Decision Tree Algorithm with Class Overlapping-Balancing Entropy for Class Imbalanced Problem

Artit Sagoolmuang and Krung Sinapiromsaran

***Abstract*—The problem of handling a class imbalanced problem by modifying decision tree algorithm has received widespread attention in recent years. A new splitting measure, called class overlapping-balancing entropy (OBE), is introduced in this paper that essentially pay attentions to all classes equally. Each step, the proportion of each class is balanced via the assigned weighted values. They not only depend on equalizing each class, but they also take into account the overlapping region between classes. The proportion of weighted values corresponding to each class is used as the component of Shannon's entropy for splitting the current dataset. From the experimental results, OBE significantly outperforms the conventional splitting measures like Gini index, gain ratio and DCSM, which are used in the well-known decision tree algorithms. It also exhibits superior performance compared to AE and ME that are designed for handling the class imbalanced problem specifically.**

***Index Terms*—Classification problem, class imbalanced learning, class overlapping-balancing entropy, decision tree algorithm.**

I. Introduction

A decision tree is recognized as one of the top 10 classification models [1]. The success of using the decision tree can be explained by three characteristics. First, a decision tree algorithm consumes small computational time for constructing the model, especially during the predicting step. Second, a decision tree has the easy interpretation for humans that it has been used for ranking variable importance. Third, a decision tree is robust with respect to anomalies and missing values. However, like most well-known classifiers, a decision tree algorithm must face the hassle of classifying a dataset with extremely unequal class distribution [2]. This problem plays an important role in knowledge discovery and data mining for the past several years, which is known as a class imbalanced problem. It widely appears in several real-world situations such as fraud detection [3], [4], disease diagnosis [5], [6], network intrusion detection [7], industrial systems monitoring [8] and sentiment analysis [9]. To minimize the accuracy of classification, the decision tree algorithm often build a tree that predicts most unknown instances to be the class containing a large number of instances, called the majority class. Hence, instances from the class containing a tiny number of instances, called the minority class, tend to be incorrectly classified. In the real-world problem, the smaller class is frequently more important and receives much attention to correctly classify. For example, in fraud detection, there is a small number of fraudulent transactions, but they are significant and must be discovered. In the same way as disease diagnosis, the prediction of disease patients is more critical than normal people.

Many methods have been presented to deal with the class imbalanced problem using various techniques [10], [11]. The idea of developing the algorithm to build the decision tree classifier that is suitable for classifying an imbalanced dataset is one of the methods that have received wide attention. Normally, the improvement of decision tree algorithm usually focuses on modifying the splitting measures to separate dataset in each node. Traditional splitting measures, especially Gini index [12] and Shannon's entropy [13], have been improved using many concepts in recent years. Asymmetric entropy (AE) [14], off-centered entropy (OCE) [15], [16] and AECID [17] apply the concept of non-symmetry instead of the symmetric one. They shift the maximum value of entropy from the middle of extreme proportions as the symmetric entropy, to be biased toward the minority class. In addition, the skew-insensitive splitting measures are suggested for dealing with the class imbalanced problem, such as DKM [18], [19] and HDDT [20], [21]. They can condone a considerable difference between the number of instances in the minority class and the majority class. Lastly and most importantly, the concept of modifying the components of the Gini index and the Shannon's entropy calculation to be inclined towards minority class are introduced in CART+Resampling [22] and minority entropy (ME) [23], respectively. They discard majority instances that do not affect the split decision of minority instances. CART+Resampling applies the sampling method directly, while ME ignores majority instances outside the minority range which has the similar effect as the sampling method.

This paper suggests the modification of Shannon's entropy components like ME for continuous attribute. The splitting measure designed to handle the class imbalanced problem is proposed, called class overlapping-balancing entropy (OBE). It assigns a larger weight to an instance that lies outside the overlapping region between two classes than the weights of other instances. Moreover, the sum of weights among all classes are set equal to one to make them balance. Then, the proportion of weights corresponding to each class is

Manuscript received February 10, 2020; revised March 5, 2020.

The authors are with the Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, Bangkok 10330, Thailand (e-mail: a.sagoolmuang@gmail.com, krung.s@chula.ac.th).

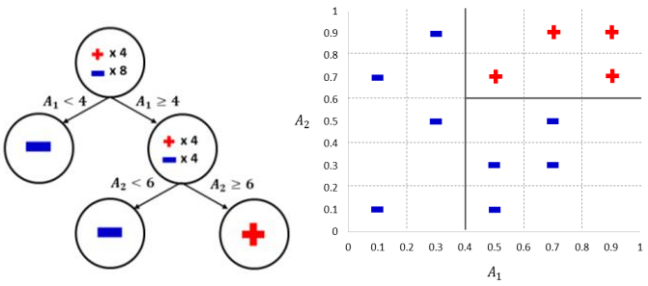
employed as the component of computing Shannon's entropy.

The remaining of this paper is outlined as follows. In Section II, a brief review of the decision tree classifier is shown. Next, Section III demonstrates the detail of the proposed splitting measure, OBE, along with its properties. Then, the discussion of experimental results is presented in Section IV. Finally, Section V concludes this research.

II. Decision Tree Classifier

Background knowledge regarding the proposed method is demonstrated in this section. It begins with formulating the classification problem, then explaining the construction of the decision tree classifier.

Initially, the problem formulation relating to this paper is defined. Given $\mathbf{D} = \{(\vec{x}_i, y_i) \mid i = 1,2,\ldots,m\} \subseteq \mathbb{R}^n \times C$ be a labeled dataset of binary classification problem, where C is a set of binary classes $C = \{+1, -1\}$. Then, $\mathbf{D}$ can be separated into 2 partitions, i.e., $\mathbf{D} = \mathbf{D}_+ \cup \mathbf{D}_-$ where $\mathbf{D}_+ = \{(\vec{x}_i, y_i) \in \mathbf{D} \mid y_i = +1 \text{ for } i = 1,2,\ldots,m\}$ and $\mathbf{D}_- = \{(\vec{x}_i, y_i) \in \mathbf{D} \mid y_i = -1 \text{ for } i = 1,2,\ldots,m\}$ having size m_+ and m_- respectively, such that $m_+ + m_- = m$.



