situation, datasets with small IRs tend to have better classification performances compared with those with large IR, especially when p is small. This also makes sense because when all other settings are fixed, we expect datasets with smaller imbalance extent easier to classify.

2.1.2. Simulation 2: additional discriminative and non-discriminative features with different mixing proportions

In real-world data, features are usually not all discriminative or non-discriminative as in simulation 1. Here we simulate data with a mixture of discriminative and non-discriminative features and study the change of classification performance when the mixing proportion changes. Some settings here are similar to those of simulation 1.

We simulate datasets with six different values of p: 2, 10, 50, 100, 500 and 1000. For each p, we simulate two classes with N_{min} and N_{maj} instances, respectively. We test five pairs of $(N_{\text{min}}, N_{\text{maj}})$: (10, 50), (10, 100), (10, 500), (10, 1000) and (10, 5000), corresponding to IR = 5, IR = 10, IR = 50, IR = 100 and IR = 500, respectively.

The two classes are from two multivariate normal distributions, $N(\boldsymbol{\mu}_{\text{min}}, \boldsymbol{\Sigma}_{\text{min}})$ and $N(\boldsymbol{\mu}_{\text{maj}}, \boldsymbol{\Sigma}_{\text{maj}})$. For $p = 2$, we set $\mu_{min} = (1, 1)^T$, $\boldsymbol{\mu}_{\text{maj}} = (2, 2)^T$ and $\boldsymbol{\Sigma}_{\text{min}} = \boldsymbol{\Sigma}_{\text{maj}} = \begin{pmatrix} 0.5 & 0 \\ 0 & 0.5 \end{pmatrix}$, which is also used as a base line. We add both discriminative and non-discriminative features with different mixing proportions for $p > 2$. For discriminative features, we set their means as 1 and 2 for the two classes, respectively, while for non-discriminative features, we set their means for both classes as 1s. For each $p > 2$, we randomly select k% features from the $p - 2$ features (except for the first two features) as discriminative features and the rest are non-discriminative features. The covariance matrices for all settings are $\boldsymbol{\Sigma}_{\text{min}} = \boldsymbol{\Sigma}_{\text{maj}} = \text{diag}(\underbrace{0.5, \dots, 0.5}_{p})$. Thus for each $p > 2$, we add k% discriminative features and $(100 - k)$% non-discriminative features. We test three values of

k: 10, 50 and 90.

For fixed p, IR and k, we simulate 10 datasets with 10 different seeds and record the average of F1 scores from SVM as the classification result for a specific parameter setting.

The F1 scores for the simulated data with different k's are shown in Fig. 2. The patterns are similar to those in Fig. 1. For fixed IR and k, the F1 score increases as p increases, and for a fixed IR, as k increases, the classification performance becomes better. They suggest that the classification performance becomes better when more discriminative features are added given all other settings fixed. This also supports our proposal for adjusting IR by the number of discriminative features, considering the correlation between an imbalance measure and the dimensionality of an imbalanced dataset.

To sum up, we draw the following conclusions from the simulation studies. First, for datasets with the same IR but different dimensionalities p, we can observe different classification performances. Thus it is not suitable to use the same IR to describe the imbalance extent of these datasets, considering the required negative correlation between the imbalance extent measure and the classification performance. Secondly, for a fixed IR, the larger the number of discriminative features, the better the classification performance. This makes sense because additional discriminative features can bring more discriminative information to enlarge the separation between classes and help classification. This observation is also consistent with the motivation for applying the kernel trick that with a proper expansion of the feature space we may obtain better classification results.

2.2. *Adjusted IR*

The above empirical results suggest that as the number of discriminative features increases, we need a smaller imbalance extent to be negatively correlated with the classification performance. Therefore in this letter, we propose a new imbalance measure, the adjusted imbalance ratio (adjusted IR), to consider the effect of dimensionality on the classification performance.

In the rest of this section, we first introduce how to determine which features are discriminative by a simple yet effective statistical hypothesis testing method, the Pearson correlation test. We then propose the adjusted IR based on a penalty term involving the number of discriminative features.

