Frontiers of Information Technology & Electronic Engineering
www.jzus.zju.edu.cn; engineering.cae.cn; www.springerlink.com
ISSN 2095-9184 (print); ISSN 2095-9230 (online)
E-mail: jzus@zju.edu.cn

One-against-all-based Hellinger distance decision tree for multiclass imbalanced learning*

Minggang DONG[1,2], Ming LIU[1,2], Chao JING[†‡1,2,3]

[1]*School of Information Science and Engineering, Guilin University of Technology, Guilin 541004, China*

[2]*Guangxi Key Laboratory of Embedded Technology and Intelligent System, Guilin 541004, China*

[3]*Guangxi Key Laboratory of Trusted Software, Guilin University of Electronic Technology, Guilin 541004, China*

[†]E-mail: jingchao@glut.edu.cn

Received Aug. 17, 2020; Revision accepted Feb. 14, 2021; Crosschecked Nov. 9, 2021

Abstract: Since traditional machine learning methods are sensitive to skewed distribution and do not consider the characteristics in multiclass imbalance problems, the skewed distribution of multiclass data poses a major challenge to machine learning algorithms. To tackle such issues, we propose a new splitting criterion of the decision tree based on the one-against-all-based Hellinger distance (OAHD). Two crucial elements are included in OAHD. First, the one-against-all scheme is integrated into the process of computing the Hellinger distance in OAHD, thereby extending the Hellinger distance decision tree to cope with the multiclass imbalance problem. Second, for the multiclass imbalance problem, the distribution and the number of distinct classes are taken into account, and a modified Gini index is designed. Moreover, we give theoretical proofs for the properties of OAHD, including skew insensitivity and the ability to seek a purer node in the decision tree. Finally, we collect 20 public real-world imbalanced data sets from the Knowledge Extraction based on Evolutionary Learning (KEEL) repository and the University of California, Irvine (UCI) repository. Experimental and statistical results show that OAHD significantly improves the performance compared with the five other well-known decision trees in terms of Precision, F-measure, and multiclass area under the receiver operating characteristic curve (MAUC). Moreover, through statistical analysis, the Friedman and Nemenyi tests are used to prove the advantage of OAHD over the five other decision trees.

Key words: Decision trees; Multiclass imbalanced learning; Node splitting criterion; Hellinger distance; One-against-all scheme

https://doi.org/10.1631/FITEE.2000417 **CLC number:** TP301

1 Introduction

Numerous lines of evidence have shown that decision tree is one of the most popular classifiers (Wu et al., 2008; Sharmin et al., 2019), with characteristics of high efficiency, simplicity, and interpretability (Cieslak et al., 2012). In a decision tree, there is a tree structure model that consists of one root node, multiple internal nodes, and some leaf nodes. Given a training data set, the decision tree uses the splitting criterion to partition the data set into child nodes (internal and leaf nodes) recursively until a stopping condition is meet.

Originally, the decision tree was developed to solve the problem of balanced classification. There are two representative decision trees, CART and C4.5 (Breiman et al., 1984; Quinlan, 1986), which have shown pronounced capability in dealing with

‡ Corresponding author

* Project supported by the National Natural Science Foundation of China (Nos. 61802085 and 61563012), the Guangxi Provincial Natural Science Foundation, China (Nos. 2021GXNSFAA220074 and 2020GXNSFAA159038), the Guangxi Key Laboratory of Embedded Technology and Intelligent System Foundation, China (No. 2018A-04), and the Guangxi Key Laboratory of Trusted Software Foundation, China (No. kx202011)

ORCID: Minggang DONG, https://orcid.org/0000-0001-7078-3942; Chao JING, https://orcid.org/0000-0002-4695-8746

balanced classification problems. Due to the importance of the splitting criterion, based on the impurity, a splitting criterion is proposed for each decision tree. These decision trees have benefited from their proposed splitting criterion while dealing with the balanced classification problem. However, these methods face a problem in processing data sets with distinct classes. Chandra et al. (2010) proposed the distinct class based splitting measure (DCSM) relying on the number of distinct classes. In this splitting criterion, DCSM takes into account not only the distribution of each class but also the number of distinct classes in a partition. By doing so, DCSM has better performance in classification. Meanwhile, strict proof has been provided to show the advantages of DCSM in terms of its properties such as being convex and well-behaved (Safavian and Landgrebe, 1991; Kotsiantis, 2013; Osei-Bryson, 2014).

