Pattern Recognition Letters
journal homepage: www.elsevier.com

LRID: A new metric of multi-class imbalance degree based on likelihood-ratio test

Rui Zhu[a,**], Ziyu Wang[b,e], Zhanyu Ma[c], Guijin Wang[d], Jing-Hao Xue[e]

[a] *School of Mathematics, Statistics and Actuarial Science, University of Kent, Parkwood Road, Canterbury, CT2 7FS, UK*
[b] *Department of Security and Crime Science, University College London, London WC1E 6BT, UK*
[c] *The Pattern Recognition and Intelligent System Laboratory, Beijing University of Posts and Telecommunications, Beijing 100876, China*
[d] *Department of Electronic Engineering, Tsinghua University, Beijing 100084, China*
[e] *Department of Statistical Science, University College London, London, WC1E 6BT, UK*

ABSTRACT

In this paper, we introduce a new likelihood ratio imbalance degree (LRID) to measure the class-imbalance extent of multi-class data. Imbalance ratio (IR) is usually used to measure class-imbalance extent in imbalanced learning problems. However, IR cannot capture the detailed information in the class distribution of multi-class data, because it only utilises the information of the largest majority class and the smallest minority class. Imbalance degree (ID) has been proposed to solve the problem of IR for multi-class data. However, we note that improper use of distance metric in ID can have harmful effect on the results. In addition, ID assumes that data with more minority classes are more imbalanced than data with less minority classes, which is not always true in practice. Thus ID cannot provide reliable measurement when the assumption is violated. In this paper, we propose a new metric based on the likelihood-ratio test, LRID, to provide a more reliable measurement of class-imbalance extent for multi-class data. Experiments on both simulated and real data show that LRID is competitive with IR and ID, and can reduce the negative correlation with F1 scores by up to 0.55.

1. Introduction

Imbalanced learning is an important research topic in machine learning (He and Garcia, 2009; Wang and Yao, 2012; Xue and Titterington, 2008; Xue and Hall, 2015). Imbalanced data are the data that have unequal class distributions: majority classes have much more samples than minority classes. Minority classes in imbalanced data can be easily misclassified by standard learning algorithms, which can lead to heavy costs in practice.

A lot of imbalanced learning algorithms have been developed over the past decade. To design algorithms that can deal with the class-imbalance problem, several approaches are widely adopted, such as the resampling approach (Zhu et al., 2017; Castellanos et al., 2018), the cost-sensitive approach (Cheng et al., 2016) and the ensemble approach (Sun et al., 2015; Lusa et al., 2016; Tang and He, 2017; Yuan et al., 2018). Most of imbalanced learning algorithms are designed to solve binary classification problems, while multi-class imbalanced learning still needs much development (Wang and Yao, 2012).

In imbalanced learning, the class-imbalance extent is an important measurement to describe how imbalanced the data are (Ortigosa-Hernández et al., 2017). Usually, the more imbalanced the data, the larger the harmful effect on the classification results. An algorithm can be identified as better than others if it performs better on the data that are more imbalanced. Moreover, the class-imbalance extent can be included in the design of a learning algorithm to improve the learning performance.

Imbalance ratio (IR) is the most commonly adopted metric for class-imbalance extent (He and Garcia, 2009). It is calculated as the ratio of the number of samples in the largest majority class to that in the smallest minority class. Although it is a good imbalance metric for binary-class data, IR cannot provide a high resolution description of imbalance extent for multi-class data because it only considers the information of the largest class and the smallest class and ignores the information of classes in between.

Ortigosa-Hernández et al. (Ortigosa-Hernández et al., 2017) first propose a new metric, imbalance degree (ID), to provide a

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

high resolution imbalance-extent measurement for multi-class data. ID is a sum of two components: 1) the normalised distance between the class distribution of the given data and that of the exactly balanced data, which takes values in $[0, 1]$, and 2) $m - 1$, where m is the number of minority classes. By measuring the difference between class distributions in the first component, ID makes use of information in all classes and can provide a higher resolution measurement than IR. The second component in ID ensures that the class-imbalance extent of data with more minority classes is definitely higher than that with less minority classes because ID takes values in $[m - 1, m]$.

In this paper, we note two problems of ID in the two components. First, although the first component can capture the information from all classes, the distance metric adopted can have large effect on the result. Several distance metrics are tested in Ortigosa-Hernández et al. (2017). However, which distance metric is suitable for the problem at hand is unknown. Second, the argument that the class-imbalance extent is higher for the data with more minority classes seems reasonable at the first glance, however, it is not always true. For example, given two datasets with three classes, one dataset has class frequencies of $(1, 1000, 1000)$ and the other dataset has class frequencies of $(1000, 1000, 1003)$. Clearly, the second dataset is roughly balanced while the first dataset is imbalanced. However, the ID of the second dataset is larger than that of the first dataset because the second one has two minority classes. Thus it is not reliable to use the number of minority classes in ID without considering how minor the classes are.

To solve the above two problems, we propose a new class-imbalance extent metric for multi-class data, the likelihood ratio imbalance degree (LRID). We employ a natural and effective statistic, the log-likelihood ratio (Rice, 2006), to measure the difference between the class distribution of the imbalanced data and that of the exactly balanced data. Thus, LRID does not suffer from the problem of choosing proper distance metrics in practice. The number of minority classes is also not needed in LRID. Thus the second problem in ID is also solved by LRID. Experiments on both simulated data and real data demonstrate the effectiveness of LRID to measure the imbalance extent of multi-class data.

