Bayes Imbalance Impact Index: A Measure of Class Imbalanced Data Set for Classification Problem

Yang Lu, *Member, IEEE*, Yiu-Ming Cheung, *Fellow, IEEE*, and Yuan Yan Tang, *Life Fellow, IEEE*

***Abstract*— Recent studies of imbalanced data classification have shown that the imbalance ratio (IR) is not the only cause of performance loss in a classifier, as other data factors, such as small disjuncts, noise, and overlapping, can also make the problem difficult. The relationship between the IR and other data factors has been demonstrated, but to the best of our knowledge, there is no measurement of the extent to which class imbalance influences the classification performance of imbalanced data. In addition, it is also unknown which data factor serves as the main barrier for classification in a data set. In this article, we focus on the Bayes optimal classifier and examine the influence of class imbalance from a theoretical perspective. We propose an instance measure called the Individual Bayes Imbalance Impact Index (IBI^3) and a data measure called the Bayes Imbalance Impact Index (BI^3). IBI^3 and BI^3 reflect the extent of influence using only the imbalance factor, in terms of each minority class sample and the whole data set, respectively. Therefore, IBI^3 can be used as an instance complexity measure of imbalance and BI^3 as a criterion to demonstrate the degree to which imbalance deteriorates the classification of a data set. We can, therefore, use BI^3 to access whether it is worth using imbalance recovery methods, such as sampling or cost-sensitive methods, to recover the performance loss of a classifier. The experiments show that IBI^3 is highly consistent with the increase of the prediction score obtained by the imbalance recovery methods and that BI^3 is highly consistent with the improvement in the F1 score obtained by the imbalance recovery methods on both synthetic and real benchmark data sets.**

***Index Terms*— Bayes classifier, class imbalance learning, data complexity, imbalance measure, imbalance recovery methods.**

Manuscript received January 24, 2019; revised July 3, 2019; accepted September 26, 2019. Date of publication November 1, 2019; date of current version September 1, 2020. This work was supported in part by the National Natural Science Foundation of China under Grant 61672444 and Grant 61272366, in part by Hong Kong Baptist University (HKBU), Research Committee, Initiation Grant—Faculty Niche Research Areas (IG-FNRA) 2018/19, under Grant RC-FNRA-IG/18-19/SCI/03, in part by the Innovation and Technology Fund of Innovation and Technology Commission of the Government of the Hong Kong SAR under Project ITS/339/18, in part by the Faculty Research Grant of HKBU under Project FRG2/17-18/082, and in part by the SZSTI under Grant JCYJ20160531194006833. *(Corresponding author: Yiu-Ming Cheung.)*

Y. Lu is with the Fujian Key Laboratory of Sensing and Computing for Smart City, School of Informatics, Xiamen University, Xiamen 361005, China, and also with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: lylylytc@gmail.com).

Y.-M. Cheung is with the Department of Computer Science, Hong Kong Baptist University, Hong Kong (e-mail: ymc@comp.hkbu.edu.hk).

Y. Y. Tang is with the Faculty of Science and Technology, UOW College Hong Kong/Community College of City University, Hong Kong, and also with the Department of Computer and Information Science, Faculty of Science and Technology, University of Macau, Macau, China (e-mail: yytang@cityu.edu.hk).

Color versions of one or more of the figures in this article are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TNNLS.2019.2944962

I. Introduction

THE classification of the binary imbalanced data is a challenging problem in the field of machine learning [1]. The classification accuracy deteriorates when the number of samples in one class overwhelms another class. Neglecting all the minority class samples has little effect on the overall accuracy because the minority class only takes only a small percentage. This problem usually occurs in detection tasks, such as cancerous diagnosis [2], insider threats [3], and prediction of software defects [4], where the recognition target is the minority class, which draws more interests in the application domain even though it has a relatively small number of samples. Various imbalance recovery methods have recently been proposed with the objective of improving the accuracy of the minority class without heavily sacrificing that of the majority class. A comprehensive review of these imbalance recovery methods is given in [5] and [6]. These methods attempt to recover the performance loss caused by imbalance via preprocessing the training data or modifying the decision-making procedure of an algorithm so that the minority class receives the same importance as the majority class during modeling and prediction.

However, before applying the imbalance recovery methods on an imbalanced data set, we should first address the questions of whether the so-called "imbalanced" issue should be considered in an imbalanced data set and whether the imbalanced recovery method should be used. To do so, we should first define the specific meaning of an imbalanced data set because perfectly balanced data sets are very rare in practice. The imbalance ratio (IR), which is the ratio between the number of the majority class samples and the minority class samples, is typically used to reflect the classification difficulty caused by class imbalance [7], and the assumption is that the higher the IR, the more difficult it is to predict the minority class samples. However, recent empirical studies have shown that the IR is not the main determinant in class imbalance learning problem [8]. A higher IR will only further deteriorate the classification accuracy if the other data complexities have also influenced the classification result. For example, Fig. 1 shows three imbalanced data sets with the same IR. The imbalance recovery methods provide different levels of accuracy improvement on the minority class in these data sets. The two classes of the data set shown in Fig. 1(a) are completely separated, so regardless of the severity of the imbalance, all of the samples will be correctly classified. Conversely, the two classes of the data set in Fig. 1(b) completely and uniformly

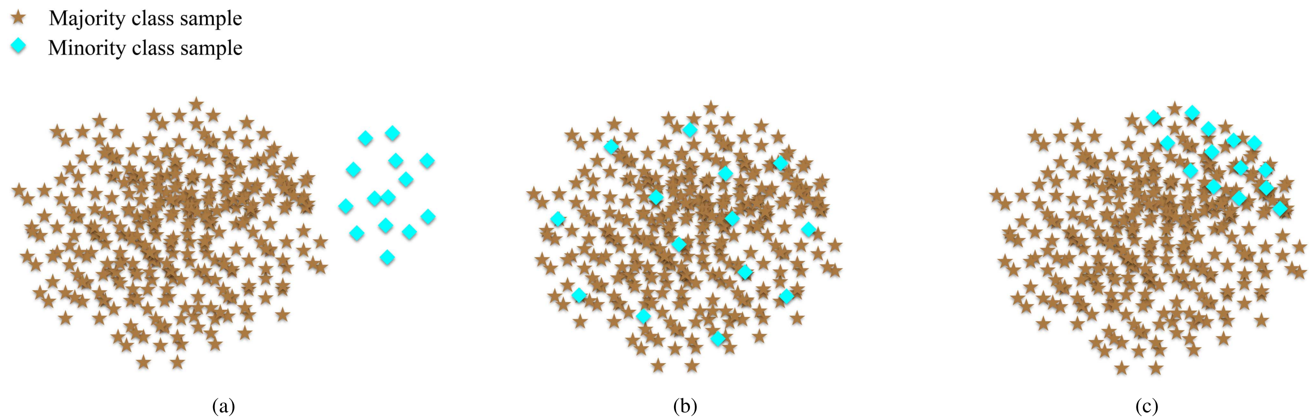

Fig. 1. Three imbalanced data sets with the same number of majority and the minority class samples. The minority class and the majority class are (a) separable, (b) totally overlapped, and (c) partially overlapped.

overlap. Even when imbalance recovery methods are applied, the best result is that a maximum of half of the minority class samples can be recovered, at the cost of reducing the accuracy of half of the majority class samples. For the case in Fig. 1(c), the minority class partially overlaps with the majority class. If imbalance recovery methods are applied, most of the minority class samples can be correctly classified with only a small loss in the accuracy of the majority class. In summary, if we only use IR to measure the difficulty of an imbalanced data set, all three data sets in Fig. 1 will be deemed to have the same difficulty for classification. The imbalance recovery methods cannot improve the classification of the data sets in Fig. 1(a), and the extent of improvement also differs for the data sets in Fig. 1(b) and (c). Therefore, if a data set cannot be improved by any imbalance recovery method, it is not necessary to consider the imbalance issue for this data set. Sometimes, the imbalance recovery methods may both increase the computational burden and deteriorate the performance if the cost of improving the minority class accuracy is to sacrifice more majority class accuracy.

It is also worth noting that IR is not the only factor that jeopardizes the classification accuracy [9], [19], as the poor result can also be generated from both low IR and high IRs. Three other data factors should be considered as well when dealing with the imbalanced data set. Basically, there are three data factors that are typically related to the class imbalance problem and should, therefore, also be considered when working with an imbalanced data set [8].

1) *Small Disjuncts:* When the data in the same class are represented by different clusters, the underrepresented small cluster will further hamper the classification if an imbalance exists in the data set.
2) *Noise:* The existence of noises in either the majority class or the minority class will bring further difficulty, particularly for the sampling-based imbalance recovery methods [10].
3) *Overlapping:* The degree of overlapping significantly affects the accuracy of the minority class because if the minority class samples in the overlapping region are sacrificed, greater overall accuracy is usually obtained.

Most studies have used experimental methods to empirically analyze the relationship between the three data factors and imbalance, and as far as we are aware, no theoretical analysis of this relationship has been conducted. The only conclusion that has been drawn is that if other data factors, such as overlapping, small disjuncts, and noise, are present to the same degree, a higher IR may lead to a further deterioration in performance [9], [19]. However, the data factors will differ in different data sets, and thus, using IR alone to represent the difficulty of the imbalanced data set will be insufficient and inaccurate. Thus, given an imbalanced data set with low performance, one has no idea whether the performance loss is due to the imbalance or to other factors. To determine the extent of the effect of imbalance, we propose two measures with which we can isolate other data factors and address the research problem of the impact of imbalance. We refer to these as the Individual Bayes Imbalance Impact Index (IBI^3) and the Bayes Imbalance Impact Index (BI^3) that estimate the degree of deterioration caused solely by imbalance at the instance level and the data level, respectively. IBI^3 is calculated by quantizing the difference in the prediction score of a given minority class sample between the imbalanced and balanced situations. BI^3 is the averaged IBI^3 over all minority class samples and can, therefore, be used to describe the effect of imbalance on the data set. For the previous example, the data set in Fig. 1(a) will have a very small BI^3 value and that in Fig. 1(c) will have a larger BI^3 value than that in Fig. 1(b). Therefore, BI^3 can be used as a judgment index, instead of purely referring to IR, to determine whether we must consider the imbalance issue and whether imbalance recovery methods should be applied before training on the data set. That is, BI^3 has positive correlation with the benefit of applying imbalance recovery methods. The higher BI^3 is, the more the performance can be improved via imbalance recovery methods. We experimentally verify the effectiveness of IBI^3 and BI^3 by correlation analysis with the different standard classifiers and imbalance recovery methods. The experimental results show that IBI^3 is highly correlated with the increase of the prediction score on minority class samples, and BI^3 is highly correlated with the improvement in the F1 score for the whole data in both synthetic and real benchmark data sets. Therefore, BI^3 is a suitable measure to describe how the data are influenced by imbalance. Our study makes the following contributions.

1) It is the first attempt to examine the data factors of an imbalanced data set from a theoretical perspective.
2) The proposed IBI^3 is the first instance complexity measure to show how a minority class sample is influenced by imbalance.
3) The proposed BI^3 can be used as a data complexity measure to describe the imbalance degree, instead of referring only to IR.
4) The influence of the imbalance can be estimated without training and testing, so it can then be determined whether a specific imbalance recovery method should be applied.