However, these methods have limitations in solving the multiclass imbalance problem. Using the prior probability of classes to compute the splitting criteria in these decision trees leads to a poor performance of minority classes (Flach, 2003; Akash et al., 2019). In the multiclass imbalanced classification problem, the minority classes occupy a small portion but are more important than the majority classes; thus, there is a challenge for traditional decision trees to deal with data sets containing minority classes. Therefore, it is crucial to improve the performance of the minority classes under the multiclass imbalanced classification problem.

To cope with the imbalanced classification problem, some decision trees have been proposed to improve the performance (Cieslak and Chawla, 2008; Liu et al., 2010; Boonchuay et al., 2017; Akash et al., 2019; Su and Cao, 2019). Liu et al. (2010) designed a class confidence proportion decision tree, introducing a new measure (class confidence proportion) without bias to the majority classes. However, this method can be used only to solve the two-class imbalance problem. To solve the multiclass imbalance problem, Akash et al. (2019) presented a weighted internode Hellinger distance (iHDw) based decision tree. Considering the multiclass imbalance problem, iHDw adopts the weighted squared Hellinger distance to measure the difference in class distribution between the parent node and child nodes rather than using the prior probability of classes. In this way, the splitting criterion with iHDw can avoid being sensitive to class imbalance. Moreover, Su and Cao (2019) integrated the Hellinger distance (Kailath, 1967; Cichocki and Amari, 2010) and KL divergence (Feng et al., 2019; Wan et al., 2020) into the lazy decision tree to solve the imbalanced classification problem. Cieslak and Chawla (2008) proposed a new decision tree, called the Hellinger distance decision tree (HDDT), which can solve the imbalance problem. In their work, the Hellinger distance (Kailath, 1967; Cichocki and Amari, 2010) was regarded as a splitting criterion in HDDT, showing better performance in the imbalanced classification problem; however, there is a limitation in their work using the Hellinger distance, because the Hellinger distance has a problem in identifying the difference between two different splits in the multiclass imbalance problem. Moreover, the splitting criterion does not consider the class distribution or the number of distinct classes.

We hereby consider a situation to illustrate the defect of the Hellinger distance. For example, for a certain node in the decision tree, we suppose that there are a number of samples with five classes notated with A, B, C, D, and E. The number of samples in each class is 100, 40, 20, 10, and 10. We assume that there are two splits for dividing these classes. One split divides all the samples in classes A and B (totaling 140) to the left child node, and the remaining 40 samples in the other classes are classified to the right child node. For the other split, 100 samples in class A and half of the number of samples in the other classes are categorized under the left child node with 140 samples; meanwhile, the leftover 40 samples from classes B to E are classified under the right child node. Because HDDT can be used only to compute the distance between two classes, it regards the multiclass problem as a two-class problem. If we suppose that class A is the positive class, the other classes are divided into the negative class. Therefore, we can obtain uniform class distribution by these two splits, obtaining the same results through the two splits by HDDT. Obviously, we can find that HDDT is inappropriate for solving the multiclass imbalanced classification problem. The main reason is that the Hellinger distance can solve only the two-class classification problem; it has limitations in identifying some differences in the multiclass imbalance problem. Furthermore, this process does not tackle the issues of the multiclass distribution and

the number of distinct classes, which are critical in the multiclass imbalance problem.