Our contributions are two-fold. First, we are the first to explore the likelihood-ratio test to measure the imbalance degree, which is demonstrated as a simple and effective measurement in experiments. Second, we show that the performances of the imbalance degree metrics can also be affected by the intrinsic properties of data, which has not been discussed in literature.

The rest of the paper is organised as follows. In Section 2, we first formulate the imbalance problem and discuss the problems of IR and ID. We then propose LRID as a more effective and reliable metric that can be easily applied in practice. In Section 3, we compare IR, ID and LRID using both simulated data and real data. Lastly, in Section 4, we present some concluding remarks.

2. Methodology

In this section, we first formulate the imbalance problem based on the multinomial distribution following Ortigosa-Hernández et al. (2017). Then we introduce two measurements in literature, the imbalance ratio (IR) and the imbalance degree (ID), and discuss their advantages and disadvantages for multi-class data. Lastly, to solve the problems in IR and ID, we propose a new measurement, likelihood ratio imbalance degree (LRID), that can effectively measure the imbalance extent for multi-class data.

2.1. *Formulate the imbalanced problem by using multinomial distribution*

Here we show how an imbalanced problem can be formulated by using the multinomial distribution. A generative classification model learns the joint distribution, $p(\mathbf{x}, y) = p(y)p(\mathbf{x}|y)$, where $\mathbf{x} \in \mathbb{R}^{p\times 1}$ is the data vector, y is its label and $p(y)$ is the prior knowledge on the probability of label y. Suppose there are C possible outcomes for y: $\mathbf{y} = [y_1, y_2, \dots, y_C]$. Then each outcome y_c is associated with a probability p_c and we have $\sum_{c=1}^{C} p_c = 1$. Thus the frequencies of the possible labels, $\mathbf{n} = [n_1, n_2, \dots, n_C]$, can be modelled by using a multinomial distribution, $Multinomial(N, \mathbf{p})$, with parameters N and $\mathbf{p} = [p_1, p_2, \dots, p_C]$.

In practice, the parameter N in $Multinomial(N, \mathbf{p})$ is usually taken as known: the total number of observations $\sum_{c=1}^{C} n_c$. The parameter p_c is usually estimated as the fraction of the number of observations in the cth class: $\hat{p}_c = \frac{n_c}{N}$. Thus the class distribution vector $\mathbf{p}$ is estimated by $\hat{\mathbf{p}} = [\hat{p}_1, \hat{p}_2, \dots, \hat{p}_C]$.

For exactly balanced data, $p_c = \frac{1}{C}$ $\forall c$. However, for imbalanced data, there are differences between the frequencies of classes, i.e. n_c's are not all the same, which makes $\hat{p}_c$'s different. The classes with $\hat{p}_c \geq \frac{1}{C}$ are defined as the majority classes while those with $\hat{p}_c < \frac{1}{C}$ are defined as the minority classes.

If we use $\mathbf{b} = [\frac{1}{C}, \frac{1}{C}, \dots, \frac{1}{C}]$ to denote the class distribution vector for exactly balanced data, then a metric to measure the class-imbalance extent can be a single value that can summarise the difference between $\hat{\mathbf{p}}$ and $\mathbf{b}$.

2.2. *Imbalance ratio*

Imbalance ratio (IR) measures the class-imbalance extent using the extreme values in $\hat{\mathbf{p}}$:

$$\text{IR} = \frac{\hat{p}_{\max}}{\hat{p}_{\min}}, \tag{1}$$

where $\hat{p}_{\max}$ and $\hat{p}_{\min}$ are the maximum and minimum values in $\hat{\mathbf{p}}$, respectively. Clearly, for multi-class data, p_c's between $\hat{p}_{\max}$ and $\hat{p}_{\min}$ are ignored in IR. Class distributions with the same $\hat{p}_{\max}$ and $\hat{p}_{\min}$ while different p_c's in between have the same IR. Thus IR is considered as a low-resolution metric to describe class-imbalance extent for multi-class data (Ortigosa-Hernández et al., 2017).

2.3. *Imbalance degree*

To solve the problem of IR, Ortigosa-Hernández et al. (2017) propose the following high-resolution metric to summarise the difference between $\hat{\mathbf{p}}$ and $\mathbf{b}$:

$$\text{ID} = \frac{d(\hat{\mathbf{p}}, \mathbf{b})}{d(\mathbf{p}_m, \mathbf{b})} + (m - 1), \tag{2}$$

where m is the number of minority classes, $\mathbf{p}_m$ describes the situation where there are exactly m minority classes in a dataset, and $d(\mathbf{p}_m, \mathbf{b})$ is the maximum distance between $\mathbf{b}$ and all possible $\mathbf{p}_m$. Ortigosa-Hernández et al. (2017) show that the maximum distance is achieved when $\mathbf{p}_m$ is a C-dimensional class distribution vector with m zeros, $(C-m-1)$ $\frac{1}{C}$s and one $1-\frac{C-m-1}{C}$: $[\underbrace{0,\dots,0}_{m}, \underbrace{\frac{1}{C},\dots,\frac{1}{C}}_{C-m-1}, \underbrace{1-\frac{C-m-1}{C}}_{1}]$.

The first term in (2) is the normalised distance between $\hat{\mathbf{p}}$ and $\mathbf{b}$ with values in $[0,1]$, which utilises information of all classes. Thus ID considers detailed information in $\hat{\mathbf{p}}$ and is a high-resolution metric.

However, the distance metric used in the first term can have large effect on the results and there is no rule about how to choose a proper distance metric in practice. In this paper, we aim to provide a simple and effective metric that can be easily applied in practice, without trialling different distance metrics or parameters.

