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 the superior performance of LRID.

1. Introduction

Imbalanced learning is an important research topic in the machine learning community (He and Garcia, 2009; Wang and Yao, 2012; Xue and Titterington, 2008; Xue and Hall, 2015). Imbalanced data are data that have unequal class distributions: majority classes have much more samples than minority classes. Minority classes in imbalanced data can be easily misclassified using 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 (Nekooeimehr and Lai-Yuen, 2016; Ha and Lee, 2016; Zhu et al., 2017; Castellanos et al., 2018), the cost-sensitive approach (Cheng et al., 2016; Castro and Braga, 2013) 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 further 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 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 for the smallest minority class. Although it is a good imbalance metric for binary-class data, IR cannot provide 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.

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

Ortigosa-Hernández et al. (Ortigosa-Hernández et al., 2017) first propose a new metric, imbalance degree (ID), to provide a 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 class. 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 as a class-imbalance measurement for multi-class data in the two components separately. 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.

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 disadvatages 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 imbalance problem using multinomial distribution

Given data vector $\mathbf{x} \in \mathbb{R}^{p\times 1}$ and its label y, a generative classification model learns the joint distribution:

$$p(\mathbf{x}, y) = p(y)p(\mathbf{x}|y), \tag{1}$$

where $p(y)$ is the prior knowledge on the probability of label y. Suppose there are C possible outcomes for y: $\mathbf{y} = [y_1, y_2, \ldots, y_C]$. Then each outcome y_i is associated with a probability p_c and we have $\sum_{c=1}^{C} p_c = 1$. Thus the frequencies of the possible labels, $\mathbf{n} = [n_1, n_2, \ldots, n_C]$, can be modelled using a multinomial distribution, $Multinomial(N, \mathbf{p})$, with parameters N and $\mathbf{p} = [p_1, p_2, \ldots, p_C]$.

Given a dataset, $\{\mathbf{x}_i^c \mid i = 1, \ldots, n_c, c = 1, \ldots, C\}$, we take N as a known parameter: 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: $\hat{p}_c = \frac{n_c}{N}$. We denote the estimation of $\mathbf{p}$ as $\hat{\mathbf{p}} = [\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_C]$.

For exactly balanced data, $p_c = \frac{1}{C}$ $(c = 1, 2, \ldots, C)$. We use $\mathbf{b} = [\frac{1}{C}, \frac{1}{C}, \ldots, \frac{1}{C}]$ to denote the class distribution vector for exactly balanced data. For imbalanced data, the class with $\hat{p}_c \geq \frac{1}{C}$ is defined as the majority class while that with $\hat{p}_c < \frac{1}{C}$ is defined as the minority class. Therefore, 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 $\hat{\mathbf{p}}$:

$$\mathrm{IR} = \frac{\hat{p}_{\max}}{\hat{p}_{\min}}, \tag{2}$$

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 $\mathbf{b}$:

$$\mathrm{ID} = \frac{d(\hat{\mathbf{p}}, \mathbf{b})}{d(\mathbf{p}_m, \mathbf{b})} + (m - 1), \tag{3}$$

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 $d(\mathbf{p}_m, \mathbf{b})$

is the distance between $\mathbf{b}$ and a class distribution vector with m zeros, $(C-m-1)$ $\frac{1}{C}$s and one $1-\frac{C-m-1}{C}$:

$$[\underbrace{0,\ldots,0}_{m},\underbrace{\frac{1}{C},\ldots,\frac{1}{C}}_{C-m-1},\underbrace{1-\frac{C-m-1}{C}}_{1}].$$

The first term in (3) 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 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 testing 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 Section 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 more problems in measuring class-imbalance extent. 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) $\hat{\mathbf{p}}_2 = [\frac{1}{3.1}, \frac{1}{3.1}, 1-\frac{2}{3.1}]$. $\mathrm{ID}(\hat{\mathbf{p}}_1) = c_1 + 0 \in [0,1]$ and $\mathrm{ID}(\hat{\mathbf{p}}_2) = c_2 + 1 \in [1,2]$, 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 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 $\mathbf{n} = [n_1, n_2, \ldots, n_C]$, 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

$$-2\ln\frac{L(\mathbf{b}|\mathbf{n})}{L(\hat{\mathbf{p}}|\mathbf{n})}, \tag{4}$$