To address the insufficiencies of the methods mentioned above, in this study we propose a one-against-all-based Hellinger distance (OAHD) decision tree to solve the multiclass imbalanced classification problem. Due to the defect in the Hellinger distance in calculating the distance between multiple classes, we introduce the scheme of one-against-all (Anand et al., 1995) to the process of computing the splitting criterion of OAHD. During the computing process, we adopt the decomposition scheme so that OAHD can be extended to deal with the multiclass problem. We also consider the issue of purity of node in the decision tree. To tackle this issue, we take into account the number of distinct classes and the distribution of the multiclass imbalance problem without considering the prior probability of the classes. Meanwhile, we modify the Gini index to incorporate it into the multiclass imbalance problem. Moreover, we strictly prove that OAHD has the skew-insensitivity property, and that a purer node is sought by the splitting criterion. Finally, we collect 20 public real-world imbalanced data sets available from the Knowledge Extraction based on Evolutionary Learning (KEEL) repository (Alcala-Fdez et al., 2011) and the University of California, Irvine (UCI) repository (Asuncion, 2007). Experimental and statistical results demonstrate that the proposed OAHD outperforms the five unpruned decision trees in terms of various metrics.

2 Background

Here, we discuss several splitting criteria for decision trees related to our proposed criterion.

2.1 Information gain

The CART decision tree (Breiman et al., 1984) and the C4.5 decision tree (Quinlan, 1986) are the classical algorithms that deal with a balanced data set. The intuition of CART is to split the attribute that reduces impurity the most. Consider a parent node u with V child nodes, and consider that there are C distinct classes in node u; then, the splitting criterion (named the Gini index) is calculated as

$$\mathrm{Gini}(x_j)=\sum_{v=1}^{V}\left[1-\sum_{k=1}^{C}\left(\frac{N^v_{\omega_k}}{N^v}\right)^2\right], \tag{1}$$

$N^v_{\omega_k}$ denotes the number of samples of a class ω_k in partition v, and N^v denotes the total number of samples in partition v.

Similar to CART, C4.5 is based on choosing a partitioning that has the largest decrease in the information gain ratio. The information gain based on an attribute x_j is defined as

$$\begin{aligned}\mathrm{Gain}(x_j)=&-\sum_{k=1}^{C}\left[\frac{N^u_{\omega_k}}{N^u}\log_2\left(\frac{N^u_{\omega_k}}{N^u}\right)\right]\\&-\sum_{v=1}^{V}\left[\frac{N^v}{N^u}\sum_{k=1}^{C}\left(-\frac{N^v_{\omega_k}}{N^v}\log_2\left(\frac{N^v_{\omega_k}}{N^v}\right)\right)\right],\end{aligned} \tag{2}$$

where N^u denotes the number of samples in node u.

Since Eq. (2) favors the attribute with a larger number of values, Quinlan (1986) normalized the information gain to introduce a new splitting criterion (the information gain ratio), which is defined as

$$\mathrm{Gain}\left(R(x_j)\right)=\frac{\mathrm{Gain}(x_j)}{-\sum\limits_{v=1}^{V}\left[\frac{N^v}{N^u}\log_2\left(\frac{N^v}{N^u}\right)\right]}. \tag{3}$$

These decision trees show better performance while dealing with the balanced classification problem. However, for a skewed class distribution, the above-mentioned methods are inadequate for improving the performance.

2.2 Distinct class based splitting measure

Chandra et al. (2010) proposed a new splitting criterion called DCSM, which not only emphasizes the proportion of each distinct class but also pays attention to the number of distinct classes in a partition. Considering a splitting node u (parent node) with V partitions (child nodes), for a given attribute x_j, the splitting criterion $M(j)$ is calculated as

$$\begin{aligned}M(j)=&\sum_{v=1}^{V}\left\{\frac{N^v}{N^u}D(v)\exp\left(D(v)\right)\right.\\&\left.\cdot\sum_{k=1}^{C}\left[a^v_{\omega_k}\exp\left(\delta^v(1-(a^v_{\omega_k})^2)\right)\right]\right\},\end{aligned} \tag{4}$$

where $D(v)$ represents the number of distinct classes in partition v, δ^v is equal to $D(v)/D(u)$, and $a^v_{\omega_k}$ is the probability of class ω_k in partition v, i.e., $N^v_{\omega_k}/N^v$.