(a) (b)

Fig. 1. An example of using the decision tree to classify a dataset. Decision tree classification model (a). Data partitioning with the decision tree algorithm (b).

A decision tree is a tree-based classification model consisting of multiple connected nodes. Each non-leaf node, including the root node and the internal node, presents a splitting condition. For each leaf node, it indicates a specific predicted class of instances. Graphically, for example, the decision tree presented in Fig. 1(a) consists of the root node at the top, one internal node, and three leaf nodes.

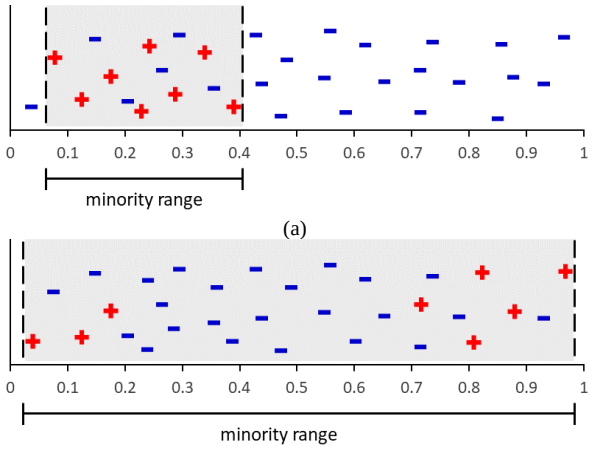
For the procedure of inducing a decision tree, at each non-leaf node, the set of instances is divided into two partitions using the selected splitting condition represented by the particular value of a specific continuous attribute. Then, the process is recursively continued until all instances in a child node have the same class labels or meeting the stopping criteria.

To select the splitting value, a greedy approach is applied. It considers all values along a variable between all instances indicated by dash lines in Fig. 1(b). The best one providing the optimal splitting measure is chosen indicated by solid lines in Fig. 1(b). Various splitting measures have been proposed based on the impurity of each partition such as the Gini index [12] and Shannon's entropy [13] which are used in the well-known decision tree algorithm like CART [24] and ID3 [25], respectively. The formulation of Shannon's Entropy, which is considered in this study, is defined by (1). Its minimum value is equal to zero appearing when all instances in the partition are in the same class. For the maximum value, it is equal to one appearing when there is a similar number of instances from each class in the partition. In addition, the most famous decision tree algorithm like C4.5 [26] applies the normalization of Information Gain, Shannon's entropy reduction after splitting, called Gain Ratio to be the splitting measure. Another interesting splitting measure is the distinct class based splitting measure (DCSM) [27]. It improves the performance of building a decision tree by considering the number of distinct classes. The partition with the smaller number of distinct classes means the purer partition.

$$Entropy(\mathbf{D}) = -\frac{|\mathbf{D}_+|}{|\mathbf{D}|}\log_2\frac{|\mathbf{D}_+|}{|\mathbf{D}|} - \frac{|\mathbf{D}_-|}{|\mathbf{D}|}\log_2\frac{|\mathbf{D}_-|}{|\mathbf{D}|} \quad (1)$$

III. Class Overlapping-Balancing Entropy

In this section, an enhanced splitting measure designed for handling the class imbalanced problem is introduced, called class overlapping-balancing entropy (OBE). It balances the dataset together with the concept of overlapping region between two classes.



(a)

(b)

Fig. 2. Two scenarios for partially and fully covered instances via the range of minority class.

A. Motivation

The motivation of this paper comes from the success of using the decision tree classifier based on minority entropy (ME) [20] to handle the class imbalanced problem. It fixed the problem of Shannon's entropy that biased towards the majority class by keeping the majority instances within the minority range. For an attribute of each internal node, ME considers only a subset of instances within the range between the smallest and the largest values of minority instances, displaying in Fig. 2 which changes the proportion of calculating the entropy. It does not change the number of minority instances (represented by positive class) but the number of majority instances (represented by negative class) decreases, see Fig. 2(a). This inevitably will split the dataset

in the region of the minority distribution. However, having exceedingly focus on the minority class causes the built decision tree to be biased. Additionally, ME does not guarantee that the effect of majority class will be reduced. It depends on the boundary of minority range. If the minority range covers all majority instances which is shown in Fig. 2(b), ME gives the same value as Shannon's entropy.

The concept of ME determining which majority instances will be dropped for the entropy calculation, is extended in this paper. OBE assigns weight to each majority instance instead of 0 or 1 assignment as ME. The weight of each instance will be assigned the value in the range (0,1] depending on the instance's position and its class. For ME, only the instances locating in the overlapping region are considered, while the rest are abandoned. On the contrary, this paper has a different notion: an instance locating in the region of a single class should be more important than an instance appearing in the region of multiple classes. That is because it represents the region of its class clearly, not an area that is shared with other classes. Then, a set of weights with respect to the instances from each class is normalized to have a total equal to one for balancing between two classes. It is used to be the calculation components of Shannon's entropy for selecting the splitting condition at each internal node. From the above reasons, it can be concluded that the weight assignment in this work is based on the following two assumptions:

1) The instances locating in the overlapping regions must be assigned smaller weight than other instances outside the overlapping regions from the same class.

2) The summation of weights corresponding to the instances belonging in one class must be equal to one.

Consequently, it is ensured that the effect of majority class will be the same as the minority class. Also, OBE does not abandon the majority instances like ME causing the built decision tree to be biased toward the minority class.