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 in (4) 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

$$\mathrm{LRID} = -2\sum_{c=1}^{C} n_c \ln\frac{b_c}{\hat{p}_c} = -2\sum_{c=1}^{C} n_c \ln\frac{N}{Cn_c}. \tag{5}$$

3. Experiments

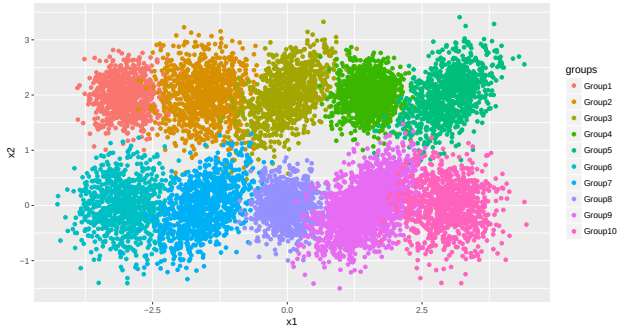
In the following experiments, we compare three imbalance degree metrics, IR, ID and LRID, on both simulated and real datasets. In ID, as suggested by the experiment results in Ortigosa-Hernández et al. (2017), the total variation distance and the Hellinger distance have the best performance and we test 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) is adopted as the classification algorithm.

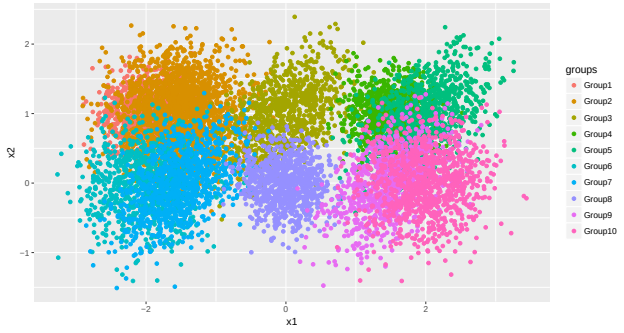
3.1. Simulated data

3.1.1. Experiment settings for simulated data

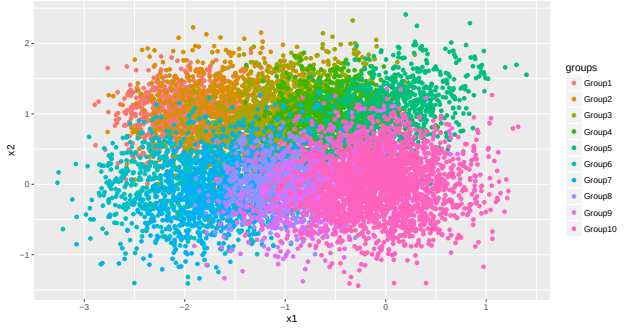
Here we design experiments to compare the performances of the class-imbalance metrics for data with different class-separation. Prati et al. (2004) show that the classification per-



(a) Well-separated



(b) Overlapped



(c) Extremely overlapped

Fig. 1: Balanced datasets for the three separation degrees.

formance of imbalanced data can also be affected by the intrinsic properties of data. Since the correltion between the classification performance and the class-imbalance metrics 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 a 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 $\boldsymbol{b} = [\frac{1}{10}, \frac{1}{10}, \ldots, \frac{1}{10}]$. 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 $\boldsymbol{p} = [\underbrace{p_{min}, \ldots, p_{min}}_{m}, \underbrace{p_{maj} \ldots, p_{maj}}_{K-m}]$, 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 number 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) data that are well separated with the means more 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) data that are overlapped 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) data that are extremely overlapped 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: $\left(\begin{smallmatrix} 0.1 & 0 \\ 0 & 0.1 \end{smallmatrix}\right)$, $\left(\begin{smallmatrix} 0.2 & 0 \\ 0 & 0.2 \end{smallmatrix}\right)$ and $\left(\begin{smallmatrix} 0.2 & 0.1 \\ 0.1 & 0.2 \end{smallmatrix}\right)$. The balanced data for the three sets of means are shown in Fig. 1. The SIs of the fully balanced well-separated data, overlapping data and extremely overlapping 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 vaules of r) and each dataset has 10000 observations and 10 classes. To make the comparison between different separation 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 to each dataset and use the F1 score as the metrics to assess the classification performances. We perform 20 random training/test splits on each dataset, with 70% training data and 30% test data.