The remainder of this article is organized as follows. Section II lists the work related to the class imbalance problem, and the data factors related to the imbalance problem are discussed. Section III describes the proposed method. Section IV presents the experiments and discussions. Finally, our concluding remarks are given in Section V.

II. Related Work

Most studies of class imbalance learning propose imbalance recovery methods, which can be basically categorized into three groups [11]. First, methods on a data level aim to manipulate the data to be balanced before training. The best-known method in this group is the Synthetic Minority Oversampling TEchnique (SMOTE) [12]. It synthesizes new samples into the minority class by interpolating the existing minority class samples with their neighbors. In addition to data synthesis, data cleaning techniques have also been used in data preprocessing. For example, Batista *et al.* [13] used the Tomek links to clean the overlapping area between classes to clarify the classification boundary after the introduction of synthetic samples. The second group of methods at the algorithm level modifies other learning methods by adapting them to the imbalanced data. The modified algorithm typically shifts the decision boundary to enhance the existence of the minority class samples. For example, Hong *et al.* [14] modified the kernel classifiers by orthogonal forward selection to optimize the model generalization for imbalanced data sets. The third group is related to the framework of cost-sensitive learning [15]. These methods assign different costs to the samples of different classes. The minority class samples are usually assigned a large cost to prevent them from easily being misclassified. The idea of cost sensitivity can also be applied to many other algorithms to turn them into imbalance recovery methods, such as decision tree [16] and SVM [17].

The imbalance recovery methods mentioned earlier assume that performance deteriorates because of class imbalance, but recent studies have shown that the imbalance is not the only cause of the performance deterioration [8], [10], [18]. At least, three other factors can render predictions inaccurate on imbalanced data sets. First, in a sparse minority class, the samples are separated into small clusters. This problem is called small disjuncts or within-class imbalance [5], which has commonly been studied together with the imbalance problem. Therefore, Japkowicz [19] generated synthetic data to study the relationships among the class disjuncts, the size of the training data, and the IR. The results show that the small disjuncts are more responsible for the decrease in accuracy than the IR by changing the degrees of these data factors. Accordingly, a solution dealing with small disjuncts called CBO has been proposed in [9]. It first conducts clustering on each class so that the oversampling is conducted on each disjunct instead of each class. In addition, Prati *et al.* [20] studied the performance of unpruned trees by considering the relationship between class imbalance and small disjuncts and proposed to use SMOTE with data cleaning methods to alleviate the performance loss from the small disjuncts.

The second data factor is noise. Noisy samples are typically defined as those from one class located deep inside another class [21]. The existence of noise samples in the minority class will make blind oversampling methods, such as SMOTE, generate more noises, so the application of oversampling on the noisy minority class may even degrade the performance further [10]. Therefore, data cleaning methods are typically used to tackle noise, such as the Tomek link [13] and ENN [22]. Collecting samples that are incorrectly classified by the kNN classifier [23] is another straightforward method of finding the noise. Van Hulse and Khoshgoftaar experimented using data with artificial noises [7] in which the class noise was injected into real data sets by randomly relabeling the samples before training. The results of all the compared classifiers showed that the minority class was severely affected by noises.

Finally, overlapping between the classes can affect classification, particularly when the data are imbalanced. Napierala and Stefanowski [18] proposed a kNN-based method to categorize the minority class examples into the four categories of safe, border, rare, and outlier. The categories depend on the ratio of the majority class samples in the k nearest neighbors (kNN) of each minority class sample. For each data set, the degree of overlap of the minority class can be obtained by investigating the proportions of the four groups. However, the analysis only shows the difficulty of classifying the minority class samples, and the degree of imbalance is not considered. García *et al*. [24] evaluated kNN when the local IR was inverse to the global IR and concluded that kNN is more dependent on the local imbalance. Anwar *et al*. [25] also proposed the use of kNN to measure the data complexity for imbalanced data with adaptively selected k. Prati *et al*. [26] observed that the performance loss is related not only to class imbalance but also to the degree of overlapping. In summary, the previous studies empirically justified their conjectures without any theoretical frameworks and no measure has as yet been proposed to assess how the data set is influenced by class imbalance, independent of other data factors.

Finally, the studies of data complexity should be considered as a related area. A list of complexity measures was proposed in [27] with different featured groups. The measures are used to study the essential structure of data and guide the selection of classifiers for specific problems. Recently, Smith *et al.* [28] have extended the study of data complexity from data level to the instance level. They proposed a group of complexity measures that can be calculated for each instance, and the correlations among them are then analyzed. These instance-level complexity measures can be used for data cleaning to filter the most difficult samples in the data. However, no specific research into the data complexity for imbalanced data has been conducted, and the existing complexity measures are not suitable to assess the influence of imbalance on the data.

III. Proposed Method

A straightforward method of establishing the influence of imbalance on a data set is to compare the model learned from the imbalanced data with that learned from its balanced case, in which the number of minority class samples equals that of the majority class and are drawn from the underlying distribution. If the distribution is known, the differences between the models built on the imbalanced data and on the balanced data will be clear because other data factors will be fixed. However, the distribution is usually unknown in practice. We can only estimate the distribution by the observed minority class samples in the data set. Thus, we propose to use the Bayes optimal classifier to estimate the difference because it has the theoretical minimum classification error and takes the class prior into account. Based on the Bayes decision theory, the difference in the theoretical classification error between the classifiers trained on the imbalanced and balanced data sets can be estimated. Thus, the impact of imbalance can be estimated while isolating other data factors that may influence the classification. First, we decompose the problem into the instance level and propose the IBI^3, which measures how each minority class sample is influenced by a class imbalance in classification. We then define the data level measure as the BI^3 by averaging IBI^3 over all minority class samples. BI^3, thus, represents the impact of imbalance on the whole data set.

A. Derivation in Normal Distribution

The details of the proposed measures are described as follows. The Bayes rule denotes that the posterior probability of a given sample $\mathbf{x}$ in class c is

$$p(y=c|\mathbf{x}) = \frac{p(\mathbf{x}|y=c)p(y=c)}{p(\mathbf{x})}. \tag{1}$$

The decision of the optimal Bayes classifier for the binary classification problem is as follows:

$$f(\mathbf{x}) = \arg\max_{c=\{+1,-1\}} p(y=c|\mathbf{x}). \tag{2}$$

$p(\mathbf{x})$ is the same for both classes, and in practice the prior probability is usually estimated by the frequency of each class. The decision can then be formulated as:

$$f(\mathbf{x}) = \begin{cases} +1, & f_p(\mathbf{x}) > f_n(\mathbf{x}) \\ -1, & \text{otherwise} \end{cases} \tag{3}$$

where

$$f_p(\mathbf{x}) = N_p\, p(\mathbf{x}|+) \tag{4}$$

$$f_n(\mathbf{x}) = N_n\, p(\mathbf{x}|-) \tag{5}$$

N_p and N_n are the numbers of samples in the positive class and negative classes, respectively, and $f_p(\mathbf{x})$ and $f_n(\mathbf{x})$ are the posterior scores, which are proportional to the posterior probabilities. $y=+1$ and $y=-1$ are simplified as $+$ and $-$ in the conditional probability. The majority class is typically denoted as negative and the minority class as positive. When the class is imbalanced, namely, $N_p \ll N_n$, the Bayes optimal decision may be dominated by the frequency so that some or even all minority class samples may be misclassified. The optimal Bayes error is the sum of all misclassified samples regardless of the class, so under the imbalance circumstance, sacrificing the accuracy of the minority class samples helps minimize the total error. However, in most of the imbalanced data applications, a low error rate does not represent good performance. To account for the importance of the minority class, measurements such as the F1 score, G-mean, and AUC are commonly used instead of the error rate [5]. Thus, an alternative decision function that is not influenced by the prior probability can be written as

$$f'(\mathbf{x}) = \begin{cases} +1, & f'_p(\mathbf{x}) > f_n(\mathbf{x}) \\ -1, & \text{otherwise} \end{cases} \tag{6}$$

where

$$f'_p(\mathbf{x}) = N_n\, p(\mathbf{x}|+). \tag{7}$$

The decision function $f'(\mathbf{x})$ directly compares the value between $p(\mathbf{x}|+)$ and $p(\mathbf{x}|-)$. This is, in fact, the decision function with minimal Bayes error when the classes are balanced. The influence of imbalance on the data set can be reflected by the difference between f'_p and f_p, where f_p is proportional to the minority class posterior probability under the real imbalanced case and f'_p is estimated under the balanced case. However, direct comparison of f_p and f'_p is meaningless because the decision hyperplane is also determined by f_n. Therefore, we define IBI^3 as the difference between the normalized posterior probabilities of the imbalanced case and the estimated balanced case

$$\text{IBI}^3(\mathbf{x}) = p(+|\mathbf{x}, f') - p(+|\mathbf{x}, f) \tag{8}$$

$$= \frac{f'_p(\mathbf{x})}{f_n(\mathbf{x}) + f'_p(\mathbf{x})} - \frac{f_p(\mathbf{x})}{f_n(\mathbf{x}) + f_p(\mathbf{x})}. \tag{9}$$

Fig. 2(a) shows an example of the distribution of $f_n(\mathbf{x})$, $f_p(\mathbf{x})$, and $f'_p(\mathbf{x})$ on a 1-D normally distributed binary class data with IR = 5. Fig. 2(b) shows the normalized posterior probabilities and IBI^3. The peak of IBI^3 is observed in the region between two decision hyperplanes $f(\mathbf{x})$ and $f'(\mathbf{x})$, which means that the part with the most difference between the imbalanced and balanced cases lies in the region between two hyperplanes. The minority class samples in this region are misclassified under the imbalanced case but correctly classified under the balanced case, which can be regarded as the impact on the minority class sample solely from the imbalance. If IBI^3 is low, the minority class sample $\mathbf{x}$ is either a noise sample, which is deeply located in the region of the majority class that makes both $p(+|\mathbf{x}, f')$ and $p(+|\mathbf{x}, f)$ close to 0, or a safe sample that is deeply located in the region of the minority class that makes both $p(+|\mathbf{x}, f')$ and $p(+|\mathbf{x}, f)$ close to 1. In both cases, IBI^3 is small, and the influence of the imbalance on $\mathbf{x}$ is insignificant. Thus, even if imbalance recovery methods

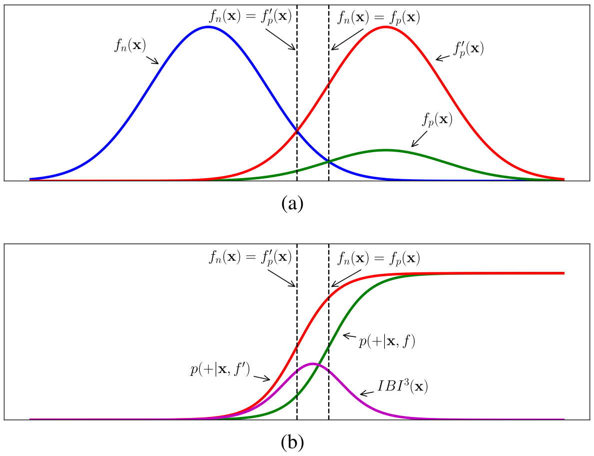

Fig. 2. Example to show the distribution of IBI^3 on two classes with normal distributions. (a) Posterior scores. (b) Normalized posterior probabilities and IBI^3. The optimal Bayes decision hyperplanes $f'(\mathbf{x})$ and $f(\mathbf{x})$ are shown by the dotted lines.

are applied, the classification results for these minority class samples with low IBI^3 values are not likely to change.