B. Class Overlapping-Balancing Entropy (OBE)

Originally, the components of Shannon's entropy comprise the proportion of instances in each class. It is defined by the number of instances in each class divided by the total number of instances in a dataset, as shown in (1). Shannon's entropy treats all instances equally counting each instance as one. ME preserves the entropy formula but it uses the subset of instances within the minority range to be computed. In this paper, an adjusted proportion in each class is defined to be consistent with the set of weights corresponding to the dataset $\mathbf{D}$, i.e.

$$\mathbf{W}^{\omega}(\mathbf{D}) = \{w_i \mid w_i = \omega(\vec{x}_i, y_i) \text{ where } (\vec{x}_i, y_i) \in \mathbf{D} \text{ for } i = 1, 2, \ldots, m\},$$

where ω is the weighting function. The number of instances in each class divided by the total number of instances which is used to compute Shannon's entropy is changed to the summation of the weights corresponding to all instances in each class divided by the total weights. The original formula of Shannon's entropy applying the proposed components is defined in (2) as follows:

$$Entropy(\mathbf{W}^{\omega}(\mathbf{D})) = -\frac{\sum_{w_i \in \mathbf{W}_+^{\omega}(\mathbf{D})} w_i}{\sum_{w_i \in \mathbf{W}^{\omega}(\mathbf{D})} w_i} \log_2 \frac{\sum_{w_i \in \mathbf{W}_+^{\omega}(\mathbf{D})} w_i}{\sum_{w_i \in \mathbf{W}^{\omega}(\mathbf{D})} w_i} - \frac{\sum_{w_i \in \mathbf{W}_-^{\omega}(\mathbf{D})} w_i}{\sum_{w_i \in \mathbf{W}^{\omega}(\mathbf{D})} w_i} \log_2 \frac{\sum_{w_i \in \mathbf{W}_-^{\omega}(\mathbf{D})} w_i}{\sum_{w_i \in \mathbf{W}^{\omega}(\mathbf{D})} w_i} \quad (2)$$

where, $\mathbf{W}_+^{\omega}(\mathbf{D}) = \{w_i \in \mathbf{W}^{\omega}(\mathbf{D}) \mid y_i = +1 \text{ for } i = 1, 2, \ldots, m\}$ and $\mathbf{W}_-^{\omega}(\mathbf{D}) = \{w_i \in \mathbf{W}^{\omega}(\mathbf{D}) \mid y_i = -1 \text{ for } i = 1, 2, \ldots, m\}$.

OBE retains the main structure of the original formula of Shannon's entropy. Progressively, using the weighted values as a component is more general and versatile. For dealing with the class imbalanced problem, the weighting function ω has been proposed based on two assumptions.

Firstly, the overlapping weighting function with respect to attribute j (denoted by α_j) is introduced in accordance with the first assumption, which is defined by (3). The weight assigning to an instance $(\vec{x}_i, y_i) \in \mathbf{D}$ is inversely proportional to the summation of the number of instances in the classes having a range covering its position, which has been scaled down by the logarithmic function. The property of determining the weighted values to the instances in $\mathbf{D}$ using the overlapping weighting function is presented in Theorem 1, which corresponds to the first assumption.

$$\alpha_j(\vec{x}_i, y_i) = (\log_2(|\mathbf{D}_{sign(y_i)}|) + \mathbf{1}(\pi_j(\vec{x}_i) \in range_j(\mathbf{D}_{sign(-y_i)})) \cdot \log_2(|\mathbf{D}_{sign(-y_i)}|))^{-1} \quad (3)$$

where,

- $\mathbf{1}(\sigma)$ is the indicator function. It obtains the value 1 if the condition σ is true. Otherwise, it is set to 0.
- $\pi_j(\vec{x}_i)$ is the projection of $\vec{x}_i$ onto attribute j.
- $range_j(\mathbf{D}_k) = \left[\min_{\vec{x}_i \in \mathbf{D}_k} \pi_j(\vec{x}_i), \max_{\vec{x}_i \in \mathbf{D}_k} \pi_j(\vec{x}_i)\right]$.

Theorem 1. For a binary class dataset $\mathbf{D} = \{(\vec{x}_i, y_i) \mid i = 1, 2, ..., m\}$, define $Overlap_j(\mathbf{D}) = range_j(\mathbf{D}_+) \cap range_j(\mathbf{D}_-)$ as the overlapping region of two class corresponding to attribute j. If two instances $(\vec{x}_a, y_a) \in \mathbf{D}$ and $(\vec{x}_b, y_b) \in \mathbf{D}$ come from the same class, i.e. $y_a = y_b$, which $\pi_j(\vec{x}_a) \in Overlap_j(\mathbf{D})$ and $\pi_j(\vec{x}_b) \notin Overlap_j(\mathbf{D})$ respectively, then $\alpha_j(\vec{x}_a, y_a) < \alpha_j(\vec{x}_b, y_b)$.

Proof. Since $\pi_j(\vec{x}_a) \in Overlap_j(\mathbf{D})$, so $\pi_j(\vec{x}_a) \in range_j(\mathbf{D}_{sign(-y_a)})$. While $\pi_j(\vec{x}_b) \notin Overlap_j(\mathbf{D})$, so $\pi_j(\vec{x}_b) \notin range_j(\mathbf{D}_{sign(-y_b)})$. Hence,

$$\begin{aligned} \log_2(|\mathbf{D}_{sign(y_a)}|) &= \log_2(|\mathbf{D}_{sign(y_b)}|) \\ \log_2(|\mathbf{D}_{sign(y_a)}|) + \log_2(|\mathbf{D}_{sign(-y_a)}|) &> \log_2(|\mathbf{D}_{sign(y_b)}|) \\ \left(\log_2(|\mathbf{D}_{sign(y_a)}|) + \log_2(|\mathbf{D}_{sign(-y_a)}|)\right)^{-1} &< \left(\log_2(|\mathbf{D}_{sign(y_b)}|)\right)^{-1} \\ \alpha_j(\vec{x}_a, y_a) &< \alpha_j(\vec{x}_b, y_b) \end{aligned}$$