There are two critical parts in Eq. (4). One is $D(v)\exp(D(v))$, which considers the number of

distinct classes in each child node. As the number of distinct classes increases, the first part will increase sharply. Thus, the purer partition will be preferred. The other is the summation of $a^v_{\omega_k}\exp(\delta^v(1-(a^v_{\omega_k})^2))$, $v=1,2,\ldots,V$. First, as δ^v decreases, the impurity of the partition will decrease, and then $1-(a^v_{\omega_k})^2$ decreases when there are more samples of a class compared to the total number of samples in the partition. Therefore, this measure is intended to favor the purer partition. Compared with other split measures, DCSM takes into account the number of distinct classes. The limitation of DCSM is similar to those of the above-mentioned decision trees. It has a preference for the majority classes, resulting in poor performance for the minority classes.

2.3 Hellinger distance decision tree

Cieslak and Chawla (2008) introduced a new splitting criterion (Hellinger distance) to solve the imbalanced classification problem. The main difference between HDDT and other splitting measures is that HDDT is insensitive when dealing with a skewed class distribution. In HDDT, assuming that there is a two-class problem (class X_+ and class X_-) and dividing all continuous features into V partitions, a feature is selected as a splitting attribute when it achieves the largest Hellinger distance between class X_+ and class X_-. The Hellinger distance is defined as

$$D_{\mathrm{H}}(X_+\|X_-)=\sqrt{\sum_{j=1}^{V}\left(\sqrt{\frac{|X_{+j}|}{|X_+|}}-\sqrt{\frac{|X_{-j}|}{|X_-|}}\right)^2}, \tag{5}$$

where $|X_{+j}|$ and $|X_{-j}|$ represent the numbers of samples of classes X_+ and X_- in partition j respectively, and $|X_+|$ and $|X_-|$ denote the numbers of samples of classes X_+ and X_- in all partitions respectively.

HDDT is strongly considered to be skew-insensitive because it does not use the prior probability of a class in the distance calculation (Abdi and Hashemi, 2016; Akash et al., 2019). Nevertheless, the splitting criterion essentially captures the differences of the feature values only for the two classes without considering the multiclass imbalanced classification problem. Therefore, it is necessary to design a splitting criterion to address the multiclass imbalanced classification problem.

2.4 Weighted internode Hellinger distance

Akash et al. (2019) used the weighted squared Hellinger distance to measure the difference between the parent node and each child node. Though the proposed selection weight is combined with the squared Hellinger distance, iHDw can obtain a purer child node in a split partition. iHDw can be defined as

$$\mathrm{iHDw}=q_{\mathrm{L}}D^2_{\mathrm{H}}(P_{\mathrm{L}}\|P)\omega_{\mathrm{L}}+q_{\mathrm{R}}D^2_{\mathrm{H}}(P_{\mathrm{R}}\|P)\omega_{\mathrm{R}}, \tag{6}$$

where

$$D^2_{\mathrm{H}}(P_t\|P)=1-\sum_{j=1}^{C}\sqrt{p_{tj}p_j},$$

$$\omega_t=1-\prod_{j=1}^{k}\frac{q_t p_{tj}}{p_j},\quad t\in\{\mathrm{L},\mathrm{R}\},$$

where q_{L} and q_{R} denote the ratios of the numbers of samples in the left and right child nodes to the number of samples in the parent node, respectively. The function of $D^2_{\mathrm{H}}(P_t\|P)$ is to maximize the distance between the distributions of the two child nodes to make a mutually exclusive and skew-insensitive partition. P_t and P indicate the child nodes and parent node, respectively. C is the number of distinct classes. p_{tj} and p_j represent the proportion from class j in the child node and the parent node, respectively. ω_t is not dependent on the prior probability of the classes, resulting in nonbiased majority classes, and ω_t increases linearly with the growth of dissimilarity between the parent node and the child node to divide a purer partition.

3 The proposed algorithm

3.1 Details of the algorithm

Due to the defect in current research, by extending the Hellinger distance, we propose a splitting criterion OAHD to solve the multiclass imbalanced classification problem. Moreover, for each partition, the impurity of the training patterns and the distribution of the multiple classes are considered to seek a purer child node.