2.2.1. *The Pearson correlation test*

To determine which features are discriminative, we employ the Pearson correlation test, which can effectively detect the non-zero correlation between two variables. If the correlation between a feature vector and the label vector is non-zero, then we can treat this feature as discriminative.

Given a dataset with N instances and p features, $\{(\mathbf{x}_i, y_i)\}_{i=1}^N$, where $\mathbf{x}_i = [x_{i1}, x_{i2}, \dots, x_{ip}]^T \in \mathbb{R}^{p\times 1}$ and $y_i \in \{-1, +1\}$, we denote the jth feature as $\mathbf{x}_j = [x_{j1}, x_{j2}, \dots, x_{jN}]^T \in \mathbb{R}^{N\times 1}$ ($j = 1, 2, \dots, p$) and the label vector as $\mathbf{y} = [y_1, y_2, \dots, y_N]^T$.

In the Pearson correlation test, we test the null hypothesis H_0 that the Pearson correlation ρ_j between the jth feature variable X_j and the label variable Y is zero, i.e. $\rho_j = 0$, against the alternative hypothesis H_1 that $\rho_j \neq 0$. The test statistic is defined as

$$t_j = r_j \sqrt{\frac{N-2}{1-r_j^2}}, \tag{2}$$

where r_j is the sample Pearson correlation between $\mathbf{x}_j$ and $\mathbf{y}$ and t follows a Student's t-distribution with $N-2$ degrees of freedom. We reject the null hypothesis H_0 if $t_j > t_{N-2,1-\alpha}$ or $t_j < -t_{N-2,1-\alpha}$, where α is the significance level of the test.

For $\mathbf{x}_j$, if we reject H_0, then we believe that the correlation between X_j and Y is non-zero and the jth feature is discriminative. We apply the Pearson correlation test to each feature and count the number of tests with H_0 being rejected. This number represents the number of discriminative features and we denote it as p^* to distinguish it from the total number of features p.

2.2.2. *Adjusted IR*

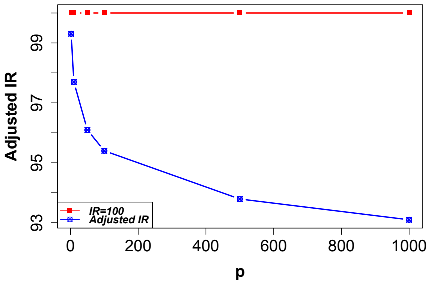


Fig. 3: Adjusted IR for different p when IR= 100.

We propose the adjusted IR as follows

$$\text{Adjusted IR} = \text{IR} - \lambda \log(p^*), \tag{3}$$

where p^* is the number of discriminative features determined by the Pearson correlation test and λ is the parameter that controls the importance of the penalty term. The penalty term $\log(p^*)$ can adjust the effect of the dimensionality on the classification performance. It is clear that for datasets with a fixed IR, the adjusted IR decreases as the number of discriminative features p^* increases. Thus the adjusted IR can have a better negative correlation with the classification performance, since the classification performance becomes better as p^* increases. The larger the value of the adjusted IR, the higher the imbalance extent of one dataset.

Fig. 3 shows the values of adjusted IR with $\lambda = 1$ for different p when IR= 100. The horizontal axis shows the value of p instead of p^* to better compare IR and the adjusted IR. Comparing the shape of the curve of adjusted IR and the curves of F1 scores in Fig. 1 and Fig. 2, we can expect good correlations between the adjusted IR and the F1 scores. We will show more results on the correlations for the adjusted IR in section 3. On the contrary, IR is a constant for all p, which makes it not a good imbalance extent measure being independent of classification performance.

We note that in some extreme cases it is possible to have $p^* = 0$, in which case $\log(p^*)$ is not defined. To address this problem, we simply set $p^* = 1$ when $p^* = 0$.

3. Experiments

In the following experiments, we compare the performances of IR and the adjusted IR, based on their correlations with the classification performances on both simulated data and real data. For the adjusted IR, we have to decide the significance level α to get p^*. Here we simply take the commonly used $\alpha = 5\%$. We also set $\lambda = 1$ for simplicity.