If we only use the first term as ID, it is possible to obtain the same ID for different class distributions $\hat{\mathbf{p}}$. Thus the second term is added to make ID an injection function that has different values for different numbers of minority/majority classes (Ortigosa-Hernández et al., 2017).

There are two problems associated with the second term. First, it is not necessary to make ID an injection function. This is because it is reasonable for different class distributions to have the same class-imbalance extent. We will show this argument empirically in Subsection 3.1.2. In addition, the argument that ID is an injection function holds only for data with the same number of classes. If two datasets have different numbers of classes but the same number of minority classes, their IDs can still be the same.

Second, introducing the second term in ID can cause extra problems in measuring class-imbalance extent. The ID of a dataset with m minority classes has value in $[m-1, m]$, as $\frac{d(\hat{\mathbf{p}},\mathbf{b})}{d(\mathbf{p}_m,\mathbf{b})} \in [0,1]$. Thus the ID of a dataset with a large m is definitely larger than that with a small m. However, it is not always true that the larger the number of minority classes, the higher the imbalance extent. Suppose we have the following two datasets with $C = 3$: 1) $\hat{\mathbf{p}}_1 = [\frac{1}{100000}, \frac{1}{2} - \frac{1}{200000}, \frac{1}{2} - \frac{1}{200000}]$ and 2) $\hat{\mathbf{p}}_2 = [\frac{1}{3.1}, \frac{1}{3.1}, 1 - \frac{2}{3.1}]$. $\text{ID}(\hat{\mathbf{p}}_1) = c_1 + 0 \in [0,1]$ and $\text{ID}(\hat{\mathbf{p}}_2) = c_2 + 1 \in [1,2]$, where c_1 and c_2 are the values of the first terms. Thus the second one is considered to be more imbalanced than the first one because its ID is larger. However, although it has two minority classes, the second dataset is roughly balanced. The first dataset is extremely imbalanced with one probability close to zero. Therefore ID fails to provide reliable class-imbalance measurement in this case.

2.4. Likelihood ratio imbalance degree

To solve the problems in ID, we propose a new metric of class-imbalance extent for multi-class data, the likelihood ratio imbalance degree (LRID).

First, since an improper distance metric may have harmful effect on ID, we propose not to use the distance metric between two distributions in the imbalance extent measurement. Instead, we explore a natural and powerful statistical inference technique, the likelihood-ratio (LR) test (Rice, 2006), to provide a single value that can well summarise the difference between $\hat{\mathbf{p}}$ and $\mathbf{b}$.

Given a dataset with C classes and $\mathbf{n} = [n_1, n_2, \dots, n_C]$, the LR test for the multinomial distribution $Multinomial(N, \mathbf{p})$ aims to test the null hypothesis that the parameters $\mathbf{p}$ equal to specific values. Here we aim to test whether $\mathbf{p}$ can be well fitted by $\mathbf{b}$, i.e. the balanced class distribution. Thus we test H_0: $\mathbf{p} = \mathbf{b}$ against H_1: $\mathbf{p} = \hat{\mathbf{p}}$. The LR test statistic is $-2\ln[L(\mathbf{b}|\mathbf{n})/L(\hat{\mathbf{p}}|\mathbf{n})]$, where $L(\cdot)$ is the likelihood function. Thus for balanced data, $L(\mathbf{b}|\mathbf{n}) = L(\hat{\mathbf{p}}|\mathbf{n})$ and the value of the test statistic is 0; while for imbalanced data, $L(\mathbf{b}|\mathbf{n}) < L(\hat{\mathbf{p}}|\mathbf{n})$ and the value of the test statistic is larger than 0. The larger the difference of the estimated class distribution $\hat{\mathbf{p}}$ from the balanced class distribution $\mathbf{b}$, the larger the value of the test statistic. Therefore the value of the test statistic can be used to measure the difference between $\hat{\mathbf{p}}$ and $\mathbf{b}$, or the class-imbalance extent. Moreover, similarly to the first term in ID, the LR test statistic considers the information of all classes and is a high-resolution measurement.

Second, as we have discussed in the previous section, the second term in ID, $(m-1)$, is an unnecessary term and brings problems to the metric. Thus, in our new metric, we propose to eliminate this term and simply use the LR test statistic as the metric. We term this metric as the likelihood-ratio imbalance degree (LRID). When $\hat{p}_c = \frac{n_c}{N}$, LRID can be written as

$$\text{LRID} = -2\sum_{c=1}^{C} n_c \ln\frac{b_c}{\hat{p}_c} = -2\sum_{c=1}^{C} n_c \ln\frac{N}{Cn_c}. \qquad (3)$$

3. Experiments

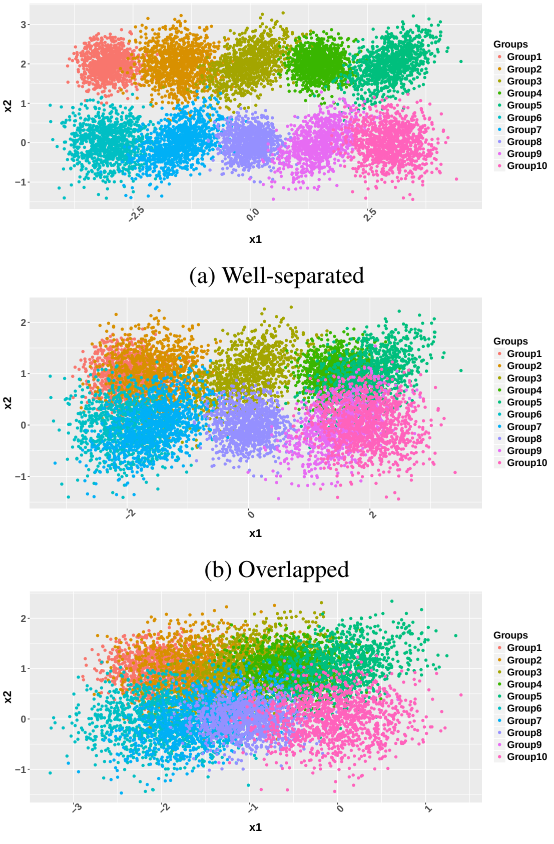
In the following experiments, we compare three imbalance degree metrics, IR, ID and LRID, on both simulated and real datasets. In ID, as suggested by the experiment results in Ortigosa-Hernández et al. (2017), the total variation distance and the Hellinger distance have the best performance and we test both distance metrics in our experiments. IDs using the total variation distance and the Hellinger distance are denoted as ID_{TV} and ID_{HE}, respectively.