The Hellinger distance is a divergence measure and a member of the α divergence family (Kailath, 1967; Cichocki and Amari, 2010). Considering two discrete probabilities P and Q, the definition of

the Hellinger distance can be given as in Eq. (5). However, for the multiclass imbalance problem, the Hellinger distance does not work. To address this problem, we propose a new splitting criterion OAHD, which is based on the Hellinger distance. We use OAHD to calculate a better split threshold between two child nodes. The purpose of OAHD is that the samples in the two child nodes are divided into mutually exclusive regions, and the node is sufficiently pure when OAHD is maximized. Considering a parent node u and the corresponding child node v, $v \in \{\mathrm{L}, \mathrm{R}\}$, for each feature j, the proposed splitting criterion is defined as follows:

$$H_{\mathrm{D}}(j) = \sum_{v} \frac{\omega_{\theta}^{v}}{\omega_{\theta}^{u}} \exp\left(\sqrt{\frac{N_{\theta}^{v}}{N_{\theta}^{u}}} - \sqrt{\frac{N_{\overline{\theta}}^{v}}{N_{\overline{\theta}}^{u}}}\right)^{2} \cdot \frac{\beta^{u} - \beta^{v}}{\beta^{u}} \exp\left(\sum_{k=1}^{C} W_{k}^{2}\right), \tag{7}$$

where

$$W_{k} = \frac{\frac{N_{k}^{v}}{N_{k}^{u}}}{\sum_{i=1}^{C} \frac{N_{i}^{v}}{N_{i}^{u}}}, \quad \frac{\omega_{\theta}^{v}}{\omega_{\theta}^{u}} = \frac{\frac{N_{\theta}^{v}}{N_{\theta}^{u}} + \frac{N_{\overline{\theta}}^{v}}{N_{\overline{\theta}}^{u}}}{\sum_{v}\left(\frac{N_{\theta}^{v}}{N_{\theta}^{u}} + \frac{N_{\overline{\theta}}^{v}}{N_{\overline{\theta}}^{u}}\right)}.$$

Here, N denotes the number of samples in a node; for example, N_{i}^{v} denotes the number of samples of class i in node v. β^{u} denotes the number of distinct classes in the parent node, and β^{v} represents the number of distinct classes, with $W_{k} \geq 1/2$ for the child node. θ and $\overline{\theta}$ denote the selected minority class and the remaining classes, respectively.

The splitting criterion is composed of two parts. The first part is the summary of $\frac{\omega_{\theta}^{v}}{\omega_{\theta}^{u}} \exp\left(\sqrt{\frac{N_{\theta}^{v}}{N_{\theta}^{u}}} - \sqrt{\frac{N_{\overline{\theta}}^{v}}{N_{\overline{\theta}}^{u}}}\right)^{2}$, which is skew-insensitive. Previous works have always decomposed a C-class imbalance problem into C two-class imbalance problems. Differently, we apply the one-against-all scheme to calculate the multiclass Hellinger distance directly. In this process, the original multiclass imbalanced data set is used to calculate OAHD directly. For each class, OAHD regards the selected minority class as the positive class (θ), and the remaining classes as the negative class ($\overline{\theta}$); by doing this, a one-against-all-based Hellinger distance of the selected minority class is obtained. Then, the threshold of the maximum one-against-all-based Hellinger distance is viewed as the optimal split threshold. $N_{\theta}^{v}/N_{\theta}^{u}$ and $N_{\overline{\theta}}^{v}/N_{\overline{\theta}}^{u}$ represent the ratios of the number of samples in the child node to the number of samples in the parent node of classes θ and $\overline{\theta}$, respectively. In previous splitting criteria, the term ω^{v}/ω^{u} always denotes the ratio of the number of samples in the child node to the number of samples in the parent node. However, due to the enormous difference in the number of samples between the majority and minority classes, this index does not work. Therefore, in our method, $\omega_{\theta}^{v}/\omega_{\theta}^{u}$ represents the proportion of the weight between the parent node and the child node without using the prior probability of the classes. As the difference in the distribution probability between classes θ and $\overline{\theta}$ increases, the partition becomes purer. Therefore, purer partition is preferred by this part.