IBI^3 is calculated for each minority class sample, and the averaged IBI^3 over all the minority class can be used to describe the imbalance impact of the data set. BI^3 for the whole data set $\mathcal{D}$ is calculated by averaging over all IBI^3 on the minority class

$$\mathrm{BI}^3(\mathcal{D}) = \frac{1}{N_p} \sum_{\substack{(\mathbf{x}_i, y_i) \in \mathcal{D}, \\ y_i = +1}} \mathrm{IBI}^3(\mathbf{x}_i). \tag{10}$$

B. Local Approximation

If the two classes are normally distributed, the likelihood functions $p(\mathbf{x}|+)$ and $p(\mathbf{x}|-)$ can be calculated by estimating the mean and variance. However, the assumption usually fails in real benchmark data sets because in addition to the distribution not being normal, small disjuncts and noises can be found among the classes. We can assume that the normality with estimated mean and variance may not be accurate enough to calculate IBI^3 and BI^3. Cover and Hart [29] showed the relationship between the error bounds of the nearest neighbor classifier and the Bayes classifier by the following theorem.

Theorem 1 (Cover and Hart, 1967): For a sufficiently large training set size N, the inequality of the error rate of the nearest neighbor classifier R_{NN} and the Bayes classifier R_{Bayes} holds

$$R_{\mathrm{Bayes}} \leq R_{\mathrm{NN}} \leq 2R_{\mathrm{Bayes}}(1 - R_{\mathrm{Bayes}}). \tag{11}$$

The upper bound of the error rate of the nearest neighbor classifier is found to be double that of the Bayes classifier, and the result is independent of the selection of the nearest neighbors k. Therefore, kNN is a good substitute to estimate the likelihood without a normality assumption. The details are given in Algorithm 1. For each minority class sample $\mathbf{x}$, we find its kNN $k\mathrm{NN}(\mathbf{x})$ and count the number of the majority class neighbors M. Thus, f_n is set at M/k, which

Algorithm 1 BI^3

Input: Dataset $\mathcal{D} = \{\mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}\}$, the number of positive samples N_p, the number of negative samples N_n, the number of nearest neighbors k_0.

1: $r \leftarrow N_n/N_p$;
2: Construct the sample set of the minority class $\mathcal{D}^+ \leftarrow \{\mathbf{x}_i^+\}$;
3: **for** $i \leftarrow 1$ to N_p **do**
4: Calculate the number of the minority class neighbors: $M \leftarrow |\{(\mathbf{x}', y') : \mathbf{x}' \in kNN(\mathbf{x}_i^+), y' = -1\}|$
5: **if** $M = k_0$ **then**
6: $M \leftarrow$ the number of the majority class samples between $\mathbf{x}_i^+$ and the nearest the minority class neighbor of $\mathbf{x}_i^+$;
7: $k \leftarrow M + 1$;
8: **else**
9: $k \leftarrow k_0$;
10: **end if**
11: $f_n \leftarrow M/k$;
12: $f_p \leftarrow (k - M)/k$;
13: $f'_p \leftarrow r(k - M)/k$;
14: Calculate IBI$^3(\mathbf{x}_i^+)$ by (9);
15: **end for**
16: Calculate BI^3 by (10);

Output: The indices IBI^3 and BI^3.

is the local probability that $\mathbf{x}$ is classified as negative, and f_p is correspondingly set at $(k - M)/k$. We assume that in the unknown balanced situation, there will be $r = N_n/N_p$ times more the minority class samples surrounded by $\mathbf{x}$. Therefore, f'_p is set at $r(k-M)/k$. To prevent the case in which all of the k neighbors of $\mathbf{x}$ are the majority class samples, which makes both f_p and f'_p equal to zero, we adopt a flexible k that is set at the minimal number to ensure that $\mathbf{x}$ has at least one the minority class neighbor. This is shown in Lines 5–10 in Algorithm 1.

An example with four binary class synthetic data sets drawn from a normal distribution with different IRs is given in Fig. 3. The IBI^3 values with $k_0 = 5$ can be visually compared in various locations of the minority class samples and with a different IR. Fig. 3 demonstrates that the minority class samples with high values of IBI^3 are mainly located in the boundary between two classes. This is consistent with the example shown in Fig. 2. The minority class samples that lie in the deep region of the majority class receives low IBI^3 because they are regarded as noises that will still be misclassified even if the two classes are balanced. Thus, their classification result is not significantly related to the imbalance. In addition, the minority class samples that are far from the majority class also receive low IBI^3 because they will be correctly classified regardless of whether the classes are imbalanced. Fig. 3 demonstrates that the IBI^3 values of the minority class samples on the boundary between two classes increase as IR increases. The influence of these minority class samples is, therefore, related to IR. The higher the IBI^3 value of a minority class sample is, the more seriously that the sample

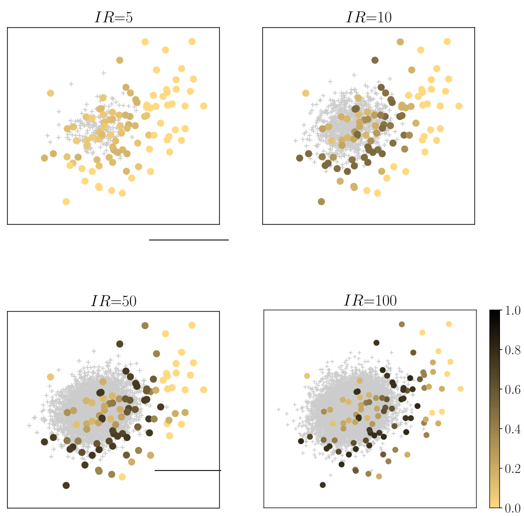

Fig. 3. Values of IBI^3 with local probability on a binary class synthetic data set drawn from a normal distribution with different IR s. The gray plus symbol is the majority class, and the colored dot is the minority class.

is influenced by imbalance and the higher the probability that the sample can be correctly classified in a balanced situation. The values of BI^3 for these four data sets are 0.0674, 0.2482, 0.3829, and 0.4588, respectively. The values of BI^3 increases as IR increases, which can be used to reflect the extent of the effect of imbalance on the data.

Remarks:

1) The minority class samples with high IBI^3 values are mainly located in the classification borderline, as shown in Fig. 3. The approach is similar to those of borderline-based methods, such as borderline-SMOTE [30], ADASYN [31], and the borderline minority class samples defined in [18]. These methods categorize the minority class samples by the percentage of majority class samples in their neighborhood. However, they do not distinguish data sets with different IRs. For example, if a minority class sample in a data set with IR $=$ 3 has one majority class sample among its five neighbors, it may not be treated as a borderline sample. However, if the same situation occurs for a minority class sample in a data set with IR $=$ 10, this sample should be treated as a borderline sample because a high IR indicates that the minority class sample may have more potential neighbors of its own class if the classes are balanced based on the underlying distribution. Therefore, the difference between IBI^3 and other borderline-based methods is that IBI^3 involves the factors of imbalance in defining the borderline minority class samples, whereas methods such as Borderline-SMOTE and ADASYN only consider the neighborhood of the minority class samples.
2) The proposed indices IBI^3 and BI^3 may not be suitable for estimation of the imbalance effect on high-dimensional imbalanced data directly, which would have intrinsic low-dimensional feature space. Under these circumstances, simply calculating the Euclidean distance by kNN on the original high-dimensional feature space would not be so appropriate to get an accurate estimations of IBI^3 and BI^3. Instead, it is still desirable that an appropriate dimension reduction technique should be applied before calculating IBI^3 and BI^3.

C. Guidance of Usage

In this section, we provide a guidance of how to use the proposed measures IBI^3 and BI^3 to deal with imbalanced data, while using IBI^3 or BI^3 to design a specific imbalance recovery algorithm is actually beyond the scope of this article and will be left for future studies.

The IBI^3 value can be used for differentiating the minority class samples for oversampling methods and cost-sensitive methods. IBI^3 indicates the impact caused by a minority class sample in terms of imbalance. Therefore, the oversampling weight can be determined by IBI^3 value. In other words, the minority class sample with a higher IBI^3 value will obtain higher probability to be oversampled. As discussed earlier, the minority class samples with low IBI^3 value are either noises or safe samples, whose classification results are likely to remain the same even when imbalance recovery methods are applied. Oversampling the minority class samples with low IBI^3 values may have limited benefit to the classification result.

The BI^3 value can be used for investigating an imbalanced data set before applying imbalance recovery methods. For researchers working on the area of imbalanced data classification, one can select data sets with high BI^3 values to conduct experiments for testing new imbalance recovery methods. Usually, researchers prefer to select imbalanced data sets by referring to IR. However, as discussed in Section I, high IR does not mean that applying imbalance recovery methods will recover more accuracy loss. Thus, the efficacy of the proposed imbalance recovery method may not be well evaluated and the experimental results may be misleading if IR is used to indicate the difficulty of imbalance. For engineers handling an imbalanced data set, one can calculate BI^3 value first to get a glimpse of the impact of imbalance on the data set. If the BI^3 value is very low (e.g., lower than 0.05 by a rule of thumb in Section IV-B), one should focus on other data clean methods instead of directly applying imbalance recovery methods.

IV. Experiments

The accuracy of the proposed measure BI^3 in the experiments is mainly evaluated by correlation analysis. We use Spearman's rank correlation coefficient [32], which is a nonparametric measure of the rank correlation between two variables that assess the degree of describing the relationship between two variables with a monotonic function. The correlation ranges from -1 to 1, where 1 or -1 indicates a perfect monotonously increasing or decreasing relationship and 0 indicates no correlation between two variables.

We use five well-known standard classifiers: RBF kernel support vector machine (SVM) [33], decision tree implemented by CART [34], kNN with $k = 5$ (5NN) [35], random forest (RF) [36], and AdaBoost [37]. We use the default parameter provided by *scikit-learn* learning library in Python [38]. The minimal number of nodes in each leaf of CART and RF is set at five to produce a probability output. We also use four imbalance recovery methods to deal with class imbalance: random oversampling (OS), random undersampling (US), SMOTE [12], and sample weighting (SW). The first three are sampling methods, and the last is a cost-sensitive method that assigns the weight of the minority class samples as the IR and the majority class sample as one. The above-mentioned methods are independent of the classifier, so they can be arbitrarily combined with standard classifiers to deal with class imbalance. We use the simplest imbalance recovery methods for the class imbalance problem because our intention is not to select the best imbalance recovery method but to show that the proposed measured index is generally consistent with the improvement made by the imbalance recovery methods. These methods are implemented by the *imbalanced-learn* toolbox [39].

The proposed measures are directly calculated on the whole data set so that each minority class sample is associated with an IBI^3 value and each data set is associated with a BI^3 value. To show the correlation with the standard classifiers using the imbalance recovery methods, we carry out tenfold cross validation with five different random partition runs for each combination of classifier and the imbalance recovery method. Thus, each minority class sample can be calculated as a test sample in its own fold and averaged by five runs. Since the proposed indices IBI^3 and BI^3 focus only on the minority class samples, we use the F1 score as the measurement because it is the harmonic mean of precision and recall on the minority class. The correlation analysis is conducted at two levels.