The performances of the three metrics are tested by following the two criteria proposed in Ortigosa-Hernández et al. (2017): 1) the resolution of the metric and 2) the correlation between the metric and the classification performance. A better metric is expected to have higher resolution and more negative correlation with classification performance (the more imbalanced, the worse the classification performance). In this paper, the classification performance is measured by the F1 score, which is a widely used metric in imbalanced learning (He and Garcia, 2009). Linear discriminant analysis (LDA) and support vector machine (SVM) are adopted as the classification algorithms.

3.1. Simulated-data experiments

3.1.1. Experiment settings for simulated data

Here we design experiments to compare the performances of the class-imbalance metrics for data with different class-



(a) Well-separated

(b) Overlapped

(c) Extremely-overlapped

Fig. 1: Balanced datasets for three separation degrees.

separation degrees. Prati et al. (2004) show that the classification performance of imbalanced data can also be affected by the intrinsic properties of data. Since the correlation between the classification performance and the class-imbalance metric is one of the criteria to measure the performance of metrics, we aim to test whether other properties of data, such as the separation of classes, can affect the performances of the metrics. We simulate three sets of data with different separation degrees of classes: well separated, overlapped and extremely overlapped. We measure the separation degrees of data by an index called 'separability index' (SI) (Greene, 2001; Thornton, 2002; Mthembu and Marwala, 2008). SI measures the proportion of observations that have the nearest neighbour with the same class, taking values between 0 and 100%. The details of the three sets of data are described as follows.

For each dataset, we simulate $N = 10000$ observations with $C = 10$ classes; that is, for fully balanced data, each class contains 1000 observations and $\boldsymbol{b} = [\frac{1}{10}, \frac{1}{10}, \dots, \frac{1}{10}]$. For imbalanced data, the number of minority classes m is set to 1 to 9. For each m, the probability vector of the multinomial distribution is set to $\boldsymbol{p} = [\underbrace{p_{min}, \dots, p_{min}}_{m}, \underbrace{p_{maj} \dots, p_{maj}}_{K-m}]$, where m minority classes have equal probabilities $p_{min} = \frac{1}{10}r$ and $K - m$ majority classes have equal probabilities $p_{maj} = (1 - \frac{m}{10}r)/(K - m)$. To control the imbalance degree, r is set to 0.01, 0.05, 0.1, 0.5 and 0.9. Thus for each number of minority classes m, we set five different numbers of observations for the minority classes, where $r = 0.01$ corresponds to the most imbalanced situation while $r = 0.9$ corresponds to the roughly balanced situation.

Two-dimensional Gaussian features are simulated for each observation. We simulate three sets of means for the ten classes: 1) well-separated data with the means clearly separate, $\{(-3, 2), (-1.5, 2), (0, 2), (1.5, 2), (3, 2), (-3, 0), (-1.5, 0), (0, 0), (1.5, 0), (3, 0)\}$; 2) overlapped data with the means less separate, $\{(-2, 1), (-1.5, 1), (0, 1), (1.5, 1), (2, 1), (-2, 0), (-1.5, 0), (0, 0), (1.5, 0), (2, 0)\}$; and 3) extremely-overlapped data with the means close together, $\{(-2, 1), (-1.5, 1), (-1, 1), (-0.5, 1), (0, 1), (-2, 0), (-1.5, 0), (-1, 0), (-0.5, 0), (0, 0)\}$. Three covariance matrices are randomly assigned to the ten classes: $\begin{pmatrix} 0.1 & 0 \\ 0 & 0.1 \end{pmatrix}$, $\begin{pmatrix} 0.2 & 0 \\ 0 & 0.2 \end{pmatrix}$ and $\begin{pmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{pmatrix}$. The balanced data for the three sets of means are shown in Fig. 1. The SIs of the fully balanced well-separated data, overlapped data and extremely-overlapped data are 100%, 59% and 41%, respectively.

Therefore, for each set of means, we simulate $9 \times 5 = 45$ datasets (9 values of m and 5 values of r) and each dataset has 10000 observations and 10 classes. To make the comparison between different separations of classes fair, we keep the frequency vectors $\mathbf{n}$ the same for the three sets of means; that is, with the same r and m, $\mathbf{n}$ is the same for different sets of means.

We apply LDA and SVM to each dataset and use the F1 score as the metric to assess the classification performances. We perform 20 random training/test splits on each dataset, with 70% training data and 30% test data. In SVM, the radial basis kernel is adopted and the parameters are set to the default values of 'svm' function in the 'e1071' R package (Meyer et al., 2017).

