Entropy-based Sampling Approaches for Multi-Class Imbalanced Problems

Lusi Li, *Student Member, IEEE*, Haibo He, *Fellow, IEEE*, and Jie Li, *Member, IEEE*

Abstract—In data mining, large differences between multi-class distributions regarded as class imbalance issues have been known to hinder the classification performance. Unfortunately, existing sampling methods have shown their deficiencies such as causing the problems of over-generation and over-lapping by oversampling techniques, or the excessive loss of significant information by undersampling techniques. This paper presents three proposed sampling approaches for imbalanced learning: the first one is the entropy-based oversampling (EOS) approach; the second one is the entropy-based undersampling (EUS) approach; the third one is the entropy-based hybrid sampling (EHS) approach combined by both oversampling and undersampling approaches. These three approaches are based on a new class imbalance metric, termed entropy-based imbalance degree (EID), considering the differences of information contents between classes instead of traditional imbalance-ratio. Specifically, to balance a data set after evaluating the information influence degree of each instance, EOS generates new instances around difficult-to-learn instances and only remains the informative ones. EUS removes easy-to-learn instances. While EHS can do both simultaneously. Finally, we use all the generated and remaining instances to train several classifiers. Extensive experiments over synthetic and real-world data sets demonstrate the effectiveness of our approaches.

Index Terms—Imbalanced learning, oversampling, undersampling, hybrid sampling, entropy

1 Introduction

Imbalanced learning has attracted a great deal of interests in the research community. Most of the well-known data mining and machine learning techniques are proposed to solve classification problems with respect to reasonably balanced class distributions [1]. However, this assumption is not always true for a skewed class distribution problem existing in many real-world data sets, in which several classes (the majorities) are over-represented by a large number of instances but some others (the minorities) are under-represented by only a few. The solutions for the class-imbalance problem using traditional learning techniques bias the dominant classes resulting in poor classification performance. For an extremely multi-class imbalanced data set, imbalanced classification performance may be provided by traditional classifiers with a near 100 percent accuracy for the majorities and with close to 0 percent accuracy for the minorities. Hence, the class-imbalance problem is considered as a significant impediment to the success of precise classifiers.

To overcome this impediment, plenty of methods have been designed recently to balance the distributions between the majorities and the minorities [2], which can be divided into two major groups: the algorithm-level methods and the data-level methods. For the former, they attempt to modify existing classification algorithms to improve learning performance. Among them, cost-sensitive methods specify the costs for misclassifying minority instances, and kernel-based methods modify the kernels to improve the learning of minority instances [3], [4]. For the latter, they aim to balance the skewed class distribution before training classifiers. Oversampling and undersampling techniques are commonly used to imbalanced learning [5].

To achieve a balance, oversampling methods create new instances by replicating original instances (e.g., random oversampling (ROS) [1]), or generating synthetic instances (e.g., synthetic minority over-sampling technique (SMOTE) [6]). In order to avoid creating error-prone decision data spaces for the minorities (i.e., over-fitting) and redundant minority instances (i.e., over-generation), newly added instances may need to enlarge the original minority space with an appropriate number and reduce the imbalance degree in data space [7], [8], [9]. However, ROS cannot mitigate this kind of imbalance and SMOTE easily leads to the generation of noisy and wrong minority instances. Besides, most of oversampling techniques, such as Safe-level-SMOTE [10], ADASYN [11], RAMO [12], MWMOTE [13], borderline-SMOTE [14], AMDO [15], and EDOS [16], are developed to solve two-class imbalance problem. If they are applied to handle multi-class imbalance problem, either class transformation or class decomposition techniques are required to convert the multi-class problem into a few two-class problems. In this case, some significant information of original multi-class data may be lost (e.g., for the correlated data in three classes, partial correlation information will be lost in each two-class task). Therefore, the created

- *L. Li and H. He are with the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island, Kingston, RI 02881. E-mail: {lli, he}@ele.uri.edu.*
- *J. Li is with the School of Electrical and Information Engineering, Chongqing University of Science and Technology, Chongqing 401331, China. E-mail: 2014008@cqust.edu.cn.*

Manuscript received 26 June 2018; revised 11 Mar. 2019; accepted 22 Apr. 2019. Date of publication 30 Apr. 2019; date of current version 6 Oct. 2020. (Corresponding author: Haibo He.)
Recommended for acceptance by X. Li.
Digital Object Identifier no. 10.1109/TKDE.2019.2913859

instances based on the two-class tasks are unable to adequately fit the whole distribution of original multi-class data.

On the other hand, undersampling methods remove a subset of majority instances to balance a data set (e.g., random undersampling (RUS) and resampling). The major advantage of undersampling is that it can ensure the realness of all training instances. However, RUS randomly selects instances from majority classes without considering if they are informative and representative [17], [18]. Resampling may raise the problem of overlapping. Furthermore, bagging and boosting are combined with sampling techniques to enhance the learning of imbalanced data [19], [20], such as OverBagg [21], RUSBoost [22], SMOTEBoost [23], EasyEnsemble (EE), BalanceCascade (BC) [17], GIREnOS, and GIREnUs [24]. Additionally, SMOTE + Tomek links and SMOTE + ENN are two approaches combined over- and under-sampling algorithms [25]. These two methods are used to remove instances from all the classes after oversampling minority instances.

In the literature, class-imbalance degree is often measured by imbalance-ratio (IR) due to its simplicity. IR refers to the ratio of the number of instances from the most majority class to that from the most minority class. However, it is not an informative measure to describe the differences among multi-classes, where there exist other classes and all the classes are needed to be considered. Thus, IR is not appropriate to measure multi-class imbalance degree.

In order to overcome this drawback, we introduce a new metric, termed entropy-based imbalance degree (EID). It has been known that information entropy can reflect the positive information content of a given data set. Thus we measure the information content of each class and obtain the differences among them, i.e., EID. In order to minimize EID to balance the data set in information content, an entropy-based hybrid sampling (EHS) approach is proposed, combining both entropy-based oversampling (EOS) and entropy-based undersampling (EUS) methods. For each original instance, we evaluate its information influence degree and remove majority instances with less information using EUS. For each synthetic minority instance, we measure if it will decrease the class entropy and only retain the qualified instance using EOS. This strategy can efficiently avoid over-fitting as well as over-generation since the introduced instances are efficient and informative to decrease the entropy until a balance is achieved. Finally, we train classifiers with the new synthetic data set. The main contributions are highlighted as follows:

1) We propose a new class imbalance metric, termed entropy-based imbalance degree. EID measures the imbalance of class-wise information contents based on their inter-class and intra-class distributions, providing a new view on the imbalance degree in imbalanced learning.
2) We develop three entropy-based sampling (i.e., oversampling, undersampling, and hybrid sampling) approaches based on the proposed EID for multi-class imbalanced learning: EOS oversamples the minorities; EUS undersamples the majorities; EHS jointly oversamples the minorities and undersamples the majorities.
3) We generate new minority instances with information around difficult-to-learn instances, and remove original majority instances with negligible information to achieve a balance between classes.

The rest of this paper is organized as follows: Section 2 reviews background on information entropy. In Section 3, we give the definitions of the proposed metric EID to measure the multi-class imbalance degree and develop three entropy-based sampling approaches for imbalanced learning. Section 4 presents the experiments on synthetic and real-world data sets. In Section 5, we provide the conclusion.

2 Background

The entropy gives a measure of uncertainty about the actual structure of a system [26]. It can be useful to characterize the information content in diverse modes and applications of various fields [27]. In information theory, the major goal for a transmitter is to convey some messages to a receiver. The "information content" of one message measures how much it resolves the uncertainty for the receiver. Generally, the information content can be considered as how much effective information the message actually contains. While in this context, the information entropy by definition is the expected average information content contained in each message. That is to say, the entropy can be viewed as how much effective information the message expects to contain. Thus, there is a positive relationship between information content and entropy [28]. The larger the entropy, the more uncertainty, the more possibilities, the more information content.

A discrete stochastic variable X is given with its all possible outcomes $\{x_1, x_2, \ldots, x_n\}$ and probability mass function $P(X)$. The information content of an outcome specifying $X = x_i$ is defined as follows:

$$I(x_i) = log_a(\frac{1}{p_i}), \tag{1}$$

where p_i is the probability of x_i in X, and a is the base of the logarithm, which is commonly assigned the values of 2, e, or 10. The information entropy of a message is defined as the expected average amount of information to be conveyed about X

$$\begin{aligned} H(X) &= E[I(X)] = E[-log_a(P(X))] \\ &= -\sum_{i=1}^{n} p_i I(x_i) = -\sum_{i=1}^{n} p_i log_a p_i, \end{aligned} \tag{2}$$

where E is the expected value operator. It can be seen that $H(X)$ is positive. Its value reaches the maximum $log_a(n)$ when X has consistent distribution, i.e., $p_i = 1/n$ $(i = 1, \ldots, n)$. Its value reaches the minimum 0 when X has certain distribution, i.e., $P(X) = 1$. In the case of $p_i = 0$, the value of $0log_a 0$ is taken to be 0. Above all, information entropy follows the fundamental properties of information: 1) continuity, i.e., small change in p_i only induces small change in the entropy $H(X)$; 2) symmetry, i.e., $H(X)$ is the same if all outcomes are re-ordered; 3) maximum, i.e.,

$$H_n(p_1, \ldots, p_n) \leq H_n(\frac{1}{n}, \ldots, \frac{1}{n}) = log_a(n). \tag{3}$$

For equiprobable outcomes, $H(X)$ increases with the number of outcomes

$$\begin{aligned} &log_a(n) < log_a(n+1) \\ &H_n(p_1,\ldots,p_n) < H_{n+1}(\frac{1}{n+1},\ldots,\frac{1}{n+1}). \end{aligned} \tag{4}$$

4) non-negativity, i.e., $H(X) \geq 0$; 5) $H(X)$ remains constant when adding or removing an outcome with zero probability

$$H_n(p_1,\ldots,p_n) = H_{n+1}(p_1,\ldots,p_n,0). \tag{5}$$

Relative entropy, known as the Kullback-Leibler divergence (KLD), is another useful measure of entropy of a data distribution [29], [30], [31]. It is often used to evaluate the difference between two non-negative functions or probability distributions. Assume $P(X)$ is the real distribution of X, and $Q(X)$ is the approximate distribution of X. $H(X)$ is the expected average information content used to represent X coinciding with $P(X)$. If we represent X in terms of $Q(X)$, expected additional information content is required. It is measured by KLD

$$D_{KL}(P||Q) = \sum_{i=1}^{n} P(x_i) log_a \frac{P(x_i)}{Q(x_i)}, \tag{6}$$

where $D_{KL}(P||Q) \geq 0$. It measures the difference between $P(X)$ and $Q(X)$. The nearer $Q(X)$ approximates to $P(X)$, the smaller $D_{KL}(P||Q)$ is. When $Q(X) = P(X)$, $D_{KL}(P||Q) = 0$. In addition, unlike $H(X)$, it is asymmetric, i.e., $D_{KL}(P||Q) \neq D_{KL}(Q||P)$.