3.1. Simulated data

Table 1: The correlations between the adjusted IR the F1 scores of SVM for simulations 1 and 2.

	Criterion	IR= 5	IR= 10	IR= 50	IR= 100	IR= 500
Simulation 1	**SRCC**	-0.80	-0.78	-0.77	-0.76	-0.77
	PCC	-0.75	-0.75	-0.75	-0.75	-0.76
Simulation 2	**SRCC**	-0.89	-0.89	-0.89	-0.89	-0.94
	PCC	-0.83	-0.87	-0.85	-0.85	-0.89

The datasets simulated in section 2 are adopted here. In simulations 1 and 2, we generate 10 datasets with 10 different seeds for each parameter setting. We take the average of the 10 adjusted IR for the 10 datasets as the adjusted IR for each specific parameter setting.

Following [6] and [7] , we calculate both the Spearman rank correlation coefficient (SRCC) and the Pearson correlation coefficient (PCC) to assess the correlation between the adjusted IR and the F1 score. For simulation 1, we combine the F1 scores and the adjusted IRs for datasets generated in situations S2, S3 and S4 with fixed IR and calculate the correlations between them. We do not include S1 in this calculation because S1 is an extreme case where the additive features are exactly the same for both classes and the standard deviations of these features are zeros which makes r_j in equation (2) not defined. For simulation 2, we combine the F1 scores and the adjusted IRs for datasets generated for $k = 10$, $k = 50$ and $k = 90$ with fixed IR and calculate the correlations between them.

The correlations between the adjusted IR and the F1 scores for the two simulations are shown in Table 1. Obviously, the adjusted IR have good negative values of SRCC and PCC for all values of IR, while IR cannot show its correlation with the classification performances in this case because it is fixed.

To compare the performances of IR and the adjusted IR more straightforwardly, we use all datasets in simulations 1 and 2 and calculate the correlations between the imbalance metrics and the classification performances. In other words, we do not fix IR in this case. The values of SRCC and PCC are shown in Table 2. It is clear that the adjusted IR has better negative correlations for both simulations and $\lambda = 1$ is a proper choice for the simulated data.

Table 2: The correlations of IR and the adjusted IR with the F1 scores of SVM for simulations 1 and 2. The values in bold faces denote the best performances.

	Criterion	IR	Adjusted IR
Simulation 1	**SRCC**	-0.08	**-0.24**
	PCC	-0.13	**-0.14**
Simulation 2	**SRCC**	-0.18	**-0.36**
	PCC	-0.17	**-0.18**

Table 3: The description of real datasets.

Name	**Class frequencies**	**p**	**p***	**IR**	**Adjusted IR**	**F1 score**
glass1	(138, 76)	9	3	1.8	0.7	0.09
pima	(500, 268)	8	7	1.9	-0.1	0.65
wisconsin	(443, 239)	9	9	1.9	-0.3	0.96
ecoli1	(259, 77)	6	3	3.4	2.3	0.73
new-thyroid1	(180, 35)	5	5	5.1	3.5	0.96
segment0	(1973, 329)	18	17	6.0	3.2	0.99
glass6	(179, 29)	9	6	6.2	4.4	0.82
yeast3	(1319, 169)	8	4	8.1	6.7	0.75
page-blocks0	(4912, 559)	10	10	8.8	6.5	0.73
vowel0	(898, 89)	13	7	10.1	8.1	0.83
led7digit_vs_1	(406, 37)	7	5	11.0	9.4	0.82
glass2	(197, 17)	9	2	11.6	10.9	0.01
cleveland-0_vs_4	(160, 13)	13	11	12.3	9.9	0.58
glass4	(201, 13)	9	5	15.5	13.9	0.28
abalone9-18	(688, 42)	8	7	16.4	14.4	0.34
dermatology-6	(330, 20)	34	21	16.5	13.5	0.98
glass5	(205, 9)	9	4	22.8	21.4	0.30
yeast4	(1432, 51)	8	4	28.1	26.9	0.00
poker-9_vs_7	(234, 8)	10	3	29.3	28.2	0.21
yeast5	(1439, 44)	8	4	32.7	31.3	0.44

Table 4: The correlations of IR and the adjusted IR with the F1 scores of SVM for real datasets. The values in bold faces denote the best performances.