Secondly, the balancing weighting function (denoted by

β) is introduced in accordance with the second assumption, which is defined by (4). A weight value w_i of an arbitrary set of weights corresponding to dataset $\mathbf{D}$, $W = \{w_i \mid (\vec{x}_i, y_i) \in \mathbf{D} \text{ and } i = 1,2,...,m\}$, is normalized by its class. The total weights with respect to each class is balanced (equal to 1) as shown in Theorem 2, corresponding to the second assumption.

$$\beta(w_i) = \frac{w_i}{\sum_{w_l \in W_{sign(y_i)}} w_l} \tag{4}$$

where, $W_k = \{w_i \in W \mid y_i = k \text{ for } i = 1,2,\ldots,m\}$.

Theorem 2. For an arbitrary set of weights corresponding to dataset $\mathbf{D}$ with $\mathbf{W}^{\omega}(\mathbf{D})$, the summation of the weights values from instances of the same class assigning by the balancing weighting function is equal to one.

Proof. For $k \in \{+,-\}$, the summation of the weights values assigning by the balancing weighting function with respect to the instances of k is equal to

$$\sum_{w_i \in \mathbf{W}_k^{\omega}(\mathbf{D})} \beta(w_i) = \sum_{w_i \in \mathbf{W}_k^{\omega}(\mathbf{D})} \frac{w_i}{\sum_{w_l \in \mathbf{W}_k^{\omega}(\mathbf{D})} w_l} = \frac{1}{\sum_{w_l \in \mathbf{W}_k^{\omega}(\mathbf{D})} w_l} \sum_{w_i \in \mathbf{W}_k^{\omega}(\mathbf{D})} w_i = 1$$

For attribute j, the class overlapping-balancing entropy (OBE) of a dataset $\mathbf{D}$ is defined in (5) according to the composite function between the overlapping weighting function and the balancing weighting function as follows:

$$OBE_j(\mathbf{D}) = Entropy(\mathbf{W}^{\beta o \alpha_j}(\mathbf{D})) \tag{5}$$

C. Workflow and Example

The workflow of assigning weight values to all instances in the dataset on a specific attribute is illustrated by Fig. 3. It starts with partitioning the dataset $\mathbf{D}$ of size m into two subsets that are the set of minority instances $\mathbf{D}_+$ of size m_+ and the set of majority instances $\mathbf{D}_-$ of size m_-. Each subset is also divided into two parts by considering the position of each instance with respect to the overlapping region. So, there are four groups of instances which are partitioned from $\mathbf{D}$, i.e. 1) the set of minority instances outside the overlapping region $\mathbf{D}_+^{\text{out}}$ of size m_+^{out}, 2) the set of minority instances inside the overlapping region $\mathbf{D}_+^{\text{in}}$ of size m_+^{in}, 3) the set of majority instances outside the overlapping region $\mathbf{D}_-^{\text{out}}$ of size m_-^{out}, and 4) the set of majority instances inside the overlapping region $\mathbf{D}_-^{\text{in}}$ of size m_-^{in}. Then, all instances in each group are assigned the overlapping weight equally, see (3). Finally, the balancing weight is used to make the total weights in each class equals to one from (4).

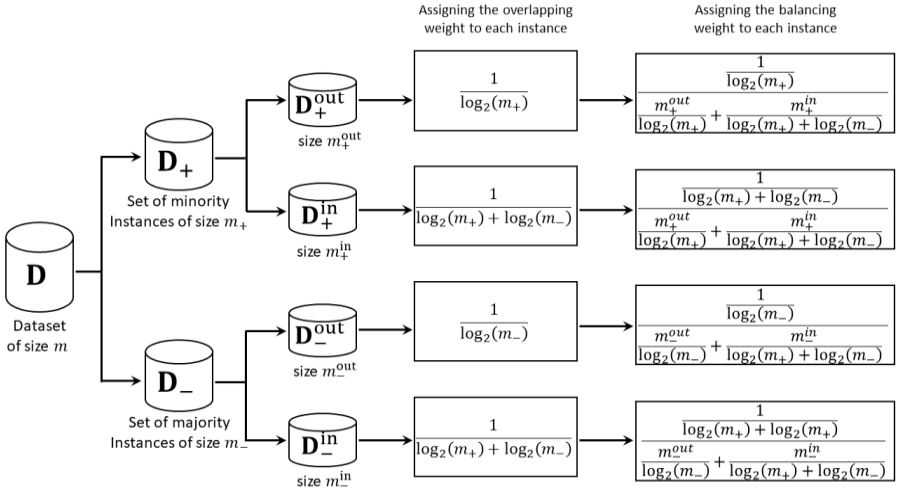


Fig. 3. The workflow of proposed weighted assignment.

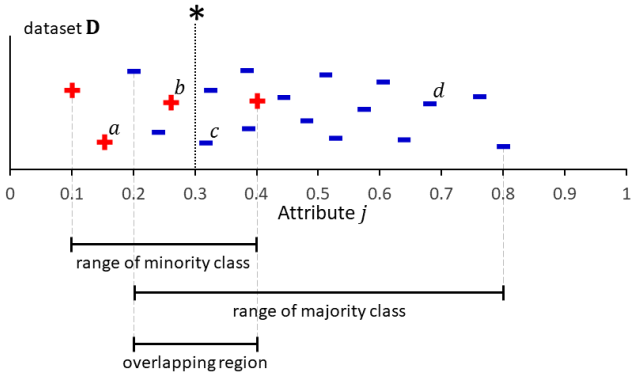


Fig. 4. An example of a binary class imbalanced dataset which is partitioned by the splitting value at 0.3 of attribute j.