3 Proposed Method

For a given multi-class imbalanced dataset, the first priority is to determine imbalance degree between the multi-majorities and the multi-minorities. Most sampling approaches use imbalance-ratio as the metric of class imbalance because of its simplicity. However, it is not an informative measure for multi-class problems. On one hand, it just describes class imbalance based on the largest class and the smallest class without considering other classes. On the other hand, the multi-class imbalance may still exist even with a balance in size. As stated in previous works [24], the number of representative (effective) minority instances, rather than that of overall minority instances, decides the classification accuracy for minority classes. Therefore, IR is inappropriate to be considered as the measure of class imbalance. In this section, we propose a novel metric to measure the class imbalance, termed entropy-based imbalance degree, instead of imbalance-ratio. In this case, we first measure the importance of instances and classes [32], [33], and then present three entropy-based sampling approaches: entropy-based oversampling approach, entropy-based undersampling approach, and entropy-based hybrid sampling approach.

3.1 EID: Entropy-Based Imbalance Degree

In information theory, entropy is defined to measure the expected average amount of information contained within a data set. It is generally used as the metric of information content. When a data set has a more entropy, the instances of this data set carry more "information" than that of another data set with a lower entropy, i.e., this data set is more uncertain, and vice verse. Therefore, entropy is a good representation of the amount of intra-class information. Moreover, KLD measures the difference between any two probability distributions. In this case, we introduce it to measure the difference of two information content, and propose a new metric, termed entropy-based imbalance degree instead of IR. Especially, the only way to eliminate the uncertainty from the outside is to introduce effective information. In this case, we propose three methods to balance the information contents among multi-classes.

The definitions of EID are as follows. The dataset D is composed of N instances $X = \{x_1, x_2, \ldots, x_N\}$ with $x_i \in R^d$, where there are m classes $C = \{c_1, c_2, \ldots, c_m\}$ with the corresponding number of instances within each class $\{N_1, N_2, \ldots, N_m\}$ ($N = N_1 + N_2 + \cdots + N_m$). In this paper, we choose a, the base of logarithm, to 2.

Definition 1 (Instance-Wise Statistic). *For each instance, $x_i \in c_r$ ($i = 1,\ldots,N$ and $r = 1,\ldots,m$), the density-based instance-wise statistic of x_i, denoted by $\lambda(x_i)$, is the inverse of average distance between x_i and its intra-class t_i neighbors*

$$\lambda(x_i) = \begin{cases} \frac{1}{t_i}\sum_{l=1}^{t_i} \frac{1}{dist(x_i, Q(x_i)_l)}, & \text{if } 0 < t_i \leq k \\ 0, & \text{if } t_i = 0 \end{cases}, \tag{7}$$

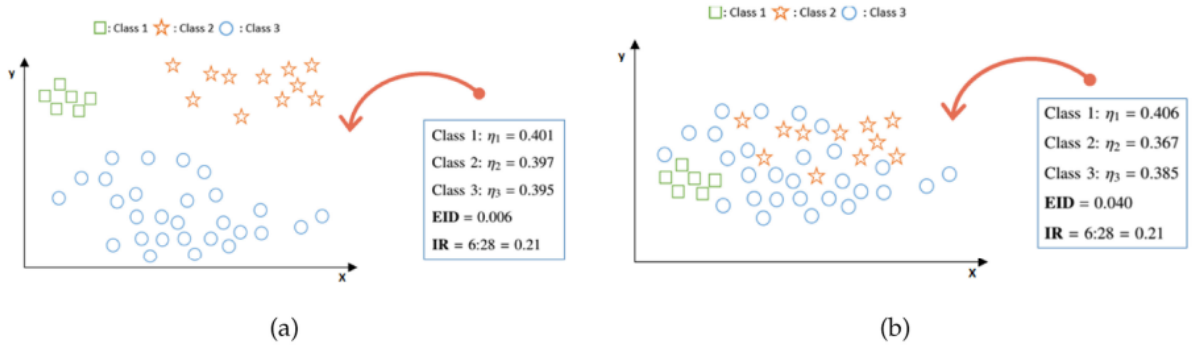
where t_i ($0 \leq t_i \leq k$) is the number of nearest intra-class neighbors in k nearest neighbors of x_i; $Q(x_i) \subseteq KNN(x_i)$, where $KNN(x_i)$ is a set of instances including k nearest neighbors of x_i over the whole data set, and $Q(x_i)$ is a subset of $KNN(x_i)$ including t_i nearest intra-class neighbors of x_i; $dist(x_i, Q(x_i)_l)$ measures the distance between x_i and lth ($l = 1, 2, \ldots, t_i$) nearest intra-class neighbors. We notice that when $t_i \in [1, k]$ the denominator is the distance from x_i to $Q(x_i)$ without other class instances, and the mean value of the opposite of the denominator is used to describe the density of x_i, which could effectively reflect the distribution of x_i. Smaller value of the denominator comes the denser intra-class distribution around x_i. The more likely x_i is a core instance with a larger instance-wise statistic. $t_i = 0$ indicates the k nearest neighbors of x_i all belong to other classes, and then x_i must be an outlier. Considering intra-class t_i neighbors of instances in their KNN benefits the evaluation of intra-class and inter-class distributions.

Definition 2 (Class-Wise Statistic). *For each class, the entropy-based class-wise statistic of $c_r \in C (r = 1, 2, \ldots, m)$, denoted by θ_r, is the expected average amount of density-based information of c_r*

$$\theta_r = -\frac{1}{N_r}\sum_{j=1}^{N_r} \gamma_j \, log_2 \, \gamma_j \quad s.t. \; \gamma_j = \frac{\lambda(x_j)}{\sum_{j=1}^{N_r} \lambda(x_j)}, \tag{8}$$

where N_r is the number of instances in class c_r, and γ_j is a percentage of overall density metric for x_j in class c_r, which can be viewed as the probability of x_j in c_r. It can be known that for class c_r, the lower entropy is, the less uncertainty is, the more density-based information content carries. Note that θ_r measures the expected intra-class density-based information content.

Definition 3 (Instance-Wise Difference Statistic). *There are two density-based information contents, θ_r and ϑ_r^i, where θ_r is real expected average information content of class c_r, and ϑ_r^i*



(a) (b)

Fig. 1. Two data sets both have three classes ($k = 5$): Class 1 (with six instances), Class 2 (with 11 instances), and Class 3 (with 28 instances) with the same IR = 6 : 28 = 0.21 and intra-class distributions but different EIDs. η_1, η_1, and η_3 represent the required average additional information contents of the three classes, respectively.

is an approximate one of class c_r with lack of instance x_i and its t_i intra-class neighbors. For any instance in class c_r ($r = 1, 2, \ldots, m$), the required average information content of class c_r to measure it using θ_r is given by

$$\upsilon(\theta_r) = -\theta_r \; log_2 \; \theta_r. \tag{9}$$

If we use ϑ_r^i to represent the instance from θ_r, the required average information content is evaluated using the following formula:

$$\upsilon(x_i|\theta_r, \vartheta_r^i) = -\theta_r \; log_2 \; \vartheta_r^i, \tag{10}$$