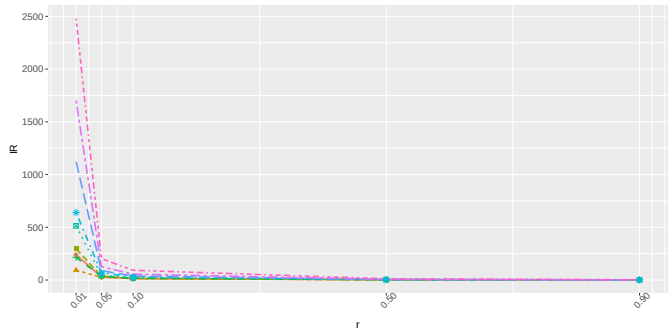
3.1.2. Results of simulated data

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 metrics 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 extent measured by r are shown in Fig. 2.

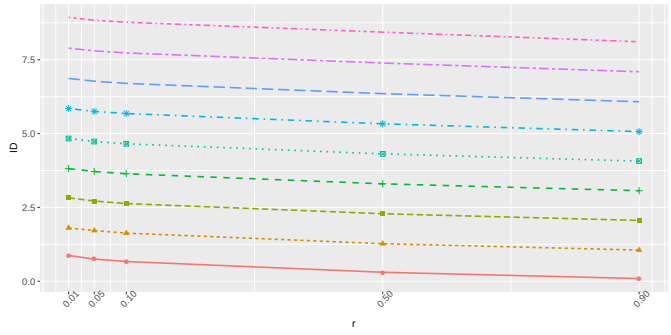
For each plot in Fig. 2, the horizontal axis shows values of r and the vertical axis shows the values of the metrics. Each line in the plot corresponds to a specific value of m.

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.

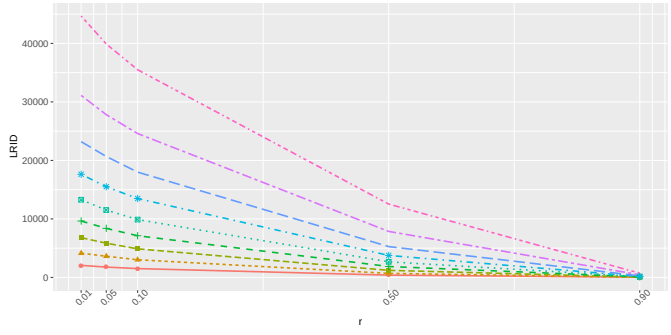
However, resolution is not the only criterion to assess the quality of the imbalance-degree metric. Although ID has the