The second part is $\frac{\beta^{u} - \beta^{v}}{\beta^{u}} \exp\left(\sum_{k=1}^{C} W_{k}^{2}\right)$, which is independent of the prior probability of the classes. Thus, it is not biased toward the majority class. If most samples of each class are split into the same child node, we assume $\beta^{v} = \beta^{u}$, $v \in \{\mathrm{L}, \mathrm{R}\}$. W_{k} is the proportion of class k in a child node. Different from the CART decision tree, in the computation of W_{k}, the proportion of the weight between a parent node and a child node of class k is used instead of the number of samples of class k, since in the absence of the prior probability of the classes, the decision tree is insensitive to the skewed distribution. As β^{v} decreases, the number of distinct classes with $W_{k} \geq 1/2$ decreases, and $(\beta^{u} - \beta^{v})/\beta^{u}$ gains a greater value. Thereby, the node becomes purer. Furthermore, for a certain class k, if W_{k} is asymptotic to one, the sum of W_{k}^{2} ($k = 1, 2, \ldots, C$) will increasingly grow with the child node becoming purer. Therefore, the combination is skew-insensitive and favors purer partition.

The main process of calculating the one-against-all-based Hellinger distance is described in Algorithm 1.

3.2 Proof of the properties of the proposed algorithm

Several properties are used to characterize the node splitting criterion, which is expected for a good splitting criterion. OAHD has the following basic properties:

1. OAHD is insensitive to the skewness of the class distribution; therefore, this splitting criterion

Algorithm 1 OAHD

Input: f (a feature in the imbalanced data set T) and j (indicating that f is the j^{th} feature of T)

Output: H_{D} (the maximum value of OAHD calculated using Eq. (7)) and T_{H} (the corresponding split threshold of H_{D})

$H_{\text{D}} \leftarrow -1$
Sort the value of feature f, and then divide it into N equal parts
for $k \leftarrow 1$ to C **do**
  for $i \leftarrow 1$ to N **do**
    $T_{\text{H}} \leftarrow \text{Threshold}(i)$
    As mentioned above, calculate the corresponding parameter
    $$\text{Cur}\_H_{\text{D}} = \sum_v \frac{\omega_\theta^v}{\omega_\theta^u} \exp\left(\sqrt{\frac{N_\theta^v}{N_\theta^u}} - \sqrt{\frac{N_{\overline{\theta}}^v}{N_{\overline{\theta}}^u}}\right)^2 \cdot \frac{\beta^u - \beta^v}{\beta^u} \exp\left(\sum_{k=1}^{C} W_k^2\right)$$
    if $\text{Cur}\_H_{\text{D}} > H_{\text{D}}$ **then**
      $H_{\text{D}} = \text{Cur}\_H_{\text{D}}$
      $T_{\text{H}} = \text{Cur}\_T_{\text{H}}$
    end if
  end for
end for

will not be biased toward the majority classes.

2. H_{D} is nonnegative, and it will obtain the lowest value, i.e., $H_{\text{D}} = 0$, which indicates that most samples of each class are split into the same child node; this will result in the node becoming more impure. Thus, a purer node can be sought by OAHD.

For the first property, we adopt the formulation of Flach (2003) in this study. This method can transform the splitting criterion into a formula containing the true positive rate (the ratio of the number of positive samples in the child node to the number of positive samples in the parent node) and the false positive rate (the ratio of the number of negative samples in the child node to the number of negative samples in the parent node), which can be used to measure the degree of sensitivity to skewness. Thus, this method can be used to measure the skew-sensitive of the splitting criterion. Vilalta and Oblinger (2000) had similar analysis. As mentioned above, without the prior probability of the classes, the Hellinger distance can also be related to the true positive rate and the false positive rate, which have been proved to be insensitive to the skewed class distribution (Vilalta and Oblinger, 2000; Flach, 2003). Thus, Eq. (5) can be written as follows:

$$D_{\text{H}} = \sqrt{(\sqrt{\text{tpr}} - \sqrt{\text{fpr}})^2 + (\sqrt{1-\text{tpr}} - \sqrt{1-\text{fpr}})^2}, \tag{8}$$

where tpr denotes the true positive rate and fpr denotes the false positive rate.