$\upsilon(x_i|\theta_r, \vartheta_r^i)$ measures the difference between θ_r and ϑ_r^i. However, $\upsilon(x_i|\theta_r, \vartheta_r^i) \neq \upsilon(x_i|\vartheta_r^i, \theta_r)$, which are asymmetric. In fact, $\upsilon(x_i|\theta_r, \vartheta_r^i) \geq \upsilon(\theta_r)$ is true with respect to Gibbs' inequality [34]. Moreover, $\upsilon(x_i|\theta_r, \vartheta_r^i) = \upsilon(\theta_r)$ when ϑ_r^i is real expected average information content of c_r, i.e., $\vartheta_r^i = \theta_r$. Then the difference between $\upsilon(x_i|\theta_r, \vartheta_r^i)$ and $\upsilon(\theta_r)$ is given by:

$$\begin{aligned}\delta(\theta_r||\vartheta_r^i) &= \upsilon(x_i|\theta_r, \vartheta_r^i) - \upsilon(\theta_r) \\ &= -\theta_r \; log_2 \; \vartheta_r^i - (-\theta_r \; log_2 \; \theta_r) \\ &= \theta_r \; log_2 \frac{\theta_r}{\vartheta_r^i},\end{aligned}$$

i.e.,

$$\delta(\theta_r||\vartheta_r^i) = \frac{-\sum_{j=1}^{N_r} \gamma_j \; log_2 \; \gamma_j}{N_r} log_2 \frac{-\frac{1}{N_r}\sum_{j=1}^{N_r} \gamma_j \; log_2 \; \gamma_j}{-\frac{1}{N_r - t_i}\sum_{j=1, j\neq L_i}^{N_r} \gamma_j^i \; log_2 \; \gamma_j^i} \tag{11}$$

$$s.t. \; \gamma_j = \frac{\lambda(x_j)}{\sum_{j=1}^{N_r} \lambda(x_j)}, \quad \gamma_j^i = \frac{\lambda(x_j)}{\sum_{j=1, j\neq L_i}^{N_r} \lambda(x_j)},$$

where L_i is the set of subscripts including x_i and $Q(x_i)$. $\delta(\theta_r||\vartheta_r^i)$ indicates the required additional information content using ϑ_r^i to represent θ_r in class c_r. If $\vartheta_r^i = \theta_r$, i.e., these two statistics are identical as well as x_i and $Q(x_i)$ does not contain any information, $\delta(\theta_r||\vartheta_r^i) = 0$. Additionally, it can be easily known that $\delta(\theta_r||\vartheta_r^i) \neq \delta(\vartheta_r^i||\theta_r)$, i.e., the calculation of instance-wise difference statistic is asymmetric. $\delta(\theta_r||\vartheta_r^i)$ is determined by both the entropy of θ_r ($\upsilon(\theta_r)$) and the expectation of ϑ_r^i in θ_r ($\upsilon(x_i|\theta_r, \vartheta_r^i)$). Thus, $\delta(\theta_r||\vartheta_r^i)$ indicates the information influence of x_i and $Q(x_i)$ for c_r, i.e., the more informative x_i and $Q(x_i)$ is, the more the required additional average information content is, the larger $\delta(\theta_r||\vartheta_r^i)$ is. We apply softmax function to map the values of $\delta(\theta_r||\vartheta_r^i)$ to [0,1]. Then the instance-wise difference statistic of $x_i \in c_r$ for the entire data set, denoted by $\varpi(\theta_r||\vartheta_r^i)$, is as follows:

$$\varpi(\theta_r||\vartheta_r^i) = \frac{e^{\delta(\theta_r||\vartheta_r^i)}}{\sum_{t=1}^{N} e^{\delta(\theta_r||\vartheta_r^t)}}. \tag{12}$$

Definition 4 (Class-Wise Difference Statistic). *For each class, $c_r \in C(r = 1, 2, \ldots, m)$, the class-wise difference statistic of c_r, denoted by η_r, is a required average additional information content for all instances in c_r*

$$\eta_r = \frac{1}{N_r}\sum_{i=1}^{N_r} \varpi(\theta_r||\vartheta_r^i). \tag{13}$$

It can be known that η_r is in [0,1], and measures the average information influence degree of c_r. The less the number of informative instances in c_r, the more η_r, the less information content c_r carries, the easier to learn c_r it is.

Definition 5 (Entropy-based Imbalance degree). *For data set D, entropy-based imbalance degree (EID) is the sum of absolute differences between each and the mean class-wise difference statistics*

$$EID = \frac{1}{m}\sum_{r=1}^{m} |\eta_r - \xi| \;\; s.t. \; \xi = \frac{1}{m}\sum_{h=1}^{m} \eta_h. \tag{14}$$

It can be shown that $EID \in$ [0,1], and $EID = 0$ when class balance is achieved. Our goal is to minimize the imbalance degree for new synthetic data set X_{new} using sampling techniques. Thus the objective function is given by:

$$\{X_{new}\}_{opt} = \arg\min_{X_{new}}(EID). \tag{15}$$

We illustrate the issues of respectively using EID and IR as multi-class imbalance measures with two examples in Fig. 1. In Figs. 1a and 1b, each data set has three classes, in which the sizes of class 1 and 2 (minority classes) are less than that of class 3 (majority class). It can be seen that the three classes in two data sets have the same intra-class distributions and different inter-class distributions, and IR between the smallest class and the largest class is 6 : 28 = 0.21. However, in Fig. 1a, it is shown that there have clear boundaries among the three classes and they can be discriminated with any simple classifier, i.e., the instances of all classes can well represent

their distributions. In Fig. 1b, there are a lot of cross-cutting among the three classes. EID in (b) are greater than that in (a). The values of η are relative due to the relationships among classes. The increases look small due to the use of logarithms and mean values but make a great deal of sense. Thus there would be less informative instances in the three classes, especially in class 2 with a relative larger increase of η. In a word, $IR_1 = IR_2$, while $EID_2 = 6.7\ EID_1$. IR cannot represent the imbalance degree of data sets.

Algorithm 1. EOS: Entropy-Based Oversampling Approach

Input: X : data set, with N instances and m classes, consisting of N_r instances in class c_r.
Output: S: qualified synthetic instances; R: classification results.

1. Calculate the imbalance degree EID^o using Eq. (16).
2. Obtain the minimum of additional information contents $\varphi = \min(\eta)$.
for r = 1:m **do**
  a. Calculate the difference between η_r and φ using $\Delta = \eta_r - \varphi$.
  while $\Delta \geq 0$ **do**
    a. Sample an instance x_i with the maximal $\varpi(\theta_r||\vartheta_r^i)$ in class c_r, and generate a new sample x_g based on x_i using Eq. (17).
    b. Add x_g in c_r: $c_r = \{c_r \cup x_g\}$, $N_r = N_r + 1$, and recalculate Δ^* for c_r.
    if $\Delta^* < \Delta$ **then**
      a. Update $\Delta = \Delta^*$, and add the qualified x_g in S.
    else
      b. Remove x_g from c_r, and reset Δ to previous values.
    end if
  end while
end for
3. Train classifier F with new synthetic data set $X' = X \cup S$, and obtain the classification results R.

3.2 EOS: Entropy-Based Oversampling Approach

Oversampling technique is effective for imbalanced learning, which is devoted to balance skewed data distribution by generating new minority instances. As aforementioned above, a large number of synthetic sample methods have been proposed (e.g., SMOTE and AdaSyn). Motivated by the success of these methods, we present an entropy-based oversampling approach on basis of EID^o metric. We first compute instance-wise statistics $\lambda(x)$ for all instances in a data set and class-wise statistics θ for overall classes using $\lambda(x)$. Then instance-wise difference statistics $\varpi(\theta_r||\vartheta_r)$ for all instances using θ and class-wise difference statistics η for all classes by $\varpi(\theta_r||\vartheta_r)$. By definitions, it can be known that instances and classes with less information and lower $\varpi(\theta_r||\vartheta_r)$ and η, are easier to learn, i.e., they are majority instances and classes. Similarly, those with more information and larger entropy are difficult to learn, i.e., they are minority instances and classes.

We describe EOS detaily in Algorithm 1. The class with the minimal additional information content φ is considered as the maximum majority class. The other classes have to generate new instances with effective information to balance data set on information contents using modified EID metric as follows:

$$\begin{aligned} &\{X_{new}\}_{opt} = \arg\min_{X_{new}}(EID^o) \\ &s.t.\ \ EID^o = \frac{1}{m}\sum_{r=1}^{m}(\eta_r - \varphi)\ \ \varphi = \min(\eta). \end{aligned} \tag{16}$$

For each class, we calculate mean value of the differences between the minimum and each value of class-wise difference statistics. It can be known that $\Delta = \eta - \varphi \geq 0$. Then we use modified SMOTE to generate new informative instances until $\Delta \leq 0$, i.e., this class and the maximum majority class achieve a balance. Finally, all classes form a balanced data set and EID^o tends to 0. In EOS, for each synthetic instance x_g in class c_r, we first sample an instance x_i with the maximal $\varpi(\theta_r||\vartheta_r^i)$ in class c_r, and generate x_g using the following formula:

$$\begin{aligned} &x_g = x_i + (x_\omega - x_i) \circ \alpha \\ &s.t.\ x_\omega = \arg\min_{x_\omega \in Q(x_i)}(\varpi(\theta_r||\vartheta_r^\omega)), \end{aligned} \tag{17}$$

where x_ω is the instance with minimal $\varpi(\theta_r||\vartheta_r)$ of the k-nearest neighbors of x_i in class c_r, the symbol "$\circ$" indicates element-wise multiplication of two vectors, and α is a random vector. Only the qualified instances, which could make Δ decrease, are allowed to add into data set. In this case, we can effectively avoid over-generation and generation of noisy and wrong instances.

Algorithm 2. EUS: Entropy-Based Undersampling Approach

Input: X : data set, with N instances and m classes, consisting of N_r instances in class c_r.
Output: U: removed data set; R: classification results.

1. Calculate the imbalance degree EID^u using Eq. (18).
2. Obtain the maximum of additional information contents $\zeta = \max(\eta)$.
for r = 1:m **do**
  a. Calculate the difference between η_r and φ using $\Delta = \zeta - \eta_r$.
  while $\Delta \geq 0$ **do**
    a. Sample an instance x_i with the minimal $\varpi(\theta_r||\vartheta_r^i)$ in class c_r, add x_i into U and remove x_i from c_r: $c_r = \{c_r - x_i\}$, $N_r = N_r - 1$.
    b. Recalculate Δ for c_r.
  end while
end for
3. Train classifier F with new synthetic data set $X' = X - U$, and obtain the classification results R.

3.3 EUS: Entropy-Based Undersampling Approach

Unlike oversampling technique, undersampling technique attempts to remove a subset of majority instances to form a balanced data set. Since a great deal of useful information may be lost, and the training for classifiers is hard on the subset of data with these under-representative information, it is necessary to implement detection and recognition of easy-to-learn instances, remove them, and retain difficult-to-learn instances.