1) *Instance-Level Correlation:* All the minority class samples in all data sets are accumulated. We calculate the correlation between IBI^3 and the increase in the prediction score made by the imbalance recovery methods on each classifier by (8). Here, f' is the classifier with imbalance recovery methods, and f is the standard classifier. Thus, we can evaluate whether IBI^3 is consistent with the improvement made by the imbalance recovery method on minority class samples.
2) *Data-Level Correlation:* All the data sets are accumulated. We calculate the BI^3 on each data set and compare it with the improvement of the F1 score made by the imbalance recovery methods. Thus, we can evaluate whether BI^3 can show the impact of imbalance on the data set in terms of improvement in the F1 score.

The number of nearest neighbors k_0 is set at five for all experiments. No adequate comparison methods are available because this is the first study to propose a measure of the degree of impact on an imbalanced data set. Thus, we compare our results with three hardness measures: kDN and CL proposed in [28] and CM proposed in [25]. These are related to kNN and the Naive Bayes classifier but do not consider imbalance. kDN measures the percentage of data point $\mathbf{x}$'s neighbors that are not in the same class as $\mathbf{x}$

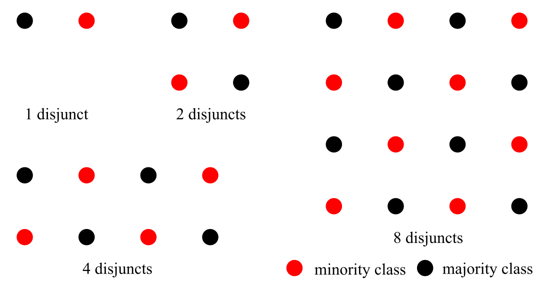

Fig. 4. Position of the majority class and the minority class with different number of disjuncts.

$$kDN(\mathbf{x}, y) = \frac{|\{(\mathbf{x}', y') : \mathbf{x}' \in k\mathrm{NN}(\mathbf{x}), y' \neq y\}|}{k} \tag{12}$$

where $k\mathrm{NN}(\mathbf{x})$ is the set of kNN of $\mathbf{x}$ and $|\cdot|$ is the size of the set. We also set $k = 5$. CL measures the global overlap between classes and the likelihood of a sample belonging to its opposite class

$$\mathrm{CL}(\mathbf{x}, y) = 1 - \prod_{i}^{d} p(\mathbf{x}_i, y) \tag{13}$$

where d is the number of dimensions and $p(\mathbf{x}_i, y)$ is the samples' likelihood on ith feature to its class y. It uses the same assumption in Naive Bayes, which is that the features are independent of each other. The original version of CL in [28] is the likelihood that a sample belongs to its own class. However, to be consistent with other methods in this article, in which the measurement is positively correlated with the instance hardness, we, therefore, use one to subtract the original CL. We average the values of kDN and CL on all minority class samples to obtain the data-level index. CM is a data-level complexity measure

$$CM(\mathbf{x}, y) = I\left(\frac{|\{(\mathbf{x}', y') : \mathbf{x}' \in k\mathrm{NN}(\mathbf{x}), y' = y\}|}{k} \leq 0.5\right) \tag{14}$$

$$CM(\mathcal{D}) = \frac{1}{N}\sum_{i=1}^{N} CM(\mathbf{x}_i, y_i) \tag{15}$$

where I is the indicator function. For the data-level correlation analysis, we also performed comparison with IR because it is usually regarded as an index for measuring the difficulty of an imbalanced data set. In summary, we compare IBI^3 with kDN and CL for instance-level correlation and compare BI^3 with kDN, CL, CM, and IR for data-level correlation.

A. Synthetic Data

We first evaluate the proposed index on synthetic binary class data sets. Three groups of synthetic data sets are generated.

1) *syn_Overlap:* The between-class distance and IR are adjusted.
2) *syn_Noise:* The noise level and IR are adjusted.
3) *syn_Disjunct:* The number of small disjuncts and IR are adjusted.

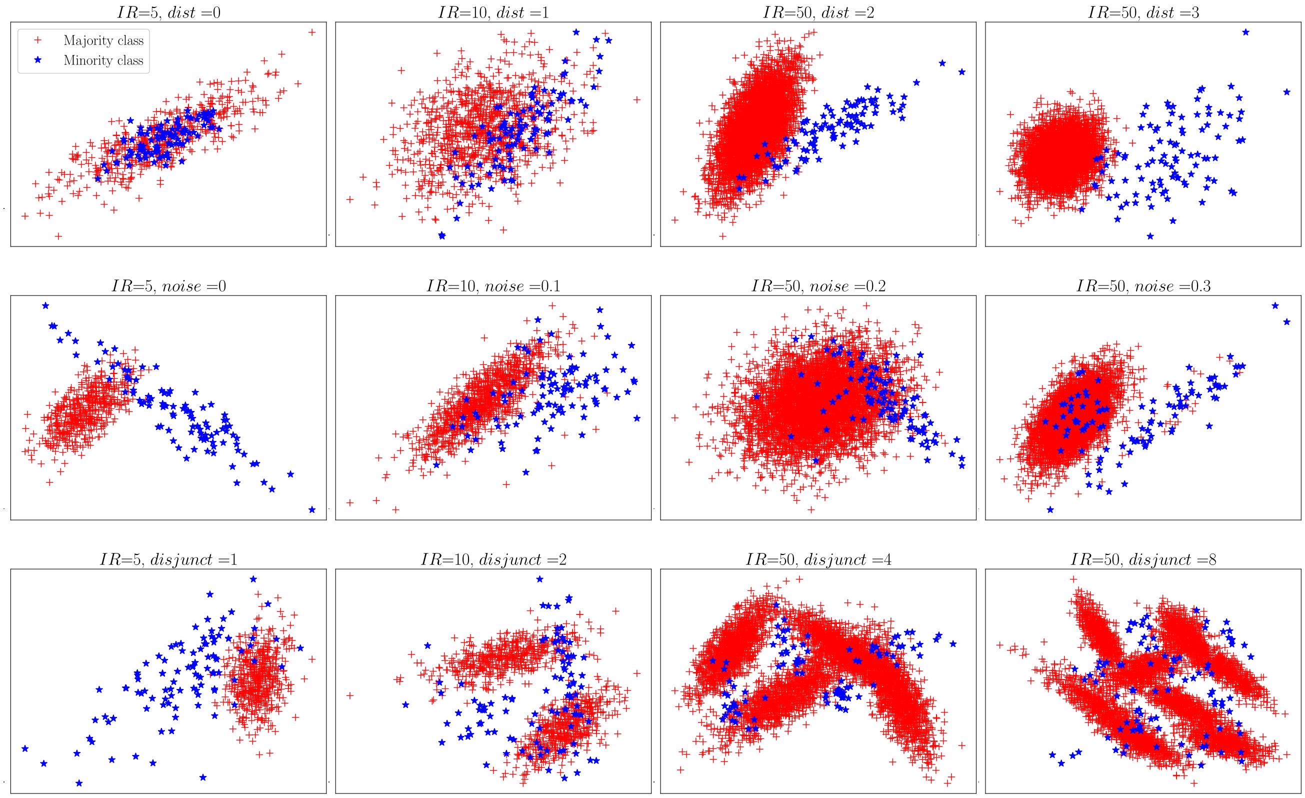

Fig. 5. Twelve synthetic binary class imbalanced data sets in data set group *syn_overlap* (top row), *syn_noise* (middle row), and *syn_disjunct* (bottom row) with different covariance combinations.

All data sets have two classes that are generated from a normal distribution with two dimensions. The number of samples in the minority class N_p is fixed at 100, and the number of samples in the majority class N_n varies in the set $\{500, 1000, 5000\}$, where IRs are 5, 10, and 50, respectively. For data set group *syn_overlap*, the distance between two classes *dist* varies in the set $\{0, 1, 2, 3\}$, and there is no noise. For data set group *syn_noise*, the noise level *noise* varies in the set $\{0, 0.1, 0.2, 0.3\}$, where 0.1 means that 10% of the minority class samples are labeled as the majority class and that the same number of the majority class samples are labeled as the minority class. The distance between the two classes for data set group *syn_noise* is fixed at two. For data set group *syn_disjunct*, the number of small disjuncts of each class *disjunct* varies in the set $\{1, 2, 4, 8\}$. For example, disjunct $= 2$ means that each class has two disjuncts. The distance between adjacent disjuncts is set at two. The position of the majority class and the minority class with the different numbers of disjuncts is shown in Fig. 4. For all synthetic data sets, the covariance matrix for each class is set to

$$\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{bmatrix} + 0.1I \tag{16}$$

where $\sigma_{11}, \sigma_{22} \in [0, 1]$ and $\sigma_{12}, \sigma_{21} \in [-1, 1]$ are uniformly random numbers. The extra term $0.1I$ ensures that the covariance matrix is positive semidefinite. The covariance matrix for the positive and negative classes is set differently, and the covariance matrix is drawn ten times to produce different combinations. Therefore, totally, there are three groups of $3 \times 4 \times 10 = 120$ data sets with various degrees of overlap, IRs, noise levels, number of disjuncts, and covariance. Four of the data sets in each data set group are shown in Fig. 5.

1) Results on Data Set Group syn_overlap: The instance-level correlation is shown in Table I. Generally, IBI^3 shows higher correlations than kDN and CL. IBI^3 shows the highest correlations on SVM with OS, US, and SMOTE, which are generally more than 0.85. A high correlation means that if the prediction score of a minority class sample can be increased by SVM with the imbalance recovery methods, its IBI^3 value is also high. Both IBI^3 and kDN use the nearest neighbors to calculate the measure. kDN has a much lower correlation than IBI^3 because the imbalance factor is not considered in kDN. The correlation of CART with OS is not high for all indices although IBI^3 achieves the highest at 0.1105, whereas the other two methods have negative correlations. Random oversampling may simply duplicate the minority class samples so that the leaf node of the decision tree is full of the duplicated the minority class samples after oversampling, which does not increase the prediction score of the minority class samples. In addition, CART with US has high correlation with IBI^3, which may suggest that US is a more effective way of increasing the minority class prediction score with CART. On 5NN, the correlations of IBI^3 of OS and SW are seen to be lower than those of US and SMOTE. OS and SW only work if the training the minority class samples are in the neighborhood of the testing minority class sample. If the testing minority class sample is surrounded by the training majority class

TABLE I

INSTANCE-LEVEL SPEARMAN RANKED CORRELATION BETWEEN THE INDICES AND THE INCREASE OF PREDICTION SCORE OF MINORITY CLASS SAMPLES ON DATA SETS GROUP *syn_overlap*

		OS	US	SMOTE	SW
kDN	SVM	0.7627	0.7840	0.7506	0.5285
	CART	-0.0061	0.7379	0.4182	0.2091
	5NN	0.2200	0.8485	0.5801	0.2925
	RF	0.0971	0.7846	0.4572	0.3515
	AdaBoost	0.2158	-0.2363	0.2187	0.2156
CL	SVM	0.6016	0.6031	0.5939	0.4431
	CART	-0.0576	0.5578	0.3964	0.2188
	5NN	0.2453	0.5930	0.4695	0.2803
	RF	0.2002	0.6312	0.4784	0.3738
	AdaBoost	0.1314	-0.2348	0.1696	0.1267
IBI^3	SVM	**0.8501**	**0.8512**	**0.8416**	**0.5977**
	CART	**0.1105**	**0.8072**	**0.5881**	**0.3522**
	5NN	**0.4995**	**0.9311**	**0.7997**	**0.5965**
	RF	**0.3215**	**0.8531**	**0.6769**	**0.5487**
	AdaBoost	**0.2841**	**-0.0944**	**0.2664**	**0.2815**

samples, it will still be misclassified because OS and SW only duplicate and increase the weight of the training minority class samples. For RF, the correlation of IBI^3 is higher than CART because the ensemble of trees is more robust and will increase the prediction score, particularly for US, which shows a correlation of 0.8531 correlation with IBI^3. For AdaBoost, the correlation is low for all indices with all imbalance recovery methods. Our investigation found that the minority class prediction score of AdaBoost is very close to 0.5 and that the imbalance recovery methods only increase the score a little, to make it just over 0.5, which would change the classification result. Therefore, AdaBoost has a low correlation with the indices.