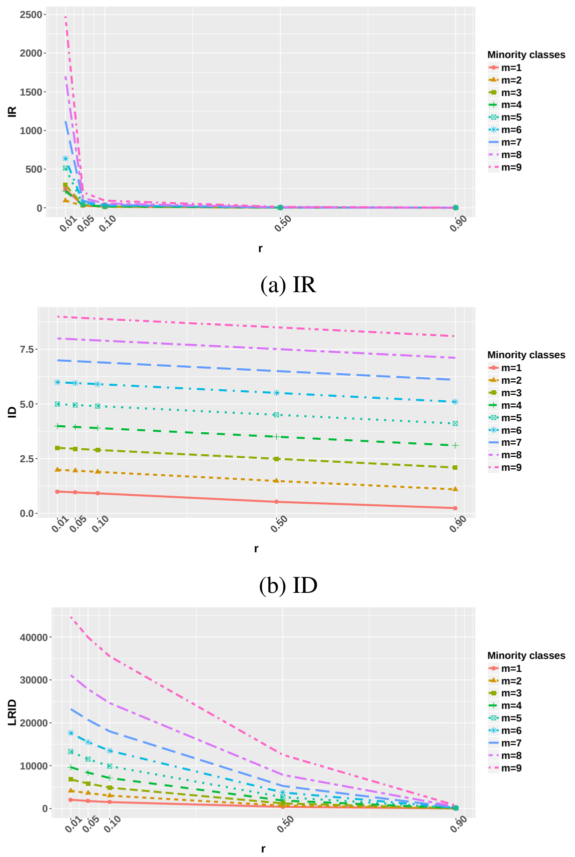
3.1.2. Results of simulated data

i) The resolution of the measurements: Since $\mathbf{n}$'s are the same for the three sets of means, the values of each class-imbalance metric are the same for the three sets of data. The values of the three metrics with different numbers of minority classes m and different imbalance extents measured by r are shown in Fig. 2. Note that we present ID with the total variation distance, ID$_{TV}$, as an example of ID in Fig. 2. ID with other distance metrics should have similar patterns as ID$_{TV}$, because the distance used in the first term of ID is normalised and is between [0,1].

For each plot in Fig. 2, the horizontal axis shows values of r and the vertical axis shows the values of the metrics. Each line in the plot corresponds to a specific value of m.

We observe different patterns for ID, IR and LRID against m and r. IR has the lowest resolution among the three measurements: the lines are close when $r \geq 0.1$ and overlap when $r \geq 0.5$, which indicates that IR cannot well distinguish data with different m. In contrast, ID has the highest resolution: the lines are equally separated, indicating that ID can well distinguish between data with different m. In addition, each line has a downward trending, indicating that the value of ID decreases as r increases. LRID has a resolution level between IR and ID: the distances between lines decrease as r increases. When $r = 0.9$, LRIDs of different numbers of minority classes are similar.

However, resolution is not the only criterion to assess the quality of an imbalance-degree metric. Although ID has the highest resolution, it is not reasonable to have very different values for different m when $r = 0.9$. This is because the datasets

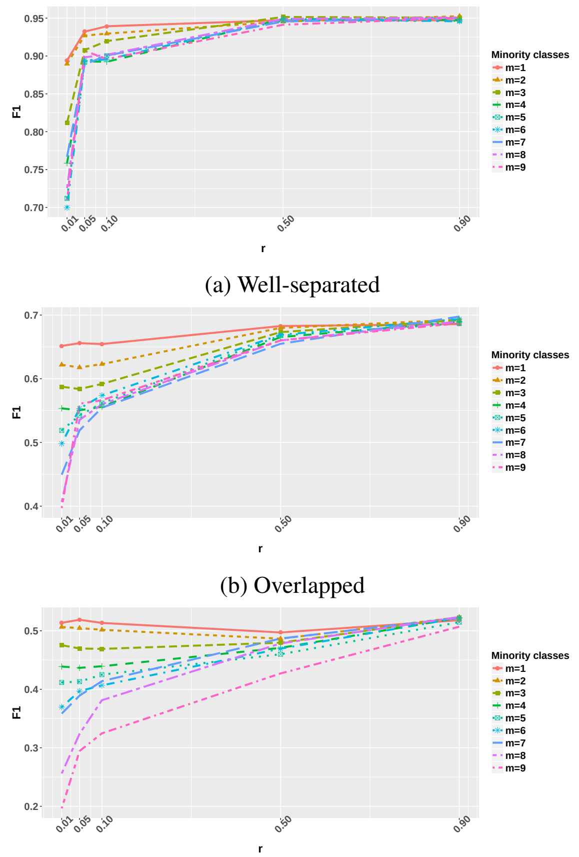


(a) IR

(b) ID

(c) LRID

Fig. 2: The values of three class-imbalance metrics against different numbers of minority classes m and different imbalance extent measured by r. ID_{TV} is presented here as an example of ID.



(a) Well-separated

(b) Overlapped

(c) Extremely-overlapped

Fig. 3: The F1 scores of LDA against different numbers of minority classes m and different imbalance extent measured by r for the three degrees of separation.

are roughly balanced when $r = 0.9$. For example, the dataset with $m = 1$ and $\mathbf{p} = [\underbrace{0.09}_{1}, \underbrace{0.101, \ldots, 0.101}_{9}]$ and the dataset with $m = 5$ and $\mathbf{p} = [\underbrace{0.09, \ldots, 0.09}_{5}, \underbrace{0.11, \ldots, 0.11}_{5}]$ are both roughly balanced and they should have similar imbalance extents. IR and LRID that have similar values for different m are more reasonable in this case. We will discuss more about how this problem will affect the correlation between ID and classification performance in the next section.

ii) The correlation with classification performances: In Ortigosa-Hernández et al. (2017), the correlations between ID and classification performances are calculated with eliminating the second term $(m - 1)$. We denote ID without $(m - 1)$ as ID^*. In this paper, we report the correlations for both ID and ID^*. The Spearman rank correlation coefficient (SRCC) and the Pearson correlation coefficient (PCC) of the metrics with the F1 scores of LDA and the F1 scores of SVM are shown in Table 1, from which we can make the following observations.