Entropy-based undersampling approach (EUS) method is summarized in Algorithm 2. EUS can adaptively

determine the easy-to-learn majority instances, i.e., they have lower informational influences ($\varpi(\theta_r||\vartheta_r^i)$), and throw away them until solving the following equation:

$$\begin{aligned} \{X_{new}\}_{opt} &= \arg\min_{X_{new}}(EID^u) \\ s.t.\ \ EID^u &= \frac{1}{m}\sum_{r=1}^{m}(\zeta - \eta_r)\ \ \zeta = \max(\eta), \end{aligned} \tag{18}$$

where EID^u is the mean value of differences between the maximum and each value of class-wise difference statistics. We conduct EUS based on the minimum class, which corresponds maximum class-wise difference statistic. It can be known that a class with large information content needs to be removed redundant information content to eliminate the uncertainty. Using the same calculations as we used in first three steps of EOS, for each class, we calculate Δ ($\Delta \geq 0$), remove easier-to-learn instances until $\Delta \leq 0$. At this point, the given data set is balanced.

Algorithm 3. EHS: Entropy-Based Hybrid Sampling Approach

Input: X : data set, with N instances and m classes, consisting of N_r instances in class c_r.
Output: S: qualified synthetic instances; U: removed data set; R: classification results.

1. Calculate the imbalance degree EID using Eq. (14).
2. Obtain the mean value of additional information contents $\xi = \frac{1}{m}\sum_{r=1}^{m}\eta_r$.
for r = 1:m **do**
  a. Calculate the difference between η_r and ξ using $\Delta = \eta_r - \xi$.
  if $\Delta > 0$ **then**
    while $\Delta > 0$ **do**
      a. Sample an instance x_i with the maximal $\varpi(\theta_r||\vartheta_r^i)$ in class c_r, and generate a new sample x_g based on x_i using Eq. (17).
      b. Add x_g in c_r: $c_r = \{c_r \cup x_g\}$, $N_r = N_r + 1$, and recalculate Δ^* for c_r.
      if $\Delta^* < \Delta$ **then**
        a. Update $\Delta = \Delta^*$, and add the qualified x_g in S.
      else
        b. Remove x_g from c_r, and reset Δ to previous values.
      end if
    end while
  else
    while $\Delta < 0$ **do**
      a. Sample an instance x_j with the minimal $\varpi(\theta_r||\vartheta_r^j)$ in class c_r, add x_j into U and remove x_j from c_r: $c_r = \{c_r - x_j\},\ N_r = N_r - 1$.
      b. Recalculate Δ for c_r.
    end while
  end if
end for
3. Train classifier F with new synthetic data set $X' = X \cup S - U$, and obtain the classification results R.

3.4 EHS: Entropy-Based Hybrid Sampling Approach

Hybrid sampling techniques combine the oversampling and undersampling techniques, adding minority instances and removing majority instances simultaneously in order to eliminate overfitting and prevent the loss of too much information effectively. Especially for multi-class imbalanced learning, if we use the minimum or the maximum of required information contents as measure of imbalance degree, they can raise the problems of overfitting and overlapping using single oversampling techniques as well as missing too much valuable information using single undersampling techniques. Therefore, we propose an entropy-based hybrid sampling approach based on EID metric.

EHS is described in detail in Algorithm 3. Unlike EOS and EUS, EHS uses the mean value of differences between the mean value and each value of class-wise difference statistics. First, we obtain EID using Eq. (14). In this case, we can use EOS to generate informative minority instances for initial $\Delta > 0$ until $\Delta \leq 0$, and use EUS to remove under-representative majority instances for initial $\Delta < 0$ until $\Delta \geq 0$. Entropy is going to be favoring whichever side has mean entropy, and a balance is achieved by minimizing EID. At this point, each class has similar information content, and the new synthetic data set can be as input to train classifiers.

As shown in Algorithm 1, 2, and 3, the proposed EOS, EUS, and EHS are based on different EIDs in terms of minimum, maximum, and mean of required information contents. Notice that the significance of our proposed methods has two aspects. First, we use the instance-wise density distributions $\lambda(x)$ with the reciprocal of mean distances to obtain the class-wise density distributions θ_r. In this case, we measure the intra-class required information contents to represent instances and the required information contents to represent classes. Second, the imbalance degree of a data set is measured by the entropy-based metric of $EIDs$ instead of IR. It is proved that we balance the data set when $EIDs$ tend to 0. Based on $EIDs$, the learning methods in these problems concentrate on difficult-to-learn instances and only allow to add qualified instances to ensure that the new constructed data set has a balanced distribution, improving classification performance.

3.5 Time Complexity Analysis

The EHS is composed of three major steps: 1) obtaining EID of a given data set; 2) initially reducing EID by generating new synthetic and informative minority instances using EOS; 3) further minimizing EID by removing majority instances with less information using EUS. Their time complexities are $O(Nd(log_2N + 1))$, $O(\alpha d m_+(log_2\alpha + 1))$, and $O(\beta d m_-(log_2\beta + 1))$, where N is the total number of original instances; d is the number of attributes; m_+ is the number of majority classes; m_- is the number of minority classes; α is sum of N and the number of generated instances; β is the final number of instances after oversampling and undersamling; $m = m_+ + m_-$; $\alpha \geq N$; $\alpha \geq \beta$. The overall time complexity of EHS is $O(Ndm(log_2N + 1))$. EHS has a high time complexity, but it yields significantly good results.

4 Experimental Results

4.1 Performance Evaluation Metrics

Imbalanced learning attempts to improve the classification performance for the minorities. As mentioned above, the overall accuracy is not a good performance evaluation metric since the majorities dominate the overall dataset. In this

paper, we use four metrics, Precision, Recall, F-Measure [35], and G-Mean [36], to evaluate the performance of classifiers. The definitions of these metrics are as follows:

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Measure = \frac{2 \times Recall \times Precision}{Recall + Precision}$$

$$G - Mean = \sqrt{Recall \times \frac{TN}{TN + FP}},$$

where true positive (TP) is the number of minority instances with correct classification; false positive (FP) is the number of majority instances with wrong classification; true negative (TN) is the number of majority instances which are classified correctly; false negative (FN) is the number of minority instances which are misclassified as the majorities.

4.2 Experimental Settings

To verify the effectiveness of the proposed EOS, EUS, and EHS methods, we carry on extensive experiments on two 2D data sets and 12 real-world data sets. 2D data sets named Spiral and Irregular are shown in the original set of Figs. 2a and 2b, which are chosen to take arbitrary-shaped and non-Gaussian data distributions. Additionally, the statistics of real data sets are summarized in Table 1, where first 8 data sets (vehicle1, segment0, page-blocks0 abbreviated as page-bk0, penbased, yeast, thyroid, shuttle, and ecoli) come from KEEL repository [37], and the other 4 data sets (msplice, letter, waveform3 abbreviated as wavefm3, and landsat) are available from UCI repository [38]. The proportions of the majorities and minorities are shown both for the binary-class and multi-class data sets. IR is the traditional overall imbalanced measure and EID is our proposed imbalanced degree. For each data set, we perform 5-fold cross validation where the original data set is randomly divided into 5 folds. Each fold is used for testing once while the remaining 4 folds are trained. In each fold, all classification methods are trained 10 times and the results are averaged over 10 runs in order to eliminate the randomness.

We select two common used base classifiers, including AdaBoost and Multilayer perceptrons (MLP). The parameters are described as follows: AdaBoost uses 100 boosting iterations; MLP is trained to 100 epoches with a learning rate of 0.1 and 10 hidden layer neurons.

In detail, the performance of our proposed approaches are compared with a number of state-of-the-art imbalanced learning techniques. We summarize these 7 approaches as follows: ADASYN, SMOTE, MWMOTE, SMOTE + Tomek links (abbreviated as SMTL), SMOTE + ENN (abbreviated as SMENN), EasyEnsemble (abbreviated as EASY), and BalanceCascade. All these methods are built with Tensorflow 0.8, and implemented on an Intel i7-6700 CPU and a single Nvidia TITAN Xp GPU with Python 2.7. The euclidean distance is used to measure the distance between instances. In oversampling techniques, all classes are oversampled until they have the same number of instances with the largest one. With respect to the parameter settings, the number of nearest neighbors is 5. The default parameter values are used for all the other compared approaches [39].