	Criterion	IR	Adjusted IR
Real data	**SRCC**	-0.41	**-0.50**
	PCC	-0.51	**-0.54**

3.2. Real data

Twenty real imbalanced datasets for binary classification are downloaded from the KEEL-dataset repository [14]. A summary of the datasets can be found in Table 3. The datasets are sorted by the values of IR in ascending order. By adjusting $\log(p^*)$, the values of adjusted IR show a different ranking order compared with IR.

Similarly to the simulated data, the linear SVM is applied to all datasets. We randomly split each dataset to a training set containing 70% instances of each class and a test set containing the rest of the dataset. We repeat the random split ten times and record the mean F1 scores which are shown in the last column of Table 3.

We report the values of SRCC and PCC between the F1 scores and the imbalance metrics in Table 4. Clearly, the adjusted IR has better negative SRCC and PCC compared with IR, which demonstrates the effectiveness of including the penalty term $\log(p^*)$ for these data, and $\lambda = 1$ remains a good choice for the real data.

4. Discussion

The adjusted IR defined in equation (3) adjusts IR by the log transformation of the number of discriminative features, $\log(p^*)$. However, it is also straightforward to use the log transformation of the total number features, $\log(p)$, without the process to identify discriminative features. In this section, we aim to show that it is necessary to adopt the number of discriminative features p^* in the penalty term, in order to have a more effective imbalance extent measure.

In this section, we denote the adjusted IR with $\log(p)$ as 'adjusted IR (p)' and that with $\log(p^*)$ as 'adjusted IR (p^*)'. Fig. 4 shows the curves of adjusted IR (p) and adjusted IR (p^*) for IR= 100. As with Fig. 3, the horizontal axis shows the values

Table 5: The correlations of the adjusted IR (p) and the adjusted IR (p^*) with the F1 scores of SVM for simulations 1 and 2. The values in bold faces denote the best performances.

	Measure	Criterion	IR = 5	IR = 10	IR = 50	IR = 100	IR = 500
Simulation 1	**Adjusted IR (p)**	**SRCC**	-0.68	-0.69	-0.68	-0.68	-0.69
		PCC	-0.71	-0.72	-0.72	-0.71	-0.71
	Adjusted IR (p^*)	**SRCC**	**-0.80**	**-0.78**	**-0.77**	**-0.76**	**-0.77**
		PCC	**-0.75**	**-0.75**	**-0.75**	**-0.75**	**-0.76**
Simulation 2	**Adjusted IR (p)**	**SRCC**	-0.74	-0.74	-0.74	-0.75	-0.81
		PCC	-0.77	-0.81	-0.82	-0.81	-0.81
	Adjusted IR (p^*)	**SRCC**	**-0.89**	**-0.89**	**-0.89**	**-0.89**	**-0.94**
		PCC	**-0.83**	**-0.87**	**-0.85**	**-0.85**	**-0.89**

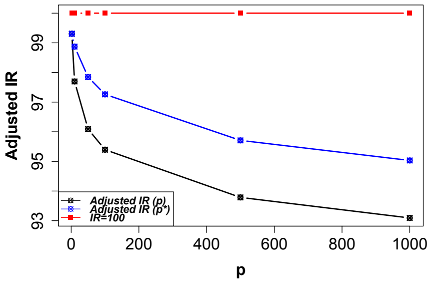


Fig. 4: The adjusted IR (p) and adjusted IR (p^*) for different p when IR= 100.

of p to make the comparison more straightforward. It is observed that the curve of adjusted IR (p^*) is above that of the adjusted IR (p), because $p^* \leq p$.

The correlations with the F1 scores for the adjusted IR (p) and the adjusted IR (p^*) are reported in Table 5. We can observe that SRCC and PCC of the adjusted IR (p^*) are better than those of the adjusted IR (p) for both simulations and for all values of IR, which shows the effectiveness of using p^* to adjust IR.