First, it is obvious that the performances of the metrics are different for different separation degrees of data. When the data are well separated, IR has the best SRCC and ID^*_{HE} has the best PCC for LDA while IR has the best SRCC and PCC for SVM. However, as the data become more overlapped, LRID becomes the best metric in terms of both SRCC and PCC for both LDA and SVM. In the cases of the largest gaps shown in Table 1, LRID can reduce SRCC by 0.55 compared with IDs and 0.16 compared with IR, and can reduce PCC by 0.34 compared with IDs and 0.42 compared with IR. It is also worth noting that IR becomes worse while IDs become better when the data are more overlapped.

Second, it is also obvious that ID_{TV} and ID_{HE} have worse performances than ID^*_{TV} and ID^*_{HE} for both LDA and SVM, except for the PCC of extremely-overlapped data for SVM. This suggests that including the second term $(m - 1)$ can be harmful in evaluating the imbalance degree for our simulated data. This observation supports our argument in Section 2.3.

Third, the correlations with the F1 scores of LDA and those with the F1 scores of SVM show similar patterns for simulated data, which suggests that the classification algorithms used in the experiments do not have large effect on the patterns of the correlations between metrics and classification performances.

To investigate further for the above observations, we plot the F1 scores against the values of r and m in Fig. 3. Since the patterns of using LDA and SVM are similar and due to the page limit, we only show the plots of the F1 scores of LDA in Fig. 3. For the data that are well separated, the number of minority classes m does not have much effect on the F1 scores when r is large, as shown in Fig. 3a. The lines for $m = 4$ to $m = 9$ are

Table 1: The correlations with the F1 scores of LDA and SVM for simulated data. ID^*_{TV} denotes $\mathrm{ID}_{TV} - (m-1)$ and ID^*_{HE} denotes $\mathrm{ID}_{HE} - (m-1)$. The values in bold faces denote the best performances.

	Data	Criterion	IR	ID_{TV}	ID_{HE}	ID^*_{TV}	ID^*_{HE}	LRID
LDA	**Well-separated**	**SRCC**	**-0.91**	-0.33	-0.33	-0.87	-0.89	-0.84
		PCC	-0.61	-0.35	-0.34	-0.59	**-0.67**	-0.58
	Overlapped	**SRCC**	-0.88	-0.41	-0.41	-0.77	-0.83	**-0.90**
		PCC	-0.66	-0.51	-0.50	-0.70	-0.76	**-0.80**
	Extremely-overlapped	**SRCC**	-0.83	-0.56	-0.56	-0.69	-0.81	**-0.95**
		PCC	-0.47	-0.65	-0.64	-0.69	-0.75	**-0.89**
SVM	**Well-separated**	**SRCC**	**-0.92**	-0.21	-0.21	-0.91	-0.90	-0.76
		PCC	**-0.79**	-0.26	-0.26	-0.58	-0.66	-0.55
	Overlapped	**SRCC**	**-0.90**	-0.40	-0.40	-0.83	-0.87	**-0.90**
		PCC	-0.77	-0.50	-0.49	-0.72	-0.78	**-0.83**
	Extremely-overlapped	**SRCC**	-0.78	-0.61	-0.61	-0.62	-0.73	**-0.94**
		PCC	-0.74	-0.68	-0.67	-0.55	-0.64	**-0.96**

almost overlapped for all values of r and the lines for all m are overlapped for $r \geq 0.5$. Hence, when data are well separated, it is reasonable for data with the same $\hat{p}_{min}$ and similar $\hat{p}_{max}$'s (i.e. with a fixed r) but different $\hat{p}_c$'s in between (i.e. with different m) to have similar imbalance extents in terms of the correlation with classification performance. Therefore we do not need a high-resolution metric under this situation and it makes sense that IR has the best performance in this case.

However, things are different when data are overlapped: the effect of the number of minority classes m becomes large on the F1 scores. The lines are more separated in Fig. 3b and Fig. 3c than in Fig. 3a. When data are overlapped, the larger the number of minority classes, the lower the F1 score. Therefore, IR does not perform well while high-resolution metrics such as ID^* and LRID can perform well in these cases.

To make the above analysis clearer, we compare the plots in Fig. 3 and Fig. 2. It is clear that the plot of IR, Fig. 2a, has the most similar shape (but opposite trend) as Fig. 3a, which explains the good performance of IR for well-separated data in terms of correlations. In addition, the plot of ID_{HE} in Fig. 2b shows the reason for its bad performance: when $r = 0.9$, datasets with different number of minority classes have very similar F1 scores, however, ID_{HE} provides very different imbalance degrees. The plot of LRID in Fig. 2c has the most similar shape with Fig. 3b and Fig. 3c, which explains its best performances in both cases.

To sum up, the simulated experiments show the following conclusions. First, the order of resolution of the three metrics is IR < LRID < ID, as shown in Fig. 2. Second, the separation of the data affects the performance of the metrics in terms of the correlations with classification performance. Data with different number of minority class m can have the same imbalance extent considering their classification performance, as shown in Fig. 3. IR and ID^*_{HE} are the best for well-separated data while LRID is the best for overlapped and extremely overlapped data. Therefore, LRID shows competitive performance compared with other two metrics in terms of both criteria: LRID has a reasonably high resolution and competitive correlations with classification performance. In practice, if we know that the data are well separated, then IR is enough to measure the imbalance extent of multi-class data. However, if we know that the data are overlapped or we are not sure about the separation level of the data, then LRID can be a good candidate.