4.3 2-D Data Sets

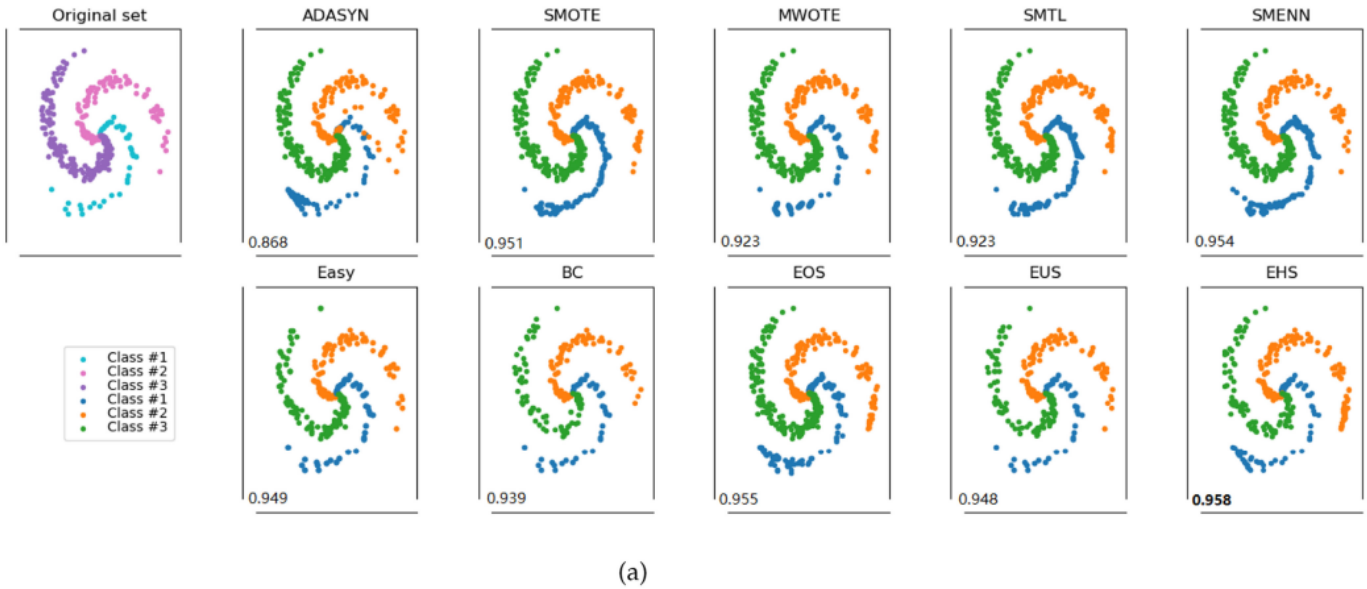
As shown in Figs. 2a and 2b, all the synthetic samples are visualized with 2D scatter plots for comparing the performance of different sampling algorithms ($k = 5$ for all the numbers of nearest neighbors). In Fig. 2a, the blue and orange dots symbolize synthetic minority class samples, respectively from class 1 and 2, corresponding to the cyan and pink dots of original set. The green dots are synthetic majority class samples from class 3 corresponding to the purple dots of original set. From the values of G-M, it can be seen that EHS has achieved the best performance compared with other sampling techniques. Fig. 2b show the results on Irregular data set. The blue and red dots symbolize synthetic minority class samples, respectively from class 1 and 4, corresponding to the cyan and gray dots of original set. The orange and green dots represent synthetic majority class samples, which correspond to the pink and purple dots of original set, respectively from class 2 and 3. The effectiveness of our proposed three methods is measured by the G-M of AdaBoost classifier, which is trained using the original data and is tested with the synthetic data. The G-Ms of different sampling methods are shown in the bottom left corner of plots. Our EHS method also outperforms other compared methods.

4.4 Multidimensional Data Sets

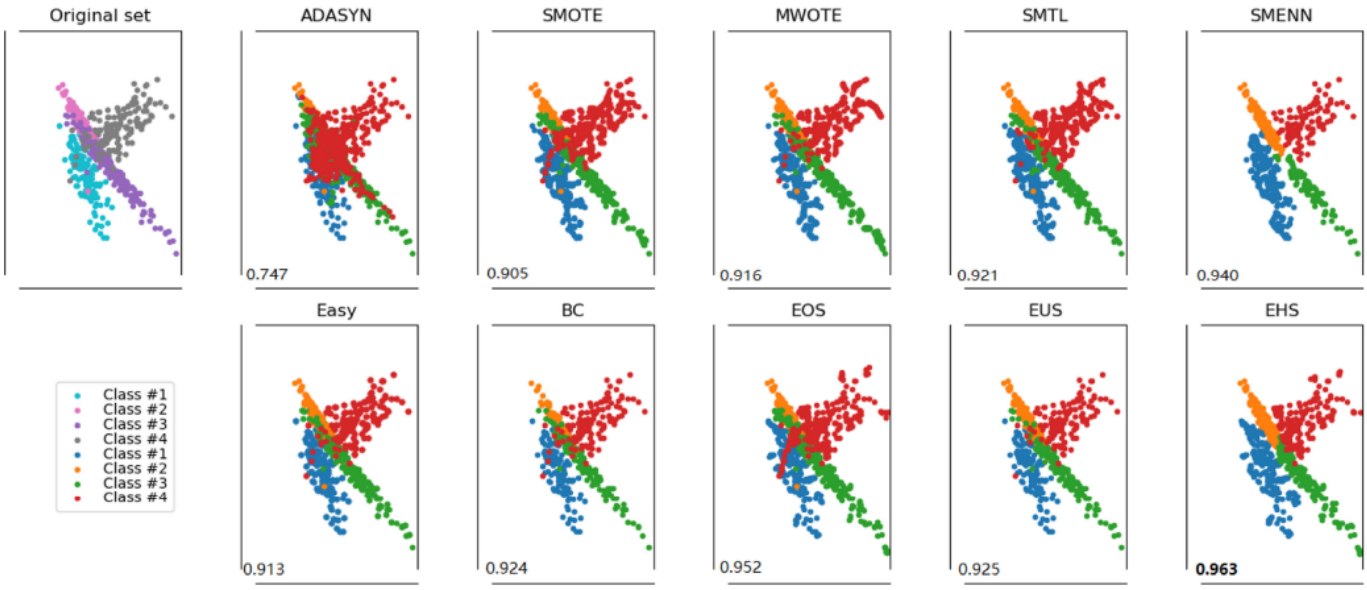
In Table 1, two measures of class imbalance degree are presented over all the data sets: modified-imbalance-ratio (MIR) and entropy-based imbalance degree. The MIR denotes the ratio of the number of instances from multi-majority classes to the number of instances from multi-minority classes, which is more representative than IR by considering all the classes. The EID defined in this paper are different in EOS, EUS, and EHS. Table 1 shows the EID of EHS. While MIR still cannot reflect the class imbalance degrees of imbalanced data sets, EID is a new view to measure the class imbalance from information content. For example, the *shuttle* data set (853) is far more imbalanced than the *penbased* data set according to MIR, while the *penbased* data set (0.211) is more imbalanced than the *shuttle* data set (0.008) according to *EID*.

Tables 2, 3, 4, and 5 summarize the average performance results respectively in precision, recall, F-measure, and G-mean of 7 compared sampling methods and our 3 methods using two base classifiers for MLP and AdaBoost over all data sets. These four tables show the average values and corresponding standard deviations. The best results are highlighted in bold. From these tables, it can be seen that our proposed methods, especially EHS, acquire better performance than the others.

In terms of precision, Table 2 shows that our proposed approaches achieve the best performance for all 12 data sets. From Table 3, our proposed methods outperform the others for 11 out of 12 data sets in terms of recall. We can see that the proposed three methods perform better for 10 out of 12 data sets in terms of F-measure in Table 4. For G-mean metric, they achieve the best performance for 11 out of 12 data sets in



(a)



(b)

Fig. 2. Scatter plots of the original data set and synthetic data of (a) Spiral data set and (b) Irregular data set generated by ADASYN, SMOTE, MWOTE, SMTL, SMENN, ESAY, BC, EOS, EUS, and EHS. In (a), the green dots from class 3 represent synthetic majority class samples corresponding to the purple dots in original set; the blue and orange dots, respectively, from class 1 and 2 represent synthetic minority class samples corresponding to the cyan and pink dots in original set. In (b), the blue and red dots, respectively, from class 1 and 4 represent synthetic minority class samples corresponding to the cyan and gray dots in original set; the orange and green dots from class 2 and 3 represent synthetic majority class samples corresponding to the pink and purple dots in original set. The values of G-M are shown in bottom left corner of plots. The bold G-Ms are the best performances.

Table 5. We also notice that different methods have different standard deviations. In order to further demonstrate the effectiveness of our proposed three methods, Welch's t-test is performed to evaluate whether EOS, EUS, and EHS can significantly outperform the other methods. The Welch's t-test takes both mean values and standard deviations into account unlike Student's t-test, which defines the statistic t if two samples have the same size n as follows:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{n}}}, \tag{19}$$

where $\overline{X_1}$ and S_1^2 are first sample mean and variance, respectively; $\overline{X_2}$ and S_2^2 are second sample mean and variance, respectively. The statistic t follows the t-distribution. The Welch Satterthwaite equation is used to estimate the degree of freedom v associated with this variance

$$v \approx \frac{(S_1^2 + S_2^2)^2 (n-1)}{S_1^4 + S_2^4}. \tag{20}$$

We summarize the t-test of Precision, Recall, F-Measure, and G-Mean between our proposed three methods and all the methods with a significant level at 0.05 in Table 6. In each row, each of our proposed method is compared with the remaining methods, and the amounts of win-tie-lose are presented over 12 data sets.