For the time complexity analysis of assigning weight values to all instances in the dataset on a specific attribute, there are six main parts to consider. First, separating the dataset $\mathbf{D}$ into two subsets takes $O(m)$ time complexity. Second, computing the overlapping region uses $O(m)$ running time. Third, spending $O(m)$ for partitioning each subset. Fourth and fifth, calculating the overlapping weight for each group takes $O(1)$ same as calculating the balancing weight. Sixth, assigning the weight to each instance based on its position uses $O(m)$. Hence, the overall time complexity is $O(m)+O(m)+2\times O(m)+4\times O(1)+4\times O(1)+O(m)=O(m)$.

For example, in Fig. 4, the dataset **D** consists of 5 minority instances and 15 majority instances. The range of each class with respect to attribute j including the overlapping region are displayed below the figure. The calculation of the weighted values of instances a, b, c and d locating in different positions are demonstrated. It begins with calculating the overlapping weights.

- For instance a, it is labeled as the minority class locating outside the overlapping region. Thus,

$$\alpha_j(a,+1)=\frac{1}{\log_2(4)}=\frac{1}{2}.$$

- For instance b, it is labeled as the minority class locating inside the overlapping region. Thus,

$$\alpha_j(b,+1)=\frac{1}{\log_2(4)+\log_2(16)}=\frac{1}{2+4}=\frac{1}{6}.$$

- For instance c, it is labeled as the majority class locating inside the overlapping region. Thus,

$$\alpha_j(c,-1)=\frac{1}{\log_2(4)+\log_2(16)}=\frac{1}{2+4}=\frac{1}{6}.$$

- For instance d, it is labeled as the majority class locating outside the overlapping region. Thus,

$$\alpha_j(d,-1)=\frac{1}{\log_2(16)}=\frac{1}{4}.$$

Then, they are normalized to balance the class weights. Initially, the total overlapping weights corresponding to each class is computed. It is equal to $2\cdot\frac{1}{2}+2\cdot\frac{1}{6}=\frac{4}{3}$ for the minority class and equal to $6\cdot\frac{1}{6}+10\cdot\frac{1}{4}=\frac{7}{2}$ for the majority class. Consequently, by applying the balancing weight, the weighted values of instance a, b, c and d are $\frac{1}{2}\times\frac{3}{4}=0.375$, $\frac{1}{6}\times\frac{3}{4}=0.125$, $\frac{1}{6}\times\frac{2}{7}=0.048$, and $\frac{1}{4}\times\frac{2}{7}=0.071$, respectively.

Moreover, the calculation of class overlapping-balancing entropy of the subset of instances having the value lower than 0.3 of attribute j (left partition) is demonstrated. There are two minority instances outside the overlapping region and one minority instance inside the overlapping region, so the total weight corresponding to the minority class is equal to $2\times0.375+0.125=0.875$. While there are two majority instances inside the overlapping region, so the total weight corresponding to the majority class is equal to $2\times0.048=0.096$. Therefore, the total weight is equal to $0.875+0.096=0.971$. Hence,

$$OBE_j(\mathbf{D})=-\frac{0.875}{0.971}\log_2\frac{0.875}{0.971}-\frac{0.096}{0.971}\log_2\frac{0.096}{0.971}=0.465.$$

IV. Experiments and Results

There are two collections of experiments to evaluate the effectiveness of the proposed class overlapping-balancing entropy (OBE). The first collection uses synthetic datasets via F-measure and G-measure comparing with Shannon's entropy. The second collection composes of twelve real-world datasets from UCI repository via precision, recall, F-measure and G-measure comparing with Gini index, gain ratio, DCSM, AE, and ME. Moreover, the Wilcoxon signed-rank test is performed.

A. Evaluation Metrics and Statistical Test

In the experiments, various evaluation metrics [28] are used to measure the performance of each method consisting of Precision (6), Recall (7), F-measure (8) and G-measure (9). They are defined from the confusion matrix as shown in Table I.

TABLE I: Confusion Matrix

	Predicted positive	Predicted negative
Actual positive	True positive (*TP*)	False negative (*FN*)
Actual negative	False positive (*FP*)	True negative (*TN*)

$$\text{Precision}=\frac{TP}{TP+FP} \quad (6)$$

$$\text{Recall}=\frac{TP}{TP+FN} \quad (7)$$

$$F-\text{measure}=2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision+Recall}} \quad (8)$$

$$G-\text{measure}=\sqrt{\text{Precision}\times\text{Recall}} \quad (9)$$

where

- *TP* is the number of minority instances that are correctly classified.
- *FP* is the number of majority instances that are incorrectly classified.
- *TN* is the number of majority instances that are correctly classified.
- *FN* is the number of minority instances that are incorrectly classified.

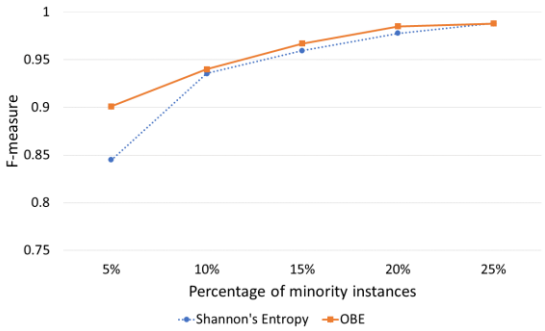
Precision presents the percentage of predicted minority instances that are correctly classified while recall presents the percentage of actual minority instances that are correctly classified. For F-measure and G-measure, they indicate the harmonic mean and geometric mean of the two metrics above, respectively.

Statistically, Wilcoxon signed-rank test with 0.1, 0.05 and 0.01 significance level (α) [29] is evaluated to show significant difference between other splitting measures and OBE. The null hypothesis (H_0) states that there is no difference between the performance of OBE and another method, while the alternative hypothesis (H_1) indicates that there is a difference between them.