3.2. Real-data experiments

Twenty UCI datasets are used in the experiments (Dheeru and Karra Taniskidou, 2017): yeast, ecoli, wine, abalone, auto mpg, glass, Hayes-Roth, pageblocks, penbased, shuttle, poker hand, new thyroid, contraceptive method choice, dermatology, statlog, zoo, soybean, wholesale, leaf and lense. The descriptions of the datasets are shown in Table 2. Similarly to the simulated data, LDA and SVM are applied to all datasets. We perform 20 random training/test split on each dataset, with 70% training data and 30% test data. In SVM, the radial basis kernel is adopted and the parameters are set to the default values of the 'svm' function in the 'e1071' R package (Meyer et al., 2017). The means of the F1 scores of LDA and SVM, and the six imbalance degree metrics are recorded for each dataset.

For the 20 datasets, we can obtain two sets of F1 score means with 20 values each set and six sets of imbalance degree metrics with 20 values each set. PCCs and SRCCs between each set of F1 scores and each set of imbalance extent metrics are calculated. Thus for each classification algorithm, we obtain six SRCCs and six PCCs.

3.2.1. Results of real data

The correlations with the F1 scores of LDA and those of SVM for the real datasets are shown in Table 3. For the F1 scores of LDA, LRID can achieve the best SRCC and PCC while for those of SVM, LRID can achieve the best SRCC and ID_{TV} has the best PCC. This result is supported by the SIs of the datasets in Table 2. Over half of the datasets show different degrees of overlapping based on the values of SI. The Abalone dataset has a very low SI of 20%. Thus LRID shows better correlation with the F1 scores based on these datasets. Although LRID does not perform the best for the PCC with the SVM F1 scores, it is still superior compared with IR and ID^*.

Similarly to those of simulated data, the results of real data also suggest that the distance metric can have some effect on the performance of ID. In contrast, the new LRID can provide competitive performance compared with ID while avoiding the difficulty to choose suitable distance metrics.

The results on real data also demonstrate that LRID is a simple and effective measurement of class-imbalance extent of multi-class data.

3.3. Computational time

The computational time of IR, ID and LRID are compared based on the two-dimensional simulated data, by using the "microbenchmark" R package (Mersmann, 2018) on a MacBook Pro with 2.9 GHz Intel Core i5. The mean computational time of 10^6 tests of IR, ID and LRID are 775, 842 and 1555 nanoseconds (10^{-9} seconds), respectively. Although the computational time of LRID is around two times of those of IR and ID, it is still around 10^{-6} seconds, which is a rather small interval of time. Considering its competitive resolution and correlation performance, the computational time of LRID is acceptable.

4. Conclusions and future work

In this paper, we propose a new metric to measure the class-imbalance extent of multi-class data based on the likelihood-

Table 2: The description of real datasets.

Name	**Classes C**	**Minority classes m**	**SI**	**Class frequencies**	**Estimated class distribution**
Yeast	10	6	50%	(463, 5, 3544, 51, 163, 244, 429, 20, 30)	(0.312, 0.003, 0.024, 0.030, 0.034, 0.110, 0.164, 0.289, 0.0135, 0.020)
Ecoli	8	5	81%	(143, 77, 2, 2, 35, 20, 5, 52)	(0.426, 0.229, 0.006, 0.006, 0.104, 0.060, 0.015, 0.155)
Wine	3	2	77%	(59, 71, 48)	(0.331, 0.399, 0.270)
Abalone	23	15	20%	(15, 57, 1, 5, 259, 391, 568, 689, 634, 487, 267 203, 126, 103, 67, 58, 42, 32, 26, 14, 6, 9, 2, 2)	(0.000, 0.000, 0.004, 0.014, 0.028, 0.062, 0.0934, 0.136, 0.165, 0.152, 0.117, 0.064, 0.0486, 0.030, 0.025, 0.016, 0.0139, 0.010, 0.008, 0.006, 0.003, 0.001, 0.002, 0.000, 0.000, 0.000, 0.000, 0.000)
Auto mpg	3	2	71%	(249, 70, 79)	(0.626, 0.176, 0.198)
Glass	6	4	72%	(70, 76, 17, 13, 9, 29)	(0.327, 0.355, 0.0794, 0.0607, 0.042, 0.136)
Hayes-Roth	3	1	48%	(50, 50, 31)	(0.386, 0.386, 0.227)
Pageblocks	5	4	95%	(492, 33, 8, 12, 3)	(0.898, 0.060, 0.005, 0.0161, 0.021)
Penbased	10	5	99%	(780, 779, 780, 719, 780, 720, 720, 778, 719, 719)	(0.104, 0.104, 0.104, 0.096, 0.104, 0.096, 0.096, 0.104, 0.0959, 0.0959)
Shuttle	7	5	99%	(34108, 37, 132, 6748, 2458, 6, 11)	(0.784, 0.0010.0030.1550.057, 0.0000.000)
Poker hand	10	8	51%	(12493, 10599, 1206, 513, 93, 54, 36, 6, 5, 5)	(0.500, 0.424, 0.048, 0.021, 0.004, 0.002, 0.001, 0.000, 0.000, 0.000)
New thyroid	3	2	95%	(150, 35, 30)	(0.700, 0.163, 0.140)
Contraceptive method choice	3	1	46%	(629, 333, 511)	(0.427, 0.226, 0.347)
Dermatology	6	3	97%	(112, 61, 72, 49, 52, 20)	(0.306, 0.167, 0.200, 0.134, 0.142, 0.055)
Statlog	4	2	44%	(26, 22, 21, 25)	(0.277, 0.234, 0.223, 0.266)
Zoo	7	5	72%	(41, 20, 5, 13, 4, 8, 10)	(0.406, 0.198, 0.050, 0.129, 0.040, 0.079, 0.100)
Soybean	4	3	91%	(10, 10, 10, 17)	(0.213, 0.213, 0.213, 0.362)
Wholesale	3	2	57%	(77, 47, 316)	(0.175, 0.107, 0.718)
Leaf	30	17	59%	(11, 10, 10, 8, 12, 8, 10, 11, 14, 13, 16, 12, 13, 12, 10 12, 11, 13, 9, 12, 11, 12, 12, 12, 11, 11, 11, 11, 11, 10)	(0.032, 0.029, 0.030, 0.024, 0.035, 0.024, 0.030, 0.032, 0.041, 0.038, 0.0472, 0.035, 0.038, 0.035, 0.029, 0.035, 0.032, 0.038, 0.027, 0.035, 0.032, 0.035, 0.035, 0.035, 0.032, 0.032, 0.032, 0.032, 0.032, 0.029)
Lense	3	2	38%	(4, 5, 15)	(0.167, 0.208, 0.625)