The data-level correlation is shown in Table II. BI^3 shows the highest correlation with the improvement in the F1 score for all classifiers and all imbalance recovery methods, where the correlations are generally greater than 0.5. For SVM, BI^3 shows high correlations with all imbalance recovery methods. All the correlations are greater than 0.77. CART, 5NN, and RF also show higher correlations than other indices. It is interesting to notice that AdaBoost generally has the second-highest correlation over all the imbalance recovery methods; however, its instance-level correlation is very low, as shown in Table I. As explained, the increase in the prediction score of AdaBoost is slight, but it changes the prediction and thus influences the F1 score. The correlations of kDN and CL are generally 0.1 less than those of BI^3 because they do not consider the imbalance in the index. They use pure data complexity to describe the effect caused by imbalance and are thus less accurate than BI^3. CM shows low correlations because it combines the neighborhood indicator values of all the majority and minority class samples. It can be used to represent the overall classification complexity of a data set but cannot show the impact of imbalance on it. IR is also compared, as an index for data-level correlation. However, most correlations between IR and the imbalance recovery methods are lower than 0.4. Thus, IR is not effective as an index for describing the influence of the class imbalance problem.

TABLE II

DATA-LEVEL SPEARMAN RANKED CORRELATION BETWEEN THE INDICES AND THE IMPROVEMENT IN THE F1 SCORE BY DIFFERENT IMBALANCE RECOVERY METHODS ON DATA SETS GROUP *syn_overlap*

		OS	US	SMOTE	SW
kDN	SVM	0.6883	0.6754	0.7036	0.6938
	CART	0.3656	0.5782	0.4497	0.4337
	5NN	0.3216	0.5628	0.4454	0.3985
	RF	0.4863	0.6647	0.5672	0.4918
	AdaBoost	0.5804	0.5601	0.5905	0.5821
CL	SVM	0.6731	0.6478	0.6894	0.6786
	CART	0.4420	0.5536	0.4860	0.4814
	5NN	0.4311	0.5477	0.4940	0.4611
	RF	0.5378	0.6148	0.5737	0.5347
	AdaBoost	0.4346	0.4156	0.4260	0.4388
CM	SVM	0.3600	0.3346	0.3753	0.3655
	CART	0.2650	0.2357	0.1693	0.2184
	5NN	0.2183	0.2407	0.1809	0.1866
	RF	0.3793	0.3270	0.2956	0.3999
	AdaBoost	0.2398	0.1664	0.2206	0.2338
IR	SVM	0.3312	0.3540	0.3324	0.3324
	CART	0.1909	0.3674	0.3494	0.2958
	5NN	0.1811	0.3671	0.3203	0.2849
	RF	0.1538	0.3459	0.3061	0.1461
	AdaBoost	0.3742	0.4403	0.4154	0.3844
BI^3	SVM	**0.7764**	**0.7710**	**0.7900**	**0.7807**
	CART	**0.4560**	**0.6883**	**0.5716**	**0.5485**
	5NN	**0.4263**	**0.6757**	**0.5682**	**0.5219**
	RF	**0.5682**	**0.7587**	**0.6709**	**0.5682**
	AdaBoost	**0.6910**	**0.6998**	**0.7101**	**0.6951**

TABLE III

BI^3 VALUES ON DATA SET GROUP *syn_overlap* AVERAGED OVER TEN DIFFERENT VARIANCES

	$dist = 0$	$dist = 1$	$dist = 2$	$dist = 3$
$IR = 5$	0.2646	0.2037	0.1055	0.0332
$IR = 10$	0.3696	0.2895	0.1580	0.0505
$IR = 50$	0.5120	0.4639	0.2593	0.1119

In summary, on data set group $syn\_overlap$, BI^3 has a high correlation with the improvement of the F1 score by imbalance recovery methods on all classifiers. BI^3 is, therefore, a proper index to describe the possible level of improvement in the F1 score by applying imbalance recovery methods. Thus, if a data set has a low BI^3 value, careful consideration should be given before applying imbalance recovery methods because any improvement may be limited or even negative. Table III shows the BI^3 values averaged over ten different variances on data set group $syn\_overlap$. When the overlapping region is reduced, BI^3 decreases as the distance between two classes increases. In addition, when IR is increasing, BI^3 is also increased. It is interesting to notice that when dist = 3 and IR = 50, where the two classes seldom overlap, the BI^3 value is comparable with dist = 2 and IR = 5. This again confirms that IR is not the only cause of classification performance degeneration and that BI^3 can more properly describe the impact brought by imbalance.

2) Results on Data Set Group syn_noise: The instance-level correlation is shown in Table IV. As the same as $syn\_overlap$, the results for IBI^3 also show the highest correlations. However, the correlations of SVM, CART, RF, and AdaBoost are generally lower than those of $syn\_overlap$ shown in Table I. The correlations of 5NN of syn_noise

TABLE IV

INSTANCE-LEVEL SPEARMAN RANKED CORRELATION BETWEEN THE INDICES AND THE INCREASE OF PREDICTION SCORE OF MINORITY CLASS SAMPLES ON DATA SETS GROUP *syn_noise*

		OS	US	SMOTE	SW
kDN	SVM	0.5958	0.6488	0.5856	0.3945
	CART	-0.0517	0.5487	0.2505	0.1050
	5NN	0.1565	0.7114	0.4406	0.2298
	RF	-0.0442	0.6193	0.2335	0.1269
	AdaBoost	0.1323	-0.4109	0.1510	0.1195
CL	SVM	0.4814	0.5104	0.4749	0.4822
	CART	0.1185	0.3116	0.1503	0.0186
	5NN	0.0068	0.3447	0.2026	0.0245
	RF	0.0587	0.4125	0.1903	0.0281
	AdaBoost	0.0039	-0.4974	0.0371	0.0266
IBI^3	SVM	**0.7283**	**0.7421**	**0.7222**	**0.4516**
	CART	**0.1836**	**0.6984**	**0.4868**	**0.3605**
	5NN	**0.5170**	**0.9150**	**0.7487**	**0.6372**
	RF	**0.3223**	**0.7763**	**0.5727**	**0.4784**
	AdaBoost	**0.2358**	**-0.1407**	**0.1957**	**0.2255**

TABLE V

DATA-LEVEL SPEARMAN RANKED CORRELATION BETWEEN THE INDICES AND THE IMPROVEMENT IN THE F1 SCORE BY DIFFERENT IMBALANCE RECOVERY METHODS ON DATA SETS GROUP *syn_noise*

		OS	US	SMOTE	SW
kDN	SVM	0.6785	0.6748	0.6750	0.6888
	CART	0.4744	0.3890	0.3046	0.4541
	5NN	0.4755	0.5358	0.4290	0.4196
	RF	0.6739	0.6245	0.5762	0.6911
	AdaBoost	0.6793	0.4907	0.6521	0.6811
CL	SVM	0.4504	0.4382	0.4459	0.4598
	CART	0.1943	0.0798	0.0039	0.1455
	5NN	0.2151	0.2783	0.1707	0.1072
	RF	0.4557	0.3545	0.3325	0.4797
	AdaBoost	0.4062	0.1945	0.3839	0.4051
CM	SVM	-0.0050	-0.0214	0.0019	0.0001
	CART	-0.2560	-0.2024	-0.3832	-0.3139
	5NN	-0.2430	-0.1333	-0.2233	-0.3631
	RF	0.0313	-0.0439	-0.0812	0.0628
	AdaBoost	-0.0750	-0.2031	-0.0503	-0.0795
IR	SVM	0.4561	0.4750	0.4496	0.4567
	CART	0.6240	0.4997	0.5161	0.6495
	5NN	0.5575	0.5094	0.5059	0.6491
	RF	0.4237	0.4688	0.4770	0.3975
	AdaBoost	0.5265	0.5463	0.4897	0.5358
BI^3	SVM	**0.7781**	**0.7806**	**0.7729**	**0.7865**
	CART	**0.6661**	**0.5588**	**0.6168**	**0.6613**
	5NN	**0.6689**	**0.6725**	**0.6033**	**0.6503**
	RF	**0.7733**	**0.7478**	**0.7114**	**0.7781**
	AdaBoost	**0.8045**	**0.6571**	**0.7720**	**0.8104**

are comparable with those of *syn_overlap* because IBI^3 is based on kNN and some minority class noise in the deep region of the majority class receives low IBI^3 value according to (8). However, the prediction score of classifiers, such as SVM and RF, on these noised points will differ significantly if imbalance recovery methods are applied. Therefore, it makes the correlations lower than those of *syn_overlap*. Similarly, kDN has lower correlations than those of *syn_overlap*. The correlations of CL are low because it is based on the Naive Bayes. When a data set has noise, the mean and variance cannot be well estimated, so the correlations are also low.

The data-level correlation is shown in Table V. Most of the correlations of BI^3 are greater than 0.6. CL has very low correlations with the improvement in the F1 score because it is sensitive to the noise. CM even generates negative correlations, which means that it is not a proper index for a description of the extent of imbalance of a noisy data set. Surprisingly, IR shows comparable correlations with kDN, which means that if the factor of overlapping is fixed, IR can still partially represent the impact of imbalance to the data set although noise exists.

TABLE VI

BI^3 VALUES ON DATA SET GROUP *syn_noise* AVERAGED OVER TEN DIFFERENT VARIANCES

	$noise = 0$	$noise = 0.1$	$noise = 0.2$	$noise = 0.3$
$IR = 5$	0.0803	0.1487	0.1988	0.2429
$IR = 10$	0.1156	0.1927	0.2529	0.3061
$IR = 50$	0.2261	0.2929	0.3446	0.3978

TABLE VII

INSTANCE-LEVEL SPEARMAN RANKED CORRELATION BETWEEN THE INDICES AND THE INCREASE OF PREDICTION SCORE OF MINORITY CLASS SAMPLES ON DATA SETS GROUP *syn_disjunct*

		OS	US	SMOTE	SW
kDN	SVM	0.5922	0.6722	0.5534	0.4776
	CART	-0.0992	0.6112	0.2176	0.0838
	5NN	0.1143	0.7934	0.4335	0.1987
	RF	-0.1479	0.7053	0.1394	0.1019
	AdaBoost	0.1968	-0.1286	0.1309	0.2074
CL	SVM	-0.2138	-0.1827	-0.1978	-0.1276
	CART	-0.0538	-0.2117	-0.1921	-0.1153
	5NN	-0.1531	-0.2443	-0.2113	-0.1560
	RF	-0.1497	-0.2627	-0.2624	-0.1684
	AdaBoost	0.0080	0.3531	0.0052	0.0040
IBI^3	SVM	**0.7752**	**0.7782**	**0.7522**	**0.5405**
	CART	**0.0596**	**0.6765**	**0.4596**	**0.2681**
	5NN	**0.4488**	**0.9101**	**0.7441**	**0.5768**
	RF	**0.1490**	**0.7726**	**0.4740**	**0.3610**
	AdaBoost	**0.3447**	**0.0151**	**0.2536**	**0.3591**

Table VI shows the BI^3 values averaged over ten different variances on data set group *syn_noise*. As the noise level increases or IR increases, the index value also increases. Both IR and the noise level affect BI^3, and this again confirms that the performance of a classifier on the imbalanced data set does not depend only on IR.