This result makes sense because for a fixed p, the classification performance becomes better when the proportion of discriminative features increases, which can be observed in Fig. 2. In this case, the adjusted IR (p) does not change and cannot reflect the change in classification performance, while the adjusted IR (p^*) is negatively correlated with the classification performance.

5. Conclusions and future work

In this letter, by studying how the classification performance of an imbalanced dataset can be affected by its dimensionality, and by showing that the classification performance can become better as the number of discriminative features increases for a fixed IR, we propose a new class-imbalance extent metric, the adjusted IR. The adjusted IR considers the effect of the dimensionality of discriminative features in imbalanced data on the classification performance, and applies the Pearson correlation test to identify the discriminative features.

For simplicity and illustrative purposes, $\lambda = 1$ already shows promising performance in the experiments to verify the effectiveness of our proposed idea. Nonetheless, for different data, a different value of λ can be expected to produce an even better measure of imbalance extent; this is an interesting direction to explore further in the future. In addition, the adjusted IR is designed to enhance IR, which is more suitable for two-class data, and thus similarly to IR, the adjusted IR only considers the information from the two extreme classes, ignores that from other classes, and is a low-resolution metric in the case of multi-class data. Extending the proposed idea to adjusting and enhancing existing metrics for multi-class data, for example, ID [6] or LRID [7], by further considering the number of discriminative features, is another interesting future research direction. Moreover, the effect of dimensions on classifying imbalanced data can also be considered when designing novel imbalanced learning algorithms.

References

[1] J.-H. Xue, D. M. Titterington, Do unbalanced data have a negative effect on LDA?, Pattern Recognition 41 (5) (2008) 1558–1571.

[2] H. He, E. A. Garcia, Learning from imbalanced data, IEEE Transactions on Knowledge and Data Engineering 21 (9) (2009) 1263–1284.

[3] J.-H. Xue, P. Hall, Why does rebalancing class-unbalanced data improve AUC for linear discriminant analysis?, IEEE Transactions on Pattern Analysis and Machine Intelligence 37 (5) (2015) 1109–1112.

[4] A. Luque, A. Carrasco, A. Martín, A. de las Heras, The impact of class imbalance in classification performance metrics based on the binary confusion matrix, Pattern Recognition 91 (2019) 216–231.

[5] X. Zhang, D. Wang, Y. Zhou, H. Chen, F. Cheng, M. Liu, Kernel modified optimal margin distribution machine for imbalanced data classification, Pattern Recognition Letters 125 (2019) 325–332.

[6] J. Ortigosa-Hernández, I. Inza, J. A. Lozano, Measuring the class-imbalance extent of multi-class problems, Pattern Recognition Letters 98 (2017) 32–38.

[7] R. Zhu, Z. Wang, Z. Ma, G. Wang, J.-H. Xue, LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test, Pattern Recognition Letters 116 (2018) 36–42.

[8] J. Yu, Y. Rui, D. Tao, Click prediction for web image reranking using multimodal sparse coding, IEEE Transactions on Image Processing 23 (5) (2014) 2019–2032.

[9] C. Gong, D. Tao, J. Yang, K. Fu, Signed laplacian embedding for supervised dimension reduction, in: Twenty-Eighth AAAI Conference on Artificial Intelligence, 2014.

[10] C. Gong, D. Tao, S. J. Maybank, W. Liu, G. Kang, J. Yang, Multi-modal curriculum learning for semi-supervised image classification, IEEE Transactions on Image Processing 25 (7) (2016) 3249–3260.

[11] H. Akaike, Information theory and an extension of the maximum likelihood principle, in: Selected papers of Hirotugu Akaike, Springer, 1998, pp. 199–213.

[12] J. Friedman, T. Hastie, R. Tibshirani, The Elements of Statistical Learning, Vol. 1, Springer series in statistics New York, 2001.

[13] D. Meyer, E. Dimitriadou, K. Hornik, A. Weingessel, F. Leisch, e1071: Misc functions of the department of statistics, probability theory group (Formerly: E1071), TU Wien, R package version 1.6-8 (2017).
URL https://CRAN.R-project.org/package=e1071

[14] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, F. Herrera, KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework., Journal of Multiple-Valued Logic & Soft Computing 17.