TABLE 1
Statistics of Experimental Data Sets

Datasets	#Instances	#Features	#Classes	% majorities	% minorities	MIR	EID
vehicle1	846	18	2	74.36	25.64	2.90	0.112
segment0	2308	19	2	85.75	14.25	6.02	0.014
page-bk0	5472	10	2	89.79	10.21	8.79	0.051
penbased	1100	16	10	66.1	33.90	1.95	0.211
yeast	1484	8	10	95.86	4.14	23.15	0.114
thyroid	720	21	3	97.36	2.64	36.94	0.149
shuttle	2175	9	7	99.88	0.12	853	0.008
ecoli	336	7	5	91.37	8.63	10.59	0.146
msplice	3175	240	3	51.91	48.09	1.08	0.019
letter	5000	16	26	49.02	50.98	0.96	0.208
wavefm3	5000	21	3	67.06	32.94	2.04	0.003
landsat	2000	36	6	66.40	33.60	1.98	0.040

TABLE 2
Averages of Precision by Different Methods on 12 Data Sets Using MLP and Adaboost

Datasets	ADASYN	SMOTE	MWMOTE	SMTL	SMENN	Easy	BC	EOS	EUS	EHS
vehicle1	0.601 ± 0.17	0.706 ± 0.11	0.746 ± 0.04	0.717 ± 0.11	0.702 ± 0.12	0.734 ± 0.07	0.796 ± 0.04	0.821 ± 0.05	0.796 ± 0.03	**0.839 ± 0.08**
segment0	0.995 ± 0.00	0.954 ± 0.10	0.995 ± 0.00	0.993 ± 0.00	0.992 ± 0.00	0.946 ± 0.11	0.990 ± 0.00	0.995 ± 0.01	0.990 ± 0.01	**0.998 ± 0.00**
page-bk0	0.956 ± 0.02	0.954 ± 0.02	0.954 ± 0.02	0.954 ± 0.01	0.957 ± 0.01	0.947 ± 0.02	0.947 ± 0.03	0.963 ± 0.02	0.965 ± 0.00	**0.978 ± 0.01**
penbased	0.756 ± 0.12	0.763 ± 0.12	0.751 ± 0.10	0.778 ± 0.09	0.751 ± 0.11	0.819 ± 0.11	0.921 ± 0.00	0.948 ± 0.01	0.937 ± 0.01	**0.953 ± 0.02**
yeast	0.514 ± 0.10	0.536 ± 0.07	0.538 ± 0.07	0.501 ± 0.10	0.517 ± 0.07	0.383 ± 0.09	0.499 ± 0.06	0.542 ± 0.06	0.547 ± 0.02	**0.592 ± 0.07**
thyroid	0.944 ± 0.04	0.920 ± 0.06	0.917 ± 0.07	0.918 ± 0.07	0.919 ± 0.06	0.931 ± 0.04	0.957 ± 0.00	0.972 ± 0.01	0.978 ± 0.01	**0.983 ± 0.01**
shuttle	0.995 ± 0.03	0.993 ± 0.01	0.994 ± 0.01	0.990 ± 0.01	0.991 ± 0.01	0.869 ± 0.04	0.958 ± 0.03	0.994 ± 0.02	**0.996 ± 0.01**	0.994 ± 0.01
ecoli	0.709 ± 0.16	0.647 ± 0.15	0.646 ± 0.17	0.644 ± 0.17	0.649 ± 0.16	0.658 ± 0.15	0.695 ± 0.19	**0.903 ± 0.12**	0.867 ± 0.17	0.889 ± 0.16
msplice	0.932 ± 0.01	0.941 ± 0.01	0.940 ± 0.01	0.938 ± 0.01	0.940 ± 0.01	0.934 ± 0.01	0.950 ± 0.00	0.952 ± 0.01	0.955 ± 0.01	**0.964 ± 0.01**
letter	0.556 ± 0.19	0.552 ± 0.21	0.566 ± 0.19	0.571 ± 0.14	0.574 ± 0.14	0.574 ± 0.17	0.741 ± 0.01	**0.862 ± 0.02**	0.860 ± 0.01	0.855 ± 0.01
wavefm3	0.851 ± 0.01	0.848 ± 0.01	0.853 ± 0.01	0.851 ± 0.01	0.853 ± 0.01	0.854 ± 0.01	0.816 ± 0.02	0.856 ± 0.02	0.842 ± 0.02	**0.881 ± 0.02**
landsat	0.718 ± 0.09	0.721 ± 0.07	0.737 ± 0.07	0.744 ± 0.06	0.732 ± 0.05	0.667 ± 0.18	0.808 ± 0.03	0.853 ± 0.03	0.855 ± 0.04	**0.872 ± 0.03**

TABLE 3
Averages of Recall by Different Methods on 12 Data Sets Using MLP and Adaboost

Datasets	ADASYN	SMOTE	MWMOTE	SMTL	SMENN	Easy	BC	EOS	EUS	EHS
vehicle1	0.611 ± 0.14	0.766 ± 0.04	0.755 ± 0.05	0.745 ± 0.06	0.746 ± 0.07	0.657 ± 0.11	0.777 ± 0.05	0.793 ± 0.06	0.810 ± 0.05	**0.836 ± 0.09**
segment0	**0.994 ± 0.00**	0.974 ± 0.06	0.995 ± 0.00	0.993 ± 0.00	0.992 ± 0.00	0.959 ± 0.07	0.990 ± 0.00	0.993 ± 0.00	0.991 ± 0.00	0.991 ± 0.00
page-bk0	0.943 ± 0.02	0.953 ± 0.02	0.954 ± 0.02	0.953 ± 0.02	0.957 ± 0.01	0.920 ± 0.04	0.911 ± 0.09	0.959 ± 0.03	0.964 ± 0.02	**0.972 ± 0.03**
penbased	0.732 ± 0.13	0.746 ± 0.14	0.726 ± 0.14	0.764 ± 0.12	0.733 ± 0.13	0.798 ± 0.12	0.808 ± 0.00	0.895 ± 0.05	0.868 ± 0.09	**0.933 ± 0.03**
yeast	0.464 ± 0.10	0.468 ± 0.11	0.462 ± 0.10	0.442 ± 0.12	0.445 ± 0.11	0.336 ± 0.05	0.451 ± 0.04	**0.476 ± 0.08**	0.468 ± 0.01	0.472 ± 0.07
thyroid	0.891 ± 0.07	0.950 ± 0.03	0.948 ± 0.03	0.948 ± 0.03	0.949 ± 0.03	0.958 ± 0.08	0.950 ± 0.01	**0.972 ± 0.02**	0.957 ± 0.03	0.959 ± 0.02
shuttle	0.996 ± 0.00	0.994 ± 0.04	0.993 ± 0.00	0.991 ± 0.01	0.992 ± 0.01	0.969 ± 0.02	0.947 ± 0.05	0.996 ± 0.04	0.995 ± 0.02	**0.998 ± 0.01**
ecoli	0.567 ± 0.12	0.583 ± 0.14	0.585 ± 0.14	0.567 ± 0.14	0.579 ± 0.13	0.394 ± 0.10	0.577 ± 0.12	0.702 ± 0.03	0.669 ± 0.05	**0.715 ± 0.06**
msplice	0.935 ± 0.01	0.940 ± 0.01	0.939 ± 0.01	0.938 ± 0.10	0.939 ± 0.01	0.932 ± 0.01	0.949 ± 0.00	0.956 ± 0.01	**0.961 ± 0.04**	0.949 ± 0.01
letter	0.537 ± 0.15	0.542 ± 0.11	0.548 ± 0.13	0.557 ± 0.11	0.555 ± 0.18	0.550 ± 0.19	0.735 ± 0.01	0.853 ± 0.02	0.861 ± 0.01	**0.880 ± 0.01**
wavefm3	0.849 ± 0.01	0.847 ± 0.01	0.851 ± 0.01	0.850 ± 0.01	0.852 ± 0.01	0.854 ± 0.01	0.815 ± 0.01	0.870 ± 0.02	0.852 ± 0.01	**0.879 ± 0.02**
landsat	0.620 ± 0.13	0.621 ± 0.14	0.676 ± 0.07	0.635 ± 0.13	0.664 ± 0.08	0.578 ± 0.15	0.766 ± 0.10	0.827 ± 0.04	**0.835 ± 0.04**	0.814 ± 0.02

The results show that both oversampling and undersampling techniques in those imbalanced learning methods exhibit their specific advantages. For some data sets, EUS could be more effective than EOS such as in page-bk0, yeast, and landsat, and less effective than EOS such as in vechicle1, segment0, ecoli, and wavefm3. The EHS, on average, outperforms the other methods.

In a word, the experimental results on synthetic and real-world data sets not only show the superiority of our proposed three methods over other compared methods, but also show the superiority of EHS over EOS and EUS. In Figs. 2a and 2b, EHS respectively achieves the highest G-Mean values and performs better than EOS and EUS. Its superiority lies in that EHS generates less minority class samples than EOS to avoid overlapping and simultaneously removes less majority class samples than EUS to avoid information loss. In Tables 2 and 4, EHS achieves better performance than EOS in 10 out of 12 and EUS in 11 out of 12. In Tables 3 and 5, EHS performs better than EOS in 10 out of 12 and EHS in 10 out of 12. Furthermore, Table 6 shows that EHS outperforms EOS and EUS on average based on the four performance evaluation metrics: Precision, Recall,

TABLE 4
Averages of F-Measure by Different Methods on 12 Data Sets Using MLP and Adaboost