Table 3: The correlations with the F1 scores of LDA and SVM for real data. ID^*_{TV} denotes $\mathrm{ID}_{TV} - (m-1)$ and ID^*_{HE} denotes $\mathrm{ID}_{HE} - (m-1)$. The values in bold faces denote the best performances.

		IR	**ID$_{TV}$**	**ID$_{HE}$**	**ID$^*_{TV}$**	**ID$^*_{HE}$**	**LRID**
LDA	**SRCC**	-0.49	-0.14	-0.14	-0.32	-0.30	**-0.60**
	PCC	-0.28	-0.28	-0.27	0.04	0.11	**-0.34**
SVM	**SRCC**	-0.44	-0.11	-0.12	-0.25	-0.25	**-0.45**
	PCC	-0.16	**-0.34**	-0.33	0.09	0.15	-0.23

ratio test, the likelihood-ratio imbalance degree (LRID). LRID can provide effective measurement of class-imbalance extent and can be easily applied in practice. In the experiments, LRID demonstrates its superior performances over IR and ID on both simulated and real data.

It is clear that the performances of the current class-imbalance metrics can be affected by the intrinsic properties of data. Designing a imbalance degree metric that can also consider the intrinsic properties of data, such as the separation degree, is of interest in the future.

Acknowledgements

The authors thank the reviewers for their constructive comments. This work was partly supported by the National Natural Science Foundation of China (NSFC) under Grant 61628301 and by the Global Engagement Funds from University College London.

References

Castellanos, F.J., Valero-Mas, J.J., Calvo-Zaragoza, J., Rico-Juan, J.R., 2018. Oversampling imbalanced data in the string space. Pattern Recognition Letters .

Cheng, F., Zhang, J., Wen, C., 2016. Cost-sensitive large margin distribution machine for classification of imbalanced data. Pattern Recognition Letters 80, 107–112.

Dheeru, D., Karra Taniskidou, E., 2017. UCI machine learning repository. URL: `http://archive.ics.uci.edu/ml`.

Greene, J., 2001. Feature subset selection using thorntons separability index and its applicability to a number of sparse proximity-based classifiers, in: Proceedings of Annual Symposium of the Pattern Recognition Association of South Africa.

He, H., Garcia, E.A., 2009. Learning from imbalanced data. IEEE Transactions on Knowledge and Data Engineering 21, 1263–1284.

Lusa, L., et al., 2016. Gradient boosting for high-dimensional prediction of rare events. Computational Statistics & Data Analysis .

Mersmann, O., 2018. microbenchmark: Accurate Timing Functions. URL: `https://CRAN.R-project.org/package=microbenchmark`. r package version 1.4-4.

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

Mthembu, L., Marwala, T., 2008. A note on the separability index. arXiv preprint arXiv:0812.1107 .

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

Prati, R.C., Batista, G.E., Monard, M.C., 2004. Class imbalances versus class overlapping: an analysis of a learning system behavior, in: Mexican International Conference on Artificial Intelligence, Springer. pp. 312–321.

Rice, J., 2006. Mathematical statistics and data analysis. Nelson Education.

Sun, Z., Song, Q., Zhu, X., Sun, H., Xu, B., Zhou, Y., 2015. A novel ensemble method for classifying imbalanced data. Pattern Recognition 48, 1623–1637.

Tang, B., He, H., 2017. GIR-based ensemble sampling approaches for imbalanced learning. Pattern Recognition 71, 306–319.

Thornton, C., 2002. Truth from trash: How learning makes sense. MIT Press.

Wang, S., Yao, X., 2012. Multiclass imbalance problems: Analysis and potential solutions. IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics) 42, 1119–1130.

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

Xue, J.H., Titterington, D.M., 2008. Do unbalanced data have a negative effect on lda? Pattern Recognition 41, 1558–1571.

Yuan, X., Xie, L., Abouelenien, M., 2018. A regularized ensemble framework of deep learning for cancer detection from multi-class, imbalanced training data. Pattern Recognition 77, 160–172.

Zhu, T., Lin, Y., Liu, Y., 2017. Synthetic minority oversampling technique for multiclass imbalance problems. Pattern Recognition 72, 327–340.