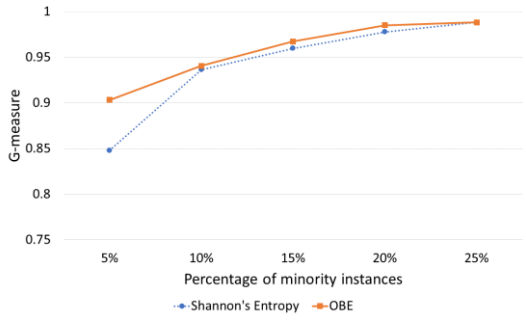
B. Experiments on Synthetic Datasets

An improvement of classifying instances for the imbalanced datasets is shown between Shannon's entropy and OBE on the synthetic datasets. Each synthetic dataset used in this section is the set of 1000 instances having 10 attributes. For each attribute, the uniform sampling was

performed within specified ranges of minority class and majority class that overlap. There are five groups of experiments having different percentages of minority instances from 5% to 25%, then repeating for 50 times for each experiment.



(a)



(b)

Fig. 5. The experimental results on synthetic datasets varying percentage of minority instances comparing with Shannon's entropy via F-measure (a) and G-measure (b).

The average results of F-measure and G-measure are shown in Fig. 5. They present the similar results that the values of both OBE and Shannon's entropy increase when the percentage of minority instances increase. Evidently, OBE significantly outperforms Shannon's entropy when the number of minority instances is tiny whereas their values will approach to 1 when a dataset is more balanced. This confirms that Shannon's entropy may not be suitable to deal with the class imbalanced problem.

C. Experiments on Real-World Datasets

The performance of classifying the real-world datasets of OBE is evaluated in this section comparing with five splitting measures. The first two traditional splitting measures are Gini index [21] used in CART algorithm [22], gain ratio used in C4.5 algorithm [11], and DCSM [23]. Importantly, two splitting measures which are proposed to handle with the class imbalanced problem like AE [12] and ME [20] are considered in the experiments. They have been shown in [20] as the best two measures for dealing with that problem.

1) Real-world datasets

The experiments were performed on twelve real-world datasets from the UCI repository [30], which are summarized in Table II. It is sorted in descending order by the percentage of instances in the minority class (%Min.) equivalent with the ascending order sort of the percentage of instances in the majority class (%Maj.). The first two columns indicate the number and the name of each dataset. For the number of instances (#Inst.) and the number of attributes (#Att.), they are shown in the third column and the fourth column, respectively. Particularly, the classes determining to be the minority class and the majority class are presented in the fifth column. In order to evaluate the performance of each method, the five-fold cross-validation is employed repeating 20 times. That is, there are up to one hundred experiments performed on each dataset.

TABLE II: REAL-WORLD DATASETS FROM UCI REPOSITORY

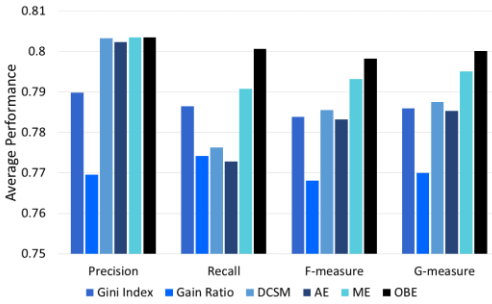
No	Datasets	#Inst	#Att	Min/Maj Class	%Min	%Maj
1	Pima	768	8	1/0	34.90	65.10
2	Wine	178	13	3/the rest	26.97	73.03
3	Haberman	306	3	2/1	26.47	73.53
4	Vehicle	846	18	bus/the rest	25.77	74.23
5	Shuttle	58000	9	the rest/1	21.40	78.60
6	Thyroid	215	5	2/the rest	16.28	83.72
7	Image	2310	19	5/the rest	14.29	85.71
8	Ecoli	336	7	imU/the rest	10.42	89.58
9	OpticDigits	5620	64	4/the rest	10.11	89.89
10	PenDigits	10992	16	5/the rest	9.60	90.40
11	Libras	360	90	15/the rest	6.67	93.33
12	PageBlocks	5473	10	2/the rest	6.01	93.99

2) Results and discussions

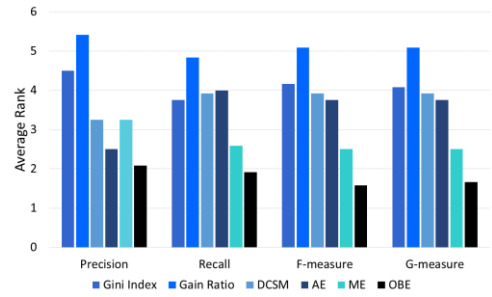
The experimental results are demonstrated in Fig. 6. The comparison of the average performance corresponding to each evaluation metric is shown in Fig. 6(a) where the higher value indicates the better performance. While Fig. 6(b) represents the comparison results by the average rank of performance corresponding to each evaluation metric where the lower value indicates the better rank. Moreover, the results of comparing the performance of OBE with other splitting measures based on Wilcoxon signed-rank test are shown in Table III. For each row, it represents testing results including the p-value when comparing OBE with each splitting measure via the specific evaluation metric. The symbol check mark denotes that OBE is significantly better than that splitting measure with the $(1-\alpha)100\%$ confidence level.

For comparing by precision, OBE yields the similar average performance to DCSM, AE and ME, which is better than Gini index and gain ratio. However, it offers the best result over the others for the average rank. From the statistical testing, it shows that OBE significantly outperforms Gini index and gain ratio with a 95% and 99% confidence level respectively. Nonetheless, it is not significantly different comparing with DCSM, AE and ME.

For comparing by recall, OBE yields the highest average performance better than other splitting measures. It also offers the best result over the others when comparing with the average rank. From the statistical testing, it shows that OBE significantly outperforms Gini index and DCSM with a 95% confidence level. It also provides a significant improvement over gain ratio and AE with a confidence level up to 99%. Nonetheless, it is not significantly different comparing with ME.