Datasets	ADASYN	SMOTE	MWMOTE	SMTL	SMENN	Easy	BC	EOS	EUS	EHS
vehicle1	0.586 ± 0.17	0.724 ± 0.07	0.741 ± 0.05	0.725 ± 0.08	0.712 ± 0.09	0.678 ± 0.10	0.783 ± 0.05	0.774 ± 0.06	0.790 ± 0.07	**0.795 ± 0.04**
segment0	0.992 ± 0.00	0.963 ± 0.08	**0.996 ± 0.01**	0.994 ± 0.00	0.993 ± 0.00	0.953 ± 0.09	0.990 ± 0.00	0.993 ± 0.02	0.990 ± 0.00	0.995 ± 0.01
page-bk0	0.947 ± 0.02	0.953 ± 0.02	0.952 ± 0.02	0.952 ± 0.02	0.956 ± 0.01	0.928 ± 0.03	0.911 ± 0.03	0.955 ± 0.00	**0.963 ± 0.04**	0.960 ± 0.03
penbased	0.716 ± 0.15	0.734 ± 0.12	0.710 ± 0.16	0.758 ± 0.13	0.718 ± 0.15	0.795 ± 0.13	0.918 ± 0.01	0.943 ± 0.03	0.931 ± 0.01	**0.961 ± 0.01**
yeast	0.461 ± 0.10	0.465 ± 0.09	0.466 ± 0.09	0.439 ± 0.11	0.448 ± 0.09	0.397 ± 0.06	0.451 ± 0.04	0.480 ± 0.01	0.496 ± 0.03	**0.616 ± 0.07**
thyroid	0.911 ± 0.07	0.932 ± 0.05	0.931 ± 0.05	0.932 ± 0.05	0.933 ± 0.05	0.916 ± 0.09	0.950 ± 0.00	0.955 ± 0.04	0.964 ± 0.02	**0.975 ± 0.01**
shuttle	**0.995 ± 0.00**	0.993 ± 0.00	0.994 ± 0.01	0.990 ± 0.01	0.991 ± 0.01	0.952 ± 0.01	0.948 ± 0.03	0.994 ± 0.02	0.993 ± 0.01	0.994 ± 0.01
ecoli	0.623 ± 0.13	0.600 ± 0.15	0.602 ± 0.16	0.591 ± 0.16	0.602 ± 0.15	0.454 ± 0.12	0.626 ± 0.15	**0.776 ± 0.12**	0.753 ± 0.15	0.772 ± 0.13
msplice	0.936 ± 0.01	0.941 ± 0.01	0.939 ± 0.01	0.938 ± 0.01	0.939 ± 0.01	0.932 ± 0.01	0.949 ± 0.01	0.951 ± 0.02	0.961 ± 0.01	**0.963 ± 0.02**
letter	0.524 ± 0.12	0.522 ± 0.13	0.544 ± 0.12	0.542 ± 0.11	0.544 ± 0.13	0.539 ± 0.15	0.735 ± 0.01	0.853 ± 0.01	0.832 ± 0.01	**0.918 ± 0.01**
wavefm3	0.849 ± 0.01	0.847 ± 0.01	0.852 ± 0.01	0.850 ± 0.01	0.852 ± 0.01	0.853 ± 0.01	0.815 ± 0.01	**0.898 ± 0.02**	0.887 ± 0.02	0.875 ± 0.01
landsat	0.629 ± 0.13	0.623 ± 0.11	0.677 ± 0.09	0.645 ± 0.12	0.678 ± 0.07	0.695 ± 0.15	0.767 ± 0.08	0.831 ± 0.04	0.942 ± 0.04	**0.958 ± 0.03**

TABLE 5
Averages of G-Mean by Different Methods on 12 Data Sets Using MLP and Adaboost

Datasets	ADASYN	SMOTE	MWMOTE	SMTL	SMENN	Easy	BC	EOS	EUS	EHS
vehicle1	0.464 ± 0.17	0.444 ± 0.16	0.581 ± 0.09	0.531 ± 0.12	0.443 ± 0.14	0.646 ± 0.12	0.737 ± 0.04	0.793 ± 0.07	0.789 ± 0.05	**0.798 ± 0.03**
segment0	0.993 ± 0.00	0.830 ± 0.10	0.994 ± 0.00	0.995 ± 0.00	0.994 ± 0.01	0.842 ± 0.13	0.989 ± 0.00	**0.997 ± 0.01**	0.993 ± 0.01	0.996 ± 0.00
page-bk0	0.907 ± 0.03	0.845 ± 0.07	0.834 ± 0.08	0.810 ± 0.11	0.842 ± 0.06	0.888 ± 0.08	0.909 ± 0.03	0.909 ± 0.02	**0.957 ± 0.02**	0.947 ± 0.02
penbased	0.796 ± 0.10	0.811 ± 0.11	0.792 ± 0.13	0.832 ± 0.16	0.797 ± 0.19	0.873 ± 0.09	0.892 ± 0.00	0.967 ± 0.01	0.932 ± 0.01	**0.975 ± 0.02**
yeast	0.605 ± 0.09	0.602 ± 0.07	0.601 ± 0.07	0.577 ± 0.09	0.586 ± 0.08	0.484 ± 0.07	0.606 ± 0.04	0.618 ± 0.01	0.624 ± 0.04	**0.672 ± 0.07**
thyroid	0.786 ± 0.14	0.492 ± 0.16	0.485 ± 0.15	0.486 ± 0.14	0.497 ± 0.16	0.756 ± 0.13	0.869 ± 0.03	0.972 ± 0.03	0.957 ± 0.01	**0.975 ± 0.02**
shuttle	0.992 ± 0.00	0.991 ± 0.01	**0.996 ± 0.01**	0.984 ± 0.02	0.986 ± 0.01	0.953 ± 0.03	0.940 ± 0.00	0.994 ± 0.01	0.990 ± 0.02	0.994 ± 0.01
ecoli	0.399 ± 0.16	0.339 ± 0.16	0.348 ± 0.17	0.344 ± 0.16	0.351 ± 0.16	0.271 ± 0.08	0.427 ± 0.21	0.680 ± 0.16	**0.691 ± 0.13**	0.634 ± 0.13
msplice	0.949 ± 0.01	0.952 ± 0.01	0.951 ± 0.01	0.951 ± 0.01	0.952 ± 0.01	0.950 ± 0.01	0.962 ± 0.00	0.967 ± 0.01	0.964 ± 0.01	**0.989 ± 0.01**
letter	0.666 ± 0.16	0.665 ± 0.16	0.681 ± 0.14	0.683 ± 0.15	0.686 ± 0.14	0.687 ± 0.13	0.709 ± 0.01	**0.919 ± 0.02**	0.883 ± 0.01	0.885 ± 0.01
wavefm3	0.886 ± 0.01	0.884 ± 0.01	0.887 ± 0.01	0.886 ± 0.00	0.888 ± 0.01	0.889 ± 0.01	0.859 ± 0.00	0.887 ± 0.02	0.884 ± 0.01	**0.908 ± 0.02**
landsat	0.721 ± 0.11	0.717 ± 0.12	0.766 ± 0.08	0.745 ± 0.10	0.773 ± 0.06	0.694 ± 0.15	0.837 ± 0.06	0.874 ± 0.04	0.877 ± 0.05	**0.881 ± 0.03**

TABLE 6
Summary of t-Tests of Precision, Recall, F-Measure, and G-Mean with Significance Level at 0.05

	Methods	ADASYN	SMOTE	MWMOTE	SMTL	SMENN	Easy	BC	EOS	EUS	EHS
Precision	EOS	10-1-1	12-0-0	10-2-0	12-0-0	12-0-0	12-0-0	12-0-0	-	6-0-6	2-1-9
	EUS	9-1-2	10-0-2	10-0-2	10-0-2	10-0-2	11-0-1	11-1-0	6-0-6	-	2-0-10
	EHS	12-0-0	12-0-0	11-1-0	12-0-0	12-0-0	12-0-0	12-0-0	9-1-2	10-0-2	-
Recall	EOS	10-1-1	12-0-0	11-0-1	11-1-0	12-0-0	12-0-0	12-0-0	-	7-0-5	5-0-7
	EUS	10-0-2	11-1-0	11-0-1	11-0-1	10-1-1	10-0-2	12-0-0	5-0-7	-	3-0-9
	EHS	11-0-1	12-0-0	11-0-1	11-0-1	11-0-1	12-0-0	11-1-0	7-0-5	9-0-3	-
F-Measure	EOS	11-0-1	12-0-0	10-1-1	11-0-1	11-0-1	12-0-0	11-0-1	-	6-0-6	3-1-8
	EUS	10-0-2	11-1-0	10-0-2	11-0-1	10-1-1	12-0-0	11-1-0	6-0-6	-	2-0-10
	EHS	11-0-1	12-0-0	10-1-1	12-0-0	12-0-0	12-0-0	12-0-0	8-1-3	10-0-2	-
G-Mean	EOS	12-0-0	12-0-0	10-1-1	12-0-0	11-0-1	11-0-1	11-1-0	-	6-0-6	1-1-10
	EUS	10-1-1	10-1-1	9-0-3	10-0-2	10-0-2	10-0-2	12-0-0	6-0-6	-	4-0-8
	EHS	12-0-0	12-0-0	11-1-0	12-0-0	12-0-0	12-0-0	12-0-0	10-1-1	8-0-4	-

The amount of win-tie-lose is given between our proposed three methods and all the methods over 12 data sets using MLP and Adaboost.

F-Measure, and G-Mean. Thus EHS has more superiority than EOS and EUS.

5 Conclusion

In this paper, we present three new entropy-based learning approaches: EOS, EUS, and EHS, for multi-class imbalance learning problems. For a given imbalanced data set, the proposed methods use new entropy-based imbalance degrees to measure the class imbalance instead of using traditional imbalance-ratio. EOS is based on the information content of the largest majority class. EOS oversamples the other classes until their information contents achieve the largest one. Similarly, EUS undersamples the other classes to balance the data set based on the information content of the smallest minority class. EHS is based on the average information content of all the classes, and oversamples the minority classes as well as undersamples the majority classes according to EID. The new EID metrics consider the imbalance of class-wise information contents and offer us a new view of the imbalance in imbalanced learning. The effectiveness of our proposed three methods are demonstrated according to the superior learning

performance both on synthetic and real-world data sets. Furthermore, since EHS can better preserve data structure than EOS and EUS by generating less new minority samples as well as removing less majority samples to balance data sets, it has more superiority than EOS and EUS.

In the future, we would like to explore the theoretical properties of our proposed imbalance measure and extend it as well as our three imbalanced learning methods for other classification problems such as image classification and transfer learning.

Acknowledgments

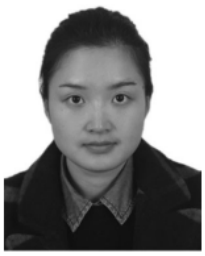
This work was supported in part by the US National Science Foundation under ECCS 1731672.