3) Results on Data Set Group syn_disjunct: The instance-level correlation is shown in Table VII. It can be seen that IBI^3 shows the highest correlation among all indices. CL shows several negative correlations because the classes in data set group *syn_disjunct* are not normally distributed if the number of disjuncts is greater than one. Among the imbalance recovery methods, US shows the highest correlation because the classes can be easily separated after US is adopted even if there are many disjuncts. For the classifiers, SVM and 5NN generally have higher correlations than the tree-based methods.

The data-level correlation is shown in Table VIII, where the correlations in *syn_disjunct* is generally higher than those in *syn_overlap* and *syn_noise*. BI^3 can, therefore, better reflect the data complexity caused by small disjuncts. kDN and BI^3 show almost the same correlations among various combinations of classifier and imbalance recovery methods, possibly because little overlap occurs between the classes in *syn_disjunct* and no noise is present. As a result, few minority class samples are located in the deep region of the

TABLE VIII

DATA-LEVEL SPEARMAN RANKED CORRELATION BETWEEN THE INDICES AND THE IMPROVEMENT IN THE F1 SCORE BY DIFFERENT IMBALANCE RECOVERY METHODS ON DATA SETS GROUP $syn\_disjunct$

		OS	US	SMOTE	SW
kDN	SVM	0.9150	0.9200	0.9153	0.9113
	CART	0.4818	**0.9678**	**0.9269**	0.6207
	5NN	0.6458	**0.9699**	**0.9429**	0.6525
	RF	0.4312	0.9106	0.8297	0.3829
	AdaBoost	**0.5179**	**0.5925**	0.5483	0.4918
CL	SVM	-0.5086	-0.5122	-0.4926	-0.5144
	CART	-0.5244	-0.5014	-0.6116	-0.5687
	5NN	-0.5855	-0.4935	-0.5456	-0.5793
	RF	-0.5784	-0.5735	-0.6637	-0.4793
	AdaBoost	-0.3795	-0.4098	-0.3926	-0.3638
CM	SVM	0.0275	0.0369	0.0206	0.0374
	CART	0.4581	-0.1160	-0.0364	0.3709
	5NN	0.4315	-0.0715	-0.0017	0.4496
	RF	0.4959	0.0088	0.1251	0.4636
	AdaBoost	0.0554	0.0376	0.0582	0.0857
IR	SVM	0.4676	0.4624	0.4748	0.4585
	CART	-0.0371	0.6014	0.5181	0.0868
	5NN	0.0376	0.5678	0.5006	0.0270
	RF	-0.1118	0.4758	0.3439	-0.1145
	AdaBoost	0.2634	0.3182	0.2720	0.2256
BI^3	SVM	**0.9205**	**0.9313**	**0.9187**	**0.9196**
	CART	**0.5272**	0.9343	0.9146	**0.6597**
	5NN	**0.7161**	0.9518	0.9390	**0.7342**
	RF	**0.4956**	**0.9198**	**0.8578**	**0.4175**
	AdaBoost	0.5119	0.5837	**0.5535**	**0.5114**

TABLE IX

BI^3 VALUES ON DATA SET GROUP $syn\_disjunct$ AVERAGED OVER TEN DIFFERENT VARIANCES

	$disj. = 1$	$disj. = 2$	$disj. = 4$	$disj. = 8$
$IR = 5$	0.1117	0.3329	0.3944	0.4153
$IR = 10$	0.1421	0.3914	0.4491	0.4620
$IR = 50$	0.2574	0.5222	0.5277	0.5125

majority class, where these samples have high kDN values, which makes the correlation different.

Table IX shows the BI^3 values averaged over ten different variances on data set group $syn\_disjunct$. BI^3 increases as the number of disjuncts and IR increase. For IR $= 50$ with disjunct $= 2, 4, 8$, the values of BI^3 are almost the same. Thus, when the classes are highly imbalanced, IR dominates the data complexity, and increasing the number of disjuncts does not further deteriorate the classification performance of the minority class.

B. Real Benchmark Data

We use 80 real data sets from the KEEL data set repository [40]. The details of the data sets are given in Table X. The IR ranges from 1.86 to 129.44 over all 80 data sets. For real benchmark data, we also compare the proposed IBI^3 and BI^3 with kDN, CL, CM, and IR, in the instance and data levels, respectively.

The instance-level correlation is shown in Table XI. IBI^3 shows greater correlations than kDN and CL because it considers the imbalance factor into the index. 5NN achieves the greatest correlation of all imbalance recovery methods because BI^3 is based on kNN and RF achieves the second-highest correlation. In terms of the imbalance recovery methods, US achieves the greatest correlation, where the correlations are greater than 0.5, except with AdaBoost.

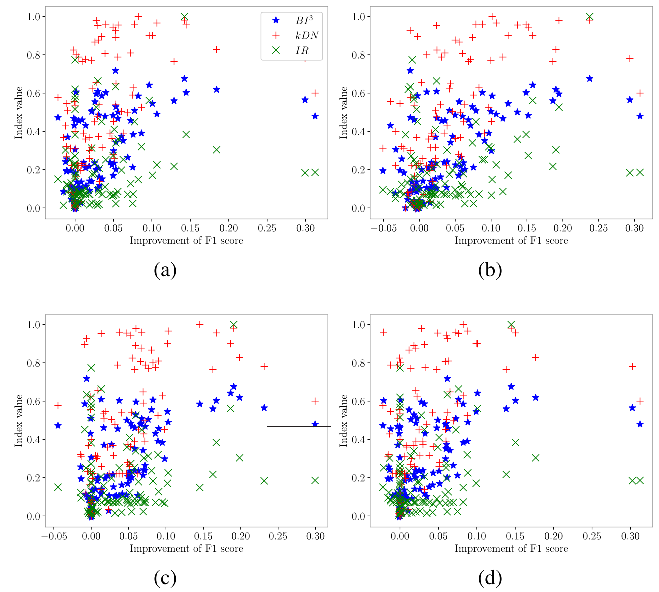

Fig. 6. BI^3, kDN, and IR over 80 KEEL real benchmark imbalanced datasets sorted along the improvement of F1 score of AdaBoost classifier with (a) OS, (b) US, (c) SMOTE, and (d) SW.

The data-level correlation is shown in Table XII. BI^3 achieves the highest correlation, and most of the correlations are greater than 0.5, which indicates a strong correlation. Thus, given a real data set, we can calculate BI^3 without training and testing to estimate the extent of improvement by using imbalance recovery methods. kDN shows greater correlation than IR in general, which means that the data complexity using the nearest neighbor can still better represent the imbalance impact on imbalanced data than referring to the IR. CM achieves low correlation, which means that CM may be a good data complex measurement for imbalanced data but not a proper index for describing the imbalance impact. 5NN achieves a high correlation at the instance level but low correlation at the data level, possibly because the imbalance recovery methods applied to 5NN simply change the prediction score but do not effectively improve the F1 score. As in the synthetic data situation, AdaBoost shows a low correlation at the instance level but a high correlation at the data level. The averaged correlation of AdaBoost over all imbalance recovery methods is higher than other classifiers, and thus, BI^3 can properly reflect the extent of improvement in the F1 score when applying imbalance recovery methods to AdaBoost.

Fig. 6 shows BI^3, kDN, and IR over 80 real benchmark data sets on the AdaBoost classifier with various imbalance recovery methods. IR is normalized to [0,1] to fit in the figure. Most of the IR points are located at the bottom, which means that the same level of IR leads to different levels of F1 score improvement. Conversely, most of the kDN points are scattered at the top, which means that kDN tends to overestimate the improvement in the F1 score because it only counts the number of neighbors with different class labels for the minority class samples. In comparison,

TABLE X

INFORMATION OF 80 IMBALANCED DATA SETS

dataset	#Inst.	#Attr.	IR	BI^3	dataset	#Inst.	#Attr.	IR	BI^3
ecoli-0_vs_1	220	7	1.86	0.01	*yeast-1_vs_7*	459	7	14.30	0.48
pima	768	8	1.87	0.10	*glass4*	214	9	15.46	0.37
iris0	150	4	2.00	0.00	*ecoli4*	336	7	15.80	0.19
glass0	214	9	2.06	0.09	*abalone9-18*	731	8	16.40	0.46
yeast1	1484	8	2.46	0.16	*dermatology-6*	358	34	16.90	0.04
haberman	306	3	2.78	0.20	*yeast-1-4-5-8_vs_7*	693	8	22.10	0.55
vehicle2	846	18	2.88	0.10	*yeast-2_vs_8*	482	8	23.10	0.24
vehicle1	846	18	2.90	0.20	*flare-F*	1066	11	23.79	0.56
*glass-0-1-2-3*_vs_4-5-6	214	9	3.20	0.10	*car-good*	1728	6	24.04	0.48
vehicle0	846	18	3.25	0.09	*car-vgood*	1728	6	25.58	0.37
ecoli1	336	7	3.36	0.14	*kr-vs-k-one_vs_draw*	2901	6	26.63	0.12
ecoli2	336	7	5.46	0.10	*kr-vs-k-one_vs_fifteen*	2244	6	27.77	0.01
segment0	2308	19	6.02	0.02	*yeast4*	1484	8	28.10	0.56
glass6	214	9	6.38	0.08	*winequality-red-4*	1599	11	29.17	0.49
yeast3	1484	8	8.10	0.22	*poker-9_vs_7*	244	10	29.50	0.47
ecoli3	336	7	8.60	0.30	*kddcup-guess_passwd_vs_satan*	1642	41	29.98	0.00
page-blocks0	5472	10	8.79	0.17	*yeast-1-2-8-9_vs_7*	947	8	30.57	0.55
ecoli-0-3-4_vs_5	200	7	9.00	0.11	*winequality-white-9_vs_4*	168	11	32.60	0.60
yeast-2_vs_4	514	8	9.08	0.22	*yeast5*	1484	8	32.73	0.35
ecoli-0-6-7_vs_3-5	222	7	9.09	0.24	*kr-vs-k-three_vs_eleven*	2935	6	35.23	0.08
ecoli-0-2-3-4_vs_5	202	7	9.10	0.11	*winequality-red-8_vs_6*	656	11	35.44	0.48
glass-0-1-5_vs_2	172	9	9.12	0.43	*abalone-17_vs_7-8-9-10*	2338	8	39.31	0.62
yeast-0-3-5-9_vs_7-8	506	8	9.12	0.34	*abalone-21_vs_8*	581	8	40.50	0.50
yeast-0-2-5-6_vs_3-7-8-9	1004	8	9.14	0.26	*yeast6*	1484	8	41.40	0.39
yeast-0-2-5-7-9_vs_3-6-8	1004	8	9.14	0.14	*winequality-white-3_vs_7*	900	11	44.00	0.53
ecoli-0-4-6_vs_5	203	6	9.15	0.11	*winequality-red-8_vs_6-7*	855	11	46.50	0.50
ecoli-0-1_vs_2-3-5	244	7	9.17	0.15	*kddcup-land_vs_portsweep*	1061	41	49.52	0.00
ecoli-0-2-6-7_vs_3-5	224	7	9.18	0.24	*abalone-19_vs_10-11-12-13*	1622	8	49.69	0.60
ecoli-0-3-4-6_vs_5	205	7	9.25	0.11	*kr-vs-k-zero_vs_eight*	1460	6	53.07	0.23
vowel0	988	13	9.98	0.03	*winequality-white-3-9_vs_5*	1482	11	58.28	0.51
ecoli-0-6-7_vs_5	220	6	10.00	0.21	*poker-8-9_vs_6*	1485	10	58.40	0.59
glass-0-1-6_vs_2	192	9	10.29	0.45	*shuttle-2_vs_5*	3316	9	66.67	0.02
ecoli-0-1-4-7_vs_2-3-5-6	336	7	10.59	0.21	*winequality-red-3_vs_5*	691	11	68.10	0.60
led7digit-0-2-4-5-6-7-8-9_vs_1	443	7	10.97	0.20	*abalone-20_vs_8-9-10*	1916	8	72.69	0.64
ecoli-0-1_vs_5	240	6	11.00	0.11	*kddcup-buffer_overflow_vs_back*	2233	41	73.43	0.04
glass-0-1-4-6_vs_2	205	9	11.06	0.47	*kddcup-land_vs_satan*	1610	41	75.67	0.02
glass2	214	9	11.59	0.46	*kr-vs-k-zero_vs_fifteen*	2193	6	80.22	0.07
cleveland-0_vs_4	173	13	12.31	0.49	*poker-8-9_vs_5*	2075	10	82.00	0.72
ecoli-0-1-4-6_vs_5	280	6	13.00	0.11	*poker-8_vs_6*	1477	10	85.88	0.61
shuttle-c0-vs-c4	1829	9	13.87	0.01	*abalone19*	4174	8	129.44	0.68