(a) IR

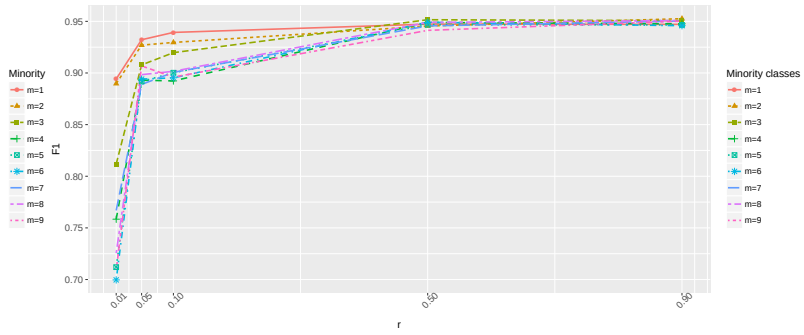


(b) ID

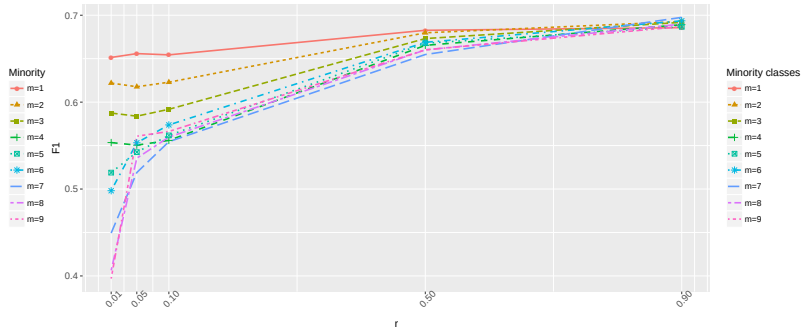


(c) LRID

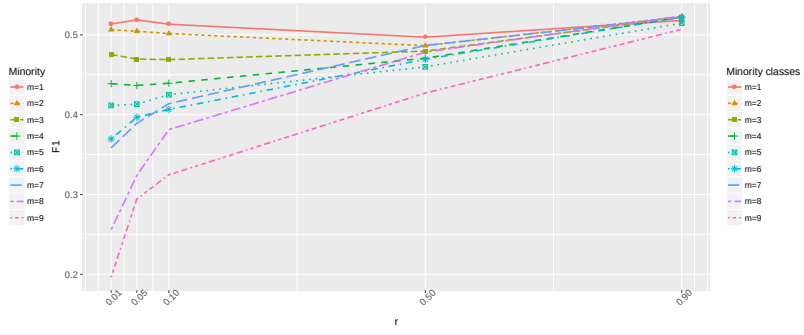
Fig. 2: The values of three class-imbalance metrics against different numbers of minority classes m and different imbalance extent measured by r.



(a) Well-separated



(b) Overlapped



(c) Extremely-overlapped

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

highest resolution, it is not reasonable to have very different values for different m when $r = 0.9$. This is because the datasets 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 extent. IR and LRID that have similar values for different m are more resonable 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) between the F1 scores and the metrics are shown in Table 1. We can make the following observations from Table 1.

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. However, as the data become more overlapped, LRID becomes the best metric in terms of both SRCC and PCC.

Second, it is obvious that ID_{TV} and ID_{HE} have worse performances than ID^*_{TV} and ID^*_{HE}, which 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.

To make further investigation for the above observations, we plot the F1 scores against the values of r and m in Fig. 3. For 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 almost

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

Data	Criterion	IR	ID_{TV}	ID_{HE}	ID^*_{TV}	ID^*_{HE}	LRID
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	-0.510	-0.50	-0.7	-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**

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)

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 $\hat{p}_{max}$ and $\hat{p}_{min}$ but different $\hat{p}_c$'s in between to have the same imbalance extent 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 more obvious, 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 rank of resolution of the three metrics is IR < LRID < ID, as shown in Fig. 2. Second, the separation of the data affect 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} is 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 resonable 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

Ten UCI datasets are used in the experiments (Dheeru and Karra Taniskidou, 2017): yeast, ecoli, wine, abalone, auto mpg, glass, Hayes-Roth, pageblocks, penbased and shuttle. The descriptions of the ten datasets are shown in Table 2. Similarly to the simulated data, LDA is applied to all datasets. We perform 20 random training/test split on each dataset, with 70% training data and 30% test data. The mean of the F1 scores is recorded for each dataset. PCC and SRCC between the imbalance extent metrics and the F1 scores are calculated.

3.2.1. Results of real data

The correlations for the real datasets are shown in Table 3. It is obvious that ID^*_{TV} and LRID can achieve the best SRCC and PCC, while IR cannot provide good correlations with classification performances for the real datasets. This result is supported by the SIs of the datasets in Table 2. Except for the last three datasets, other seven 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.

Similarly to those of simulated data, the results of real data also suggest that the distance metric can have great effect on the performance of ID^*. In addition, adding $(m-1)$ can also have harmful effect on the performance of ID. In contrast, the new LRID can provide competitive performance with the best 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.

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

	IR	**ID_{TV}**	**ID_{HE}**	**ID^*_{TV}**	**ID^*_{HE}**	**LRID**
SRCC	-0.41	-0.08	-0.48	**-0.72**	-0.22	**-0.72**
PCC	-0.30	0.13	-0.32	**-0.54**	0.13	**-0.54**

4. Conclusion

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

Acknowledgements

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 .

Castro, C.L., Braga, A.P., 2013. Novel cost-sensitive approach to improve the multilayer perceptron performance on imbalanced data. IEEE transactions on neural networks and learning systems 24, 888–899.

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.

Ha, J., Lee, J.S., 2016. A new under-sampling method using genetic algorithm for imbalanced data classification, in: Proceedings of the 10th International Conference on Ubiquitous Information Management and Communication, ACM. p. 95.

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 .

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

Nekooeimehr, I., Lai-Yuen, S.K., 2016. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. Expert Systems with Applications 46, 405–416.

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.