References

[1] H. He and E. A. Garcia, "Learning from imbalanced data," *IEEE Trans. Knowl. Data Eng.*, vol. 21, no. 9, pp. 1263–1284, Sep. 2009.
[2] Z. Wan, H. He, and B. Tang, "A generative model for sparse hyperparameter determination," *IEEE Trans. Big Data*, vol. 4, no. 1, pp. 2–10, Jan.-Mar. 2018.
[3] C.-T. Lin, T.-Y. Hsieh, Y.-T. Liu, Y.-Y. Lin, C.-N. Fang, Y.-K. Wang, G. Yen, N. R. Pal, and C.-H. Chuang, "Minority oversampling in kernel adaptive subspaces for class imbalanced datasets," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 5, pp. 950–962, May 2018.
[4] M. Ohsaki, P. Wang, K. Matsuda, S. Katagiri, H. Watanabe, and A. Ralescu, "Confusion-matrix-based kernel logistic regression for imbalanced data classification," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 9, pp. 1806–1819, Sep. 2017.
[5] Z. Wan, H. Li, H. He, and D. Prokhorov, "Model-free real-time EV charging scheduling based on deep reinforcement learning," *IEEE Trans. Smart Grid*, to be published, doi: 10.1109/TSG.2018.2879572.
[6] 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, pp. 321–357, 2002.
[7] T. Zhu, Y. Lin, and Y. Liu, "Synthetic minority oversampling technique for multiclass imbalance problems," *Pattern Recognit.*, vol. 72, pp. 327–340, 2017.
[8] K. E. Bennin, J. Keung, P. Phannachitta, A. Monden, and S. Mensah, "MAHAKIL: Diversity based oversampling approach to alleviate the class imbalance issue in software defect prediction," *IEEE Trans. Softw. Eng.*, vol. 44, no. 6, pp. 534–550, Jun. 2018.
[9] Z. Wan and H. He, "AnswerNet: Learning to answer questions," *IEEE Trans. Big Data*, to be published, doi: 10.1109/TBDATA.2018.2884486.
[10] C. Bunkhumpornpat, K. Sinapiromsaran, and C. Lursinsap, "Safe-level-SMOTE: Safe-level-synthetic minority over-sampling technique for handling the class imbalanced problem," in *Proc. Pacific-Asia Conf. Advances Knowl. Discovery Data Mining*, 2009, pp. 475–482.
[11] 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.*, 2008, pp. 1322–1328.
[12] S. Chen, H. He, and E. A. Garcia, "RAMOBoost: Ranked minority oversampling in boosting," *IEEE Trans. Neural Netw.*, vol. 21, no. 10, pp. 1624–1642, Oct. 2010.
[13] S. Barua, M. M. Islam, X. Yao, and K. Murase, "MWMOTE–Majority weighted minority oversampling technique for imbalanced data set learning," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 2, pp. 405–425, Feb. 2014.
[14] 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.*, 2005, pp. 878–887.
[15] X. Yang, Q. Kuang, W. Zhang, and G. Zhang, "AMDO: An over-sampling technique for multi-class imbalanced problems," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 9, pp. 1672–1685, Sep. 2018.
[16] L. Li, H. He, J. Li, and W. Li, "EDOS: Entropy difference-based oversampling approach for imbalanced learning," in *Proc. Int. Joint Conf. Neural Netw.*, 2018, pp. 1–8.
[17] X.-Y. Liu, J. Wu, and Z.-H. Zhou, "Exploratory undersampling for class-imbalance learning," *IEEE Trans. Syst. Man Cybern. Part B (Cybern.)*, vol. 39, no. 2, pp. 539–550, Apr. 2009.
[18] Q. Kang, X. Chen, S. Li, and M. Zhou, "A noise-filtered under-sampling scheme for imbalanced classification," *IEEE Trans. Cybern.*, vol. 47, no. 12, pp. 4263–4274, Dec. 2017.
[19] P. Lim, C. K. Goh, and K. C. Tan, "Evolutionary cluster-based synthetic oversampling ensemble (ECO-ensemble) for imbalance learning," *IEEE Trans. Cybern.*, vol. 47, no. 9, pp. 2850–2861, Sep. 2017.
[20] S. Wang, L. L. Minku, and X. Yao, "Resampling-based ensemble methods for online class imbalance learning," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 5, pp. 1356–1368, May 2015.
[21] S. Wang and X. Yao, "Diversity analysis on imbalanced data sets by using ensemble models," in *Proc. IEEE Symp. Comput. Intell. Data Mining*, 2009, pp. 324–331.
[22] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: Improving classification performance when training data is skewed," in *Proc. 19th Int. Conf. Pattern Recognit.*, 2008, pp. 1–4.
[23] A. Lazarevic, N. Chawla, L. Hall, and K. Bowyer, "SMOTE-Boost: Improving the prediction of minority class in boosting," in *Proc. 7th Eur. Conf Principles Practice Knowl. Discovery Databases*, 2002, pp. 107–119.
[24] B. Tang and H. He, "GIR-based ensemble sampling approaches for imbalanced learning," *Pattern Recognit.*, vol. 71, pp. 306–319, 2017.
[25] G. E. Batista, R. C. Prati, and M. C. Monard, "A study of the behavior of several methods for balancing machine learning training data," *ACM SIGKDD Explorations Newslett.*, vol. 6, no. 1, pp. 20–29, 2004.
[26] C. E. Shannon, "A mathematical theory of communication," *ACM SIGMOBILE Mobile Comput. Commun. Rev.*, vol. 5, no. 1, pp. 3–55, 2001.
[27] S. Li, L. Li, J. Yan, and H. He, "SDE: A novel clustering framework based on sparsity-density entropy," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 8, pp. 1575–1587, Aug. 2018.
[28] R. M. Gray, *Entropy and Information Theory*. Berlin, Germany: Springer, 2011.
[29] M. A. Nielsen and I. L. Chuang, *Quantum Computation and Quantum Information*. Cambridge, U.K.: Cambridge Univ. Press, 2010.
[30] L. Feng, H. Wang, B. Jin, H. Li, M. Xue, and L. Wang, "Learning a distance metric by balancing KL-divergence for imbalanced datasets," *IEEE Trans. Syst. Man Cybern.: Syst.*, to be published, doi: 10.1109/TSMC.2018.2790914.
[31] Z. Wan, C. Jiang, M. Fahad, Z. Ni, Y. Guo, and H. He, "Robot-assisted pedestrian regulation based on deep reinforcement learning," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/TCYB.2018.2878977.
[32] Q. Wang, J. Wan, F. Nie, B. Liu, C. Yan, and X. Li, "Hierarchical feature selection for random projection," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 30, no. 5, pp. 1581–1586, May 2019.
[33] Q. Wang, M. Chen, F. Nie, and X. Li, "Detecting coherent groups in crowd scenes by multiview clustering," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2018.
[34] D. J. MacKay, *Information Theory, Inference and Learning Algorithms*. Cambridge, U.K.: Cambridge Univ. Press, 2003.
[35] G. Hripcsak and A. S. Rothschild, "Agreement, the F-measure, and reliability in information retrieval," *J. Amer. Med. Informat. Assoc. Jamia*, vol. 12, no. 3, pp. 296–298, 2005.
[36] H. Guo, H. Liu, C. Wu, W. Zhi, Y. Xiao, and W. She, "Logistic discrimination based on G-mean and F-measure for imbalanced problem," *J. Intell. Fuzzy Syst.*, vol. 31, no. 3, pp. 1155–1166, 2016.
[37] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, and S. Garcia, "KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework," *J. Multiple-Valued Logic Soft Comput.*, vol. 17, pp. 255–287, 2011.
[38] A. Fernandez, S. Garcia, M. J. del Jesus, and F. Herrera, "A study of the behaviour of linguistic fuzzy rule based classification systems in the framework of imbalanced data-sets," *Fuzzy Sets Syst.*, vol. 159, no. 18, pp. 2378–2398, 2008.
[39] G. Lemaitre, 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.

Lusi Li received the BS and MS degrees in computer science from the Zhongnan University of Economics and Law, China, in 2014 and 2017, respectively. Now, she is currently working toward the PhD degree in electrical engineering at the University of Rhode Island, Kingston, Rhode Island. Her research interests include machine learning, data mining, and transfer learning. She is a student member of the IEEE.

Haibo He (SM'11-F'18) received the BS and MS degrees in electrical engineering from the Huazhong University of Science and Technology, China, in 1999 and 2002, respectively, and the PhD degree in electrical engineering from Ohio University, in 2006. From 2006 to 2009, he was an assistant professor with the Department of Electrical and Computer Engineering, Stevens Institute of Technology. Currently, he is the Robert Haas endowed chair professor with the Department of Electrical, Computer, and Biomedical Engineering, University of Rhode Island. His research interests include computational intelligence, machine learning and data mining, and various applications. He severed as the general chair of the IEEE Symposium Series on Computational Intelligence (SSCI 2014). He was a recipient of the IEEE International Conference on Communications Best Paper Award (2014), IEEE Computational Intelligence Society (CIS) Outstanding Early Career Award (2014), and National Science Foundation (NSF) CAREER Award (2011). Currently, he is the editor-in-chief of the *IEEE Transactions on Neural Networks and Learning Systems*. He is a fellow of the IEEE.



Jie Li received the BS and PhD degrees in automation from Chongqing University, China, in 2007 and 2011, respectively. She was an assistant professor with the Chongqing Institute of Green and Intelligent Technology, Chinese Academy of Sciences, China, in 2012. She is currently an assistant professor with the Chongqing University of Science and Technology, China. Her research interests include machine learning, deep neural network, generative adversarial network, and signal processing. She is a member of the IEEE.