TABLE XI

INSTANCE-LEVEL SPEARMAN RANKED CORRELATION BETWEEN THE INDICES AND THE PREDICTION SCORE INCREASE OF MINORITY CLASS SAMPLE OVER 80 REAL DATA SETS

		OS	US	SMOTE	SW
kDN	SVM	0.3117	0.5224	0.3157	0.1459
	CART	0.0996	0.5103	0.1941	0.2120
	5NN	0.3951	0.8252	0.5799	0.4894
	RF	0.3080	0.6825	0.3898	0.3707
	AdaBoost	0.1963	-0.0735	0.2248	0.1711
CL	SVM	0.1689	0.3802	0.2002	0.0684
	CART	0.1077	0.3216	0.1562	0.1768
	5NN	0.2889	0.4326	0.3484	0.3130
	RF	0.2610	0.4552	0.2931	0.3039
	AdaBoost	0.1336	0.1391	0.1842	0.1367
IBI^3	SVM	**0.3864**	**0.5565**	**0.4012**	**0.1481**
	CART	**0.1633**	**0.5175**	**0.2315**	**0.2703**
	5NN	**0.6018**	**0.8981**	**0.7613**	**0.7080**
	RF	**0.4520**	**0.7311**	**0.5050**	**0.4936**
	AdaBoost	**0.2795**	**0.0925**	**0.2842**	**0.2699**

BI^3 generally increases as the improvement in the F1 score increases, as shown in Fig. 6. Only a few points lie on the region so that the improvement in the F1 score is close to 0, but BI^3 has high values. The selected imbalance recovery methods are the simplest ones found in the literature and, thus, may not be effective in improving the F1 score for all the data sets.

We specifically studied two real benchmark data sets from Table X: *kddcup-land_vs_satan* and *haberman*. The data set *kddcup-land_vs_satan* has IR $= 75.67$, which is highly imbalanced, but $\text{BI}^3 = 0.02$, which means that the imbalance impact on this data set is low. Table XIII shows the F1 scores of different classifiers and the improvement in the F1 scores from the imbalance recovery methods. The F1 scores for classifiers without imbalance recovery are already very high. Therefore, the improvements from the imbalance recovery methods are very limited. Most are near or equals to 0. US even deteriorates the F1 scores for all classifiers and shows negative improvement, possibly because there is a greater decrease in precision than an increase in recall as the F1 score is the harmonic mean between precision and recall. The result obtained from data set *kddcup-land_vs_satan* shows that the minority class in the data set itself is not very difficult to classify although it is significantly outnumbered by the majority class. In contrast, the data set *haberman* has IR $= 2.78$, which is not highly imbalanced compared with data set *kddcup-land_vs_satan*, but its BI^3 value is 0.2. Table XIV shows the F1 scores and the improvements of various classifiers and imbalance recovery methods. Most of the imbalance recovery methods can, therefore, make obvious improvements on all classifiers. Most improvements in the F1 scores are greater than 0.1. In general, imbalance recovery methods should be applied to data set $haberman$ because the F1 score can be actually

TABLE XII

THE DATA-LEVEL SPEARMAN RANKED CORRELATION BETWEEN THE INDICES AND THE IMPROVEMENT IN THE F1 SCORE BY DIFFERENT IMBALANCE RECOVERY METHODS ON DATA LEVEL OVER 80 REAL DATA SETS

		OS	US	SMOTE	SW
kDN	SVM	0.4565	0.4531	0.4479	0.4607
	CART	0.4584	0.5742	0.5407	0.5052
	5NN	0.2738	0.3042	0.4527	0.3828
	RF	0.2792	0.5029	0.5597	0.1060
	AdaBoost	0.6820	0.7211	0.6499	0.5789
CL	SVM	0.2066	0.2695	0.1939	0.2010
	CART	0.2330	0.4520	0.3118	0.3037
	5NN	0.3736	0.3711	0.4473	0.3885
	RF	0.3497	0.4383	0.4769	0.2733
	AdaBoost	0.5474	0.4020	0.4020	0.5663
CM	SVM	0.1684	0.0304	0.1120	0.1774
	CART	0.0141	0.0935	-0.0015	0.0619
	5NN	0.0420	0.0651	0.0343	0.1199
	RF	0.2167	0.1704	0.1603	0.1602
	AdaBoost	0.2913	0.2989	0.3425	0.2169
IR	SVM	0.2665	0.3744	0.3343	0.2629
	CART	0.3700	0.3151	0.4267	0.3414
	5NN	0.1492	0.1033	0.2843	0.1735
	RF	-0.0500	0.1572	0.1905	-0.1863
	AdaBoost	0.2656	0.2331	0.1781	0.2366
BI^3	SVM	**0.5423**	**0.5463**	**0.5395**	**0.5448**
	CART	**0.6314**	**0.6349**	**0.6854**	**0.6561**
	5NN	**0.4497**	**0.4406**	**0.6239**	**0.5497**
	RF	**0.3828**	**0.5420**	**0.6494**	**0.2035**
	AdaBoost	**0.7278**	**0.7693**	**0.7012**	**0.6249**

TABLE XIII

IMPROVEMENT IN THE F1 SCORE ON THE DATA SET *kddcup-land_vs_satan*. THE COLUMN NONE IS THE F1 SCORE OF THE CLASSIFIER WITHOUT IMBALANCE RECOVERY METHODS

	None	OS	US	SMOTE	SW
SVM	0.9114	+0.0000	−0.5494	+0.0000	+0.0000
CART	0.9346	−0.0050	−0.5495	−0.0050	+0.0000
5NN	0.9503	+0.0000	−0.5906	+0.0000	−0.0169
RF	0.9446	+0.0358	−0.3950	+0.0356	+0.0102
AdaBoost	0.9614	+0.0051	−0.5420	+0.0000	+0.0000

TABLE XIV

IMPROVEMENT IN THE F1 SCORE ON THE DATA SET *Haberman*. THE COLUMN NONE IS THE F1 SCORE OF THE CLASSIFIER WITHOUT IMBALANCE RECOVERY METHODS

	None	OS	US	SMOTE	SW
SVM	0.0376	+0.1054	+0.4067	+0.2108	+0.1120
CART	0.3009	+0.1130	+0.1386	+0.0903	+0.1452
5NN	0.2973	+0.1201	+0.1270	+0.1091	+0.1025
RF	0.3514	+0.1676	+0.1813	+0.1482	+0.1492
AdaBoost	0.3514	+0.0533	+0.0659	+0.0671	+0.0687

improved although its IR is not very high. This example again confirms that IR is not the only cause of the performance degeneration of an imbalanced data set. In empirical terms, we, therefore, suggest that the focus should be on other data factors in an imbalanced data set if its BI^3 value is lower than 0.05.

C. Parameter Sensitivity

The number of nearest neighbors, k_0, used in the calculation of BI^3 is set at five for all experiments. In this experiment, we compare the averaged correlation of BI^3 with different settings of k_0. We also verify the effectiveness of the flexible k_0 used in Algorithm 1, compared with that only using the fixed number of k_0, which is denoted as BI^3_f. Fig. 7 shows the correlation of BI^3 averaged over all classifiers and imbalance recovery methods increases the number of nearest neighbors k_0 from 2 to 50. Both instance- and data-level correlations have the highest values of around $k = 5$. As k_0 increases from 2 to 5, the averaged correlation increases and then decreases. Thus, $k = 5$ appears to be a proper selection for BI^3. In addition, the averaged correlation of BI^3 is higher than BI^3_f over all settings of k_0 for both data- and instance-level correlations, which confirms the effectiveness of the flexible k_0.

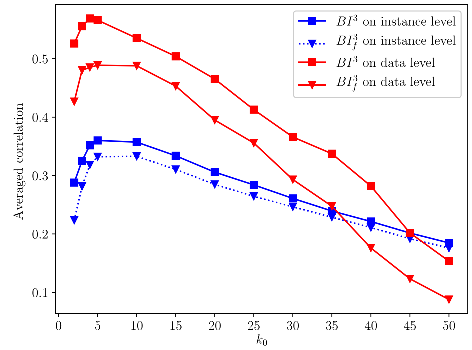

Fig. 7. Change of correlation of BI^3 and BI^3_f averaged over all classifiers and imbalance recovery methods as increasing the number of nearest neighbors k_0.

V. CONCLUSION

Most studies of class imbalance learning attempt to recover the accuracy loss caused by the IR. However, the accuracy loss is not only related to imbalance but also to many other data factors. Using IR to describe the classification difficulty of imbalanced data is inaccurate and misleading. In this article, we have proposed two measures IBI^3 and BI^3 to estimate the impact that is solely caused by imbalance at the instance and data levels, respectively. IBI^3 measures how much a single minority class sample is influenced by the imbalance. BI^3, which is the average over IBI^3, can be used as a measure of the degree of degradation in an imbalanced data set, and one can determine whether or not to apply imbalance recovery methods by referring to the BI^3 value instead of IR. The experiments on synthetic and real benchmark data sets have shown high correlations at both the instance and data levels with the improvements in the F1 score made by various imbalance recovery methods.