(a)



(b)

Fig. 6. The experimental results on real-world datasets comparing by the average performance (a) and the average rank (b).

TABLE III: THE STATISTICAL RESULTS BASED ON THE WILCOXON SIGNED-RANK TEST COMPARING OBE WITH OTHER SPLITTING MEASURES

Evaluation Metric	Splitting Measure	α 0.1	α 0.05	α 0.01	p-value
Precision	Gini index	✓	✓		0.018603
	Gain ratio	✓	✓	✓	0.002218
	DCSM				0.346522
	AE				0.476907
	ME				0.722108
Recall	Gini index	✓	✓		0.028056
	Gain ratio	✓	✓	✓	0.003702
	DCSM	✓	✓		0.041389
	AE	✓	✓	✓	0.003346
	ME				0.272095
F-measure	Gini index	✓	✓	✓	0.002209
	Gain ratio	✓	✓	✓	0.002218
	DCSM	✓	✓		0.028056
	AE	✓	✓	✓	0.007649
	ME	✓			0.065154
G-measure	Gini index	✓	✓	✓	0.004742
	Gain ratio	✓	✓	✓	0.002218
	DCSM	✓	✓		0.028056
	AE	✓	✓	✓	0.007649
	ME	✓			0.065154

For comparing by F-measure and G-measure, they exhibit the same results. OBE yields the highest average performance better than other splitting measures. It also offers the best result over the others when comparing with the average rank. From the statistical testing, it shows that OBE significantly outperforms Gini index, gain ratio and AE with a 99% confidence level. It also provides a significant improvement over DCSM and ME at 0.05 and 0.1 significant level (α) respectively.

The experimental results confirm that the conventional splitting measures like Gini index and gain ratio are not suitable to classify the imbalanced datasets. For the splitting measures proposed to deal with the class imbalanced problem like AE and ME including DCSM, they show impressive results in terms of precision which is the same for OBE. This happens because they concentrate on the region of minority class avoiding the majority classes' region. Therefore, the majority instances are less likely to be defined as the minority class, making them to obtain high precision. However, DCSM and AE along with ME show inferior results when considered in terms of recall. This happens because of focusing too much on the minority class may cause an overfit phenomenon. Hence, the value of recall corresponding to the minority class is inferior. However, this incident does not happen to OBE due to its mechanism that balancing the proportion of each class. All classes still receive attention based on their weights in the process of selecting the splitting condition, which will avoid the overfitting problem.

V. CONCLUSIONS

This paper presents a new splitting measure for inducing the decision tree classifier, named class overlapping-balancing entropy (OBE), to handle the class imbalanced problem. It arises from expanding the interesting concept of minority entropy (ME). OBE employs the proportion of the weighted values corresponding to the instances in each class as a calculation component. The overlapping weighting function (α_j) and the balancing weighting function (β) are proposed for assigning the weighted values based on the overlapping region between classes and the proportion of each class respectively. Theoretically, the weighted values given to each instance correspond to the two initial principal assumptions, i.e., 1) the weights of instances locating in the overlapping region are less than the weights of other instances from the same class and 2) the total weights in each class are equal.

The improved performance to classify an imbalanced dataset of Shannon's entropy using OBE is shown by two collections of experiments which are synthetic datasets and real-world datasets from UCI. It shows that OBE significantly outperforms the traditional splitting measures of decision trees. For all evaluation metrics, it provides significantly better results than Gini index and gain ratio. Importantly, the overfitting problem found in the splitting measure designed for an imbalanced dataset specifically like AE and ME including DCSM does not occur to OBE, which shows from the recall improvement. In term of precision, OBE also shows impressive results which has the lowest average rank compared to other splitting measures which implies that it has notably better overall performances via F-measure and G-measure.

Although OBE is highly successful in handling the class imbalanced problem, it is not able to work with a dataset containing categorical attributes including multiple classes which are the original important feature of the decision tree model. The proposed weights assignment needs to be generalized to be able to apply for a multi-class imbalanced dataset with all types of attributes.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

Author Contributions

This research paper is carried by Artit Sagoolmuang under the guidance and support of Krung Sinapiromsaran. Both authors have equally contributed to this work. Artit Sagoolmuang presented the concept of OBE, implemented the algorithm, and drafted the manuscript, while the analysis of the results and the editing of the manuscript are done by Krung Sinapiromsaran.

Acknowledgment

This research is favorably supported by the Applied Mathematics and Computational Science (AMCS) Program, Department of Mathematics and Computer Science, Faculty of Science, Chulalongkorn University, and the Graduate School of Chulalongkorn University, including the Science Achievement Scholarship of Thailand (SAST).