In addition to this work, there is still room for future research. For example, a classifier-oriented index can be proposed, which shows exactly how much the imbalance influences a specific classifier because each type of classifier has a different level of sensitivity to imbalance. Furthermore, IBI^3 can be incorporated into imbalance recovery methods, such as sampling or cost-sensitive methods, to help recover the loss caused by imbalance. In addition, taking advantage of BI^3 can guide the selection of a proper imbalance recovery method for a specific imbalanced data set. Because the recovery methods developed from the various theories and methodologies complement each other to some degree, their selection becomes particularly important as given an imbalanced data set.

References

[1] Q. Yang and X. Wu, "10 challenging problems in data mining research," *Int. J. Inf. Technol. Decis. Making*, vol. 5, no. 4, pp. 597–604, 2006.

[2] R. B. Rao, S. Krishnan, and R. S. Niculescu, "Data mining for improved cardiac care," *ACM SIGKDD Explor. Newslett.*, vol. 8, no. 1, pp. 3–10, 2006.

[3] A. Azaria, A. Richardson, S. Kraus, and V. S. Subrahmanian, "Behavioral analysis of insider threat: A survey and bootstrapped prediction in imbalanced data," *IEEE Trans. Comput. Social Syst.*, vol. 1, no. 2, pp. 135–155, Jun. 2014.

[4] S. Wang and X. Yao, "Using class imbalance learning for software defect prediction," *IEEE Trans. Rel.*, vol. 62, no. 2, pp. 434–443, Jun. 2013.

[5] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.

[6] P. Branco, L. Torgo, and R. P. Ribeiro, "A survey of predictive modeling on imbalanced domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 31:1–31:50, 2016.

[7] J. Van Hulse, T. M. Khoshgoftaar, and A. Napolitano, "Experimental perspectives on learning from imbalanced data," in *Proc. 24th Int. Conf. Mach. Learn.*, 2007, pp. 935–942.

[8] V. López, A. Fernández, S. García, V. Palade, and F. Herrera, "An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics," *Inf. Sci.*, vol. 250, pp. 113–141, Nov. 2013.

[9] T. Jo and N. Japkowicz, "Class imbalances versus small disjuncts," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 40–49, 2004.

[10] J. A. Sáez, J. Luengo, J. Stefanowski, and F. Herrera, "SMOTE–IPF: Addressing the noisy and borderline examples problem in imbalanced classification by a re-sampling method with filtering," *Inf. Sci.*, vol. 291, pp. 184–203, Jan. 2015.

[11] M. Galar, A. Fernandez, E. Barrenechea, H. Bustince, and F. Herrera, "A review on ensembles for the class imbalance problem: Bagging-, boosting-, and hybrid-based approaches," *IEEE Trans. Syst., Man, Cybern. C, Appl. Rev.*, vol. 42, no. 4, pp. 463–484, Jul. 2012.

[12] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," *J. Artif. Intell. Res.*, vol. 16, no. 1, pp. 321–357, 2002.

[13] G. E. Batista, R. C. Prati, and M. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explor. Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.

[14] X. Hong, S. Chen, and C. J. Harris, "A kernel-based two-class classifier for imbalanced data sets," *IEEE Trans. Neural Netw.*, vol. 18, no. 1, pp. 28–41, Jan. 2007.

[15] C. Elkan, "The foundations of cost-sensitive learning," in *Proc. 17th Int. Joint Conf. Artif. Intell.*, Seattle, WA, USA, 2001, pp. 973–978.

[16] C. X. Ling, V. S. Sheng, and Q. Yang, "Test strategies for cost-sensitive decision trees," *IEEE Trans. Knowl. Data Eng.*, vol. 18, no. 8, pp. 1055–1067, Aug. 2006.

[17] M. A. Davenport, R. G. Baraniuk, and C. D. Scott, "Tuning support vector machines for minimax and neyman-pearson classification," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 10, pp. 1888–1898, Oct. 2010.

[18] K. Napierala and J. Stefanowski, "Types of minority class examples and their influence on learning classifiers from imbalanced data," *J. Intell. Inf. Syst.*, vol. 46, no. 3, pp. 563–597, 2016.

[19] N. Japkowicz, "Class imbalances: Are we focusing on the right issue," in *Proc. ICML Workshop Learn. Imbalanced Data Sets II*, 2003.

[20] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "Learning with class skews and small disjuncts," in *Proc. Brazilian Symp. Artif. Intell.* Berlin, Germany: Springer, 2004, pp. 296–306.

[21] M. Kubat and S. Matwin, "Addressing the curse of imbalanced training sets: One-sided selection," in *Proc. Int. Conf. Mach. Learn.*, Nashville, TN, USA, vol. 97, 1997, pp. 179–186.

[22] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," in *Proc. Conf. Artif. Intell. Med. Eur.* Berlin, Germany: Springer, 2001, pp. 63–66.

[23] K. Napierała, J. Stefanowski, and S. Wilk, "Learning from imbalanced data in presence of noisy and borderline examples," in *Proc. Int. Conf. Rough Sets Current Trends Comput.* Berlin, Germany: Springer, 2010, pp. 158–167.

[24] V. García, R. A. Mollineda, and J. S. Sánchez, "On the k-NN performance in a challenging scenario of imbalance and overlapping," *Pattern Anal. Appl.*, vol. 11, nos. 3–4, pp. 269–280, 2008.

[25] N. Anwar, G. Jones, and S. Ganesh, "Measurement of data complexity for classification problems with unbalanced data," *Stat. Anal. Data Mining*, vol. 7, no. 3, pp. 194–211, 2014.

[26] R. C. Prati, G. E. A. P. A. Batista, and M. C. Monard, "Class imbalances versus class overlapping: An analysis of a learning system behavior," in *Proc. Mexican Int. Conf. Artif. Intell.* Berlin, Germany: Springer, 2004, pp. 312–321.

[27] T. K. Ho and M. Basu, "Complexity measures of supervised classification problems," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 24, no. 3, pp. 289–300, Mar. 2002.

[28] M. R. Smith, T. Martinez, and C. Giraud-Carrier, "An instance level analysis of data complexity," *Mach. Learn.*, vol. 95, no. 2, pp. 225–256, 2014.

[29] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, Jan. 1967.

[30] H. Han, W.-Y. Wang, and B.-H. Mao, "Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning," in *Proc. Int. Conf. Intell. Comput.*, Hefei, China, 2005, pp. 878–887.

[31] H. He, Y. Bai, E. A. Garcia, and S. Li, "ADASYN: Adaptive synthetic sampling approach for imbalanced learning," in *Proc. IEEE Int. Joint Conf. Neural Netw.*, Hong Kong, Jun. 2008, pp. 1322–1328.

[32] M. Kendall and J. D. Gibbons, *Rank Correlation Methods*, 5th ed. Oxford, U.K.: Oxford Univ. Press, 1990.

[33] V. N. Vapnik, *Statistical Learning Theory*, vol. 1. New York, NY, USA: Wiley, 1998.

[34] L. Breiman, J. Friedman, C. J. Stone, and R. A. Olshen, *Classification and Regression Trees*. Boca Raton, FL, USA: CRC Press, 1984.

[35] R. O. Duda and P. E. Hart, *Pattern Classification and Scene Analysis*, vol. 3. New York, NY, USA: Wiley, 1973.

[36] L. Breiman, "Random forests," *Mach. Learn.*, vol. 45, no. 1, pp. 5–32, 2001.

[37] Y. Freund and R. E. Schapire, "A decision-theoretic generalization of on-line learning and an application to boosting," *J. Comput. Syst. Sci.*, vol. 55, no. 1, pp. 119–139, Aug. 1997.

[38] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, Oct. 2011.

[39] G. Lemaître, F. Nogueira, and C. K. Aridas, "Imbalanced-learn: A Python toolbox to tackle the curse of imbalanced datasets in machine learning," *J. Mach. Learn. Res.*, vol. 18, no. 17, pp. 1–5, 2017. [Online]. Available: http://jmlr.org/papers/v18/16-365.html

[40] J. Alcalá-Fdez, A. Fernández, J. Luengo, J. Derrac, S. García, L. Sánchez, and F. Herrera, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Multiple-Valued Logic Soft Comput.*, vol. 17, pp. 255–281, Jun. 2011.

Yang Lu (S'13–M'19) received the B.Sc. and M.Sc. degrees in software engineering from the University of Macau, Macau, China, in 2012 and 2014, respectively, and the Ph.D. degree in computer science from Hong Kong Baptist University, Hong Kong, in 2019.

He is currently an Assistant Professor with the Department of Computer Science, School of Informatics, Xiamen University, Xiamen, China. He is also with the Department of Computer Science, Hong Kong Baptist University. His current research interests include imbalanced data learning, clustering, ensemble learning, and online learning.

Yiu-Ming Cheung (SM'06–F'18) received the Ph.D. degree from the Department of Computer Science and Engineering, The Chinese University of Hong Kong, Hong Kong.

He is currently a Full Professor with the Department of Computer Science, Hong Kong Baptist University, Hong Kong. His current research interests include machine learning, pattern recognition, visual computing, and optimization.

Prof. Cheung is an IET Fellow, a BCS Fellow, an RSA Fellow, and a Distinguished Fellow of IETI. He is also the Founding Chair of the Computational Intelligence Chapter of the IEEE Hong Kong Section and the Chair of the Technical Committee on Intelligent Informatics of the IEEE Computer Society. He also serves as an Associate Editor for the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the IEEE TRANSACTIONS ON CYBERNETICS, *Pattern Recognition*, *Knowledge and Information Systems*, and *Neurocomputing*, to name a few.

Yuan Yan Tang (S'88–M'88–SM'96–F'04–LF'16) received the B.Sc. degree in electrical and computer engineering from Chongqing University, Chongqing, China, the M.Eng. degree in electrical engineering from the Beijing Institute of Posts and Telecommunications, Beijing, China, and the Ph.D. degree in computer science from Concordia University, Montreal, QC, Canada.

He is currently a Chair Professor with the Faculty of Science and Technology, UOW College Hong Kong/Community College of City University, Hong Kong, and an Emeritus Chair Professor with the Faculty of Science and Technology, University of Macau, Macau, China. He is also a Professor, an Adjunct Professor, and an Honorary Professor with Chongqing University, Concordia University, and Hong Kong Baptist University, Hong Kong, respectively. He has authored or coauthored over 400 academic articles, over 25 monographs, books, and book chapters. His current research interests include wavelets, pattern recognition, and image processing.

Prof. Tang is an IAPR Fellow. He was the general chair, the program chair, and a committee member for many international conferences. He is the Founder and the Chair of the Pattern Recognition Committee of the IEEE TRANSACTIONS ON SYSTEMS, MAN, AND CYBERNETICS. He is the Founder and the General Chair of the series International Conferences on Wavelets Analysis and Pattern Recognition. He is the Founder and the Chair of the Macau Branch of International Association of Pattern Recognition. He is also the Founder and the Editor-in-Chief of the *International Journal of Wavelets, Multiresolution and Information Processing* and an Associate Editor of several international journals.