References

[1] X. Wu, V. Kumar, J. R. Quinlan *et al.*, "Top 10 algorithms in data mining," *Knowledge and information systems*, vol. 14, no. 1, pp. 1-37, 2008.
[2] 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," *Information Sciences*, vol. 250, pp. 113-141, 2013.
[3] Y. Sahin, S. Bulkan, and E. Duman, "A cost-sensitive decision tree approach for fraud detection," *Expert Systems with Applications*, vol. 40, no. 15, pp. 5916-5923, 2013.
[4] W. Wei, J. Li, L. Cao, Y. Ou, and J. Chen, "Effective detection of sophisticated online banking fraud on extremely imbalanced data," *World Wide Web*, vol. 16, no. 4, pp. 449-475, 2013.
[5] B. Krawczyk, M. Galar, Ł. Jeleń, and F. Herrera, "Evolutionary undersampling boosting for imbalanced classification of breast cancer malignancy," *Applied Soft Computing*, vol. 38, pp. 714-726, 2016.
[6] S.-H. Bae and K.-J. Yoon, "Polyp detection via imbalanced learning and discriminative feature learning," *IEEE Transactions on Medical Imaging*, vol. 34, no. 11, pp. 2379-2393, 2015.
[7] S. Hajian, J. Domingo-Ferrer, and A. Martinez-Balleste, "Discrimination prevention in data mining for intrusion and crime detection," in *Proc. 2011 IEEE Symposium on Computational Intelligence in Cyber Security*, 2011, pp. 47-54.
[8] E. Ramentol, I. Gondres, S. Lajes, R. Bello, Y. Caballero, C. Cornelis, and F. Herrera, "Fuzzy-rough imbalanced learning for the diagnosis of high voltage circuit breaker maintenance: The smote-frst-2t algorithm," *Engineering Applications of Artificial Intelligence*, vol. 48, pp. 134-139, 2016.
[9] P. C. Lane, D. Clarke, and P. Hender, "On developing robust models for favourability analysis: Model choice, feature sets and imbalanced data," *Decision Support Systems*, vol. 53, no. 4, pp. 712-718, 2012.
[10] M. Hlosta, R. Str´ız, J. Kupc´ık, J. Zendulka, and T. Hruska, "Constrained classification of large imbalanced data by logistic regression and genetic algorithm," *International Journal of Machine Learning and Computing*, vol. 3, no. 2, p. 214-218, 2013.
[11] R. Harliman and K. Uchida, "Data-and algorithm-hybrid approach for imbalanced data problems in deep neural network," *International Journal of Machine Learning and Computing*, vol. 8, no. 3, pp. 208-213, 2018.
[12] C. W. Gini, "Variability and mutability, contribution to the study of statistical distributions and relations. studi cconomico-giuridici della r. universita de cagliari (1912). Reviewed in: Light, rj, margolin, bh: An analysis of variance for categorical data," *J. American Statistical Association*, vol. 66, pp. 534-544, 1971.
[13] C. E. Shannon, "A mathematical theory of communication," *Bell system Technical Journal*, vol. 27, no. 3, pp. 379-423, 1948.
[14] S. Marcellin, D. A. Zighed, and G. Ritschard, "An asymmetric entropy measure for decision trees," in *Proc. 11th Conference on Information Processing and Management of Uncertainty in Knowledge-Based, Systems*, 2006.
[15] P. Lenca, S. Lallich, T.-N. Do, and N.-K. Pham, "A comparison of different off-centered entropies to deal with class imbalance for decision trees," in *Proc. Pacific-Asia Conference on Knowledge Discovery and Data Mining*, Springer, 2008, pp. 634-643.
[16] P. Lenca, S. Lallich, and B. Vaillant, "Construction of an off-centered entropy for the supervised learning of imbalanced classes: Some first results," *Communications in Statistics Theory and Methods*, vol. 39, no. 3, pp. 493-507, 2010.
[17] R. Guermazi, I. Chaabane, and M. Hammami, "Aecid: Asymmetric entropy for classifying imbalanced data," *Information Sciences*, vol. 467, pp. 373-397, 2018.
[18] T. Dietterich, M. Kearns, and Y. Mansour, "Applying the weak learning framework to understand and improve c4. 5," in *Proc. ICML*, 1996, pp. 96-104.
[19] C. Drummond and R. C. Holte, "Exploiting the cost (in) sensitivity of decision tree splitting criteria," in *Proc. ICML*, 2000, vol. 1.
[20] D. A. Cieslak and N. V. Chawla, "Learning decision trees for unbalanced data," in *Proc. Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, Springer, 2008, pp. 241-256.
[21] D. A. Cieslak, T. R. Hoens, N. V. Chawla, and W. P. Kegelmeyer, "Hellinger distance decision trees are robust and skew-insensitive," *Data Mining and Knowledge Discovery*, vol. 24, no. 1, pp. 136-158, 2012.
[22] S. Boonamnuay, N. Kerdprasop, and K. Kerdprasop, "Classification and regression tree with resampling for classifying imbalanced data," *International Journal of Machine Learning and Computing*, vol. 8, no. 4, pp. 336-340, 2018.
[23] K. Boonchuay, K. Sinapiromsaran, and C. Lursinsap, "Decision tree induction based on minority entropy for the class imbalance problem," *Pattern Analysis and Applications*, vol. 20, no. 3, pp. 769-782, 2017.
[24] L. Breiman, *Classification and Regression Trees*, Routledge, 2017.
[25] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81-106, 1986.
[26] J. R. Quinlan, *C4. 5: Programs for Machine Learning*, Elsevier, 2014.
[27] B. Chandra, R. Kothari, and P. Paul, "A new node splitting measure for decision tree construction," *Pattern Recognition*, vol. 43, no. 8, pp. 2725-2731, 2010.
[28] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management*, vol. 45, no. 4, pp. 427-437, 2009.
[29] G. W. Corder and D. I. Foreman, *Nonparametric Statistics: A Step-by-Step Approach*, John Wiley & Sons, 2014.
[30] C. Blake and C. Merz, *Uci Repository of Machine Learning Databases*, 1998.

Artit Sagoolmuang received the B.Sc. degrees (first-class honours) in mathematics from Kasetsart University, Bangkok, Thailand, in 2014 and the M.S. degree in applied mathematics and computational science from Chulalongkorn University, Bangkok, Thailand, in 2016. Since 2017, he has been a Ph.D. candidate in Applied Mathematics and Computational Science Program at Chulalongkorn University. His research interests include data mining and machine learning algorithm especially in handling the class imbalanced problem.

Krung Sinapiromsaran received his B.S. in mathematics from Chulalongkorn University, his M.S. and Ph.D. in computer science from the University of Wisconsin-Madison. He is currently an assistant professor in the Department of Mathematics, Chulalongkorn University. His ongoing research works are related to deep learning, machine learning, artificial intelligence, data mining, knowledge discovery, and optimization.