Proposition 1 OAHD has the skew-insensitive property.

Proof To explain Proposition 1, consider a data set with C classes ($k = 1, 2, \cdots, C$); the number of samples of class k in the left child node is denoted by tp_k^{L}, and tp_k^{R} is the number of samples of class k in the right child node; thus, the true positive rate tpr_k is equal to $\text{tp}_k^v \Big/ \sum_v^V \text{tp}_k^v$, $v \in \{\text{L}, \text{R}\}$. Suppose that class θ denotes the selected minority class, and that the remaining classes are recorded as class $\overline{\theta}$. From Eq. (7), for feature j, the splitting criterion of OAHD can be formulated as follows:

$$\begin{aligned} H_{\text{D}}(j) =& \frac{\text{tpr}_\theta^{\text{L}} + \text{tpr}_{\overline{\theta}}^{\text{L}}}{\sum_v^V (\text{tpr}_\theta^v + \text{tpr}_{\overline{\theta}}^v)} \exp\left(\sqrt{\text{tpr}_\theta^{\text{L}}} - \sqrt{\text{tpr}_{\overline{\theta}}^{\text{L}}}\right)^2 \\ & \cdot \frac{\beta^u - \beta^{\text{L}}}{\beta^u} \exp\left[\sum_{k=1}^{C} \left(\text{tpr}_k^{\text{L}} \Big/ \sum_{i=1}^{C} \text{tpr}_i^{\text{L}}\right)^2\right] \\ & + \frac{\text{tpr}_\theta^{\text{R}} + \text{tpr}_{\overline{\theta}}^{\text{R}}}{\sum_v^V (\text{tpr}_\theta^v + \text{tpr}_{\overline{\theta}}^v)} \exp\left(\sqrt{\text{tpr}_\theta^{\text{R}}} - \sqrt{\text{tpr}_{\overline{\theta}}^{\text{R}}}\right)^2 \\ & \cdot \frac{\beta^u - \beta^{\text{R}}}{\beta^u} \exp\left[\sum_{k=1}^{C} \left(\text{tpr}_k^{\text{R}} \Big/ \sum_{i=1}^{C} \text{tpr}_i^{\text{R}}\right)^2\right]. \end{aligned} \tag{9}$$

Since $(\beta^u - \beta^v)/\beta^u$ ($v \in \{\text{L}, \text{R}\}$) denotes the impact of the number of distinct classes on the process of splitting a node, where β^u denotes the number of distinct classes in the parent node, it is not necessary to use the prior probability of the classes. Therefore, to clarify Eq. (9), we set $(\beta^u - \beta^v)/\beta^u$ notated with D^v. $\sum_{k=1}^{C} \left(\text{tpr}_k^v \Big/ \sum_{i=1}^{C} \text{tpr}_i^v\right)^2$ ($v \in \{\text{L}, \text{R}\}$) is organized with two parts, a constant value and a true positive rate without the prior probability of the classes, where $\sum_{i=1}^{C} \text{tpr}_i^v$ is a constant value that equals S^v. Thus, it can be written as $\sum_{k=1}^{C} \left(\frac{\text{tpr}_k^v}{S^v}\right)^2 = \left(\frac{\text{tpr}_\theta^v}{S^v}\right)^2 + \sum_{k=1,k\neq\theta}^{C} \left(\frac{\text{tpr}_k^v}{S^v}\right)^2$, assuming that $\sum_{k=1,k\neq\theta}^{C} \left(\frac{\text{tpr}_k^v}{S^v}\right)^2$ is equal to DR^v (a constant

value). It is obvious that $S^{\mathrm{L}}+S^{\mathrm{R}}$ is equal to C, so we can set $S^{\mathrm{R}}=C-S^{\mathrm{L}}$. Similarly, D^{R} can be calculated from $1-D^{\mathrm{L}}$, where $\beta^v\neq\beta^u$ and $\sum_{v}^{V}(\mathrm{tpr}_\theta^v+\mathrm{tpr}_{\overline{\theta}}^v)=2$. Thus, Eq. (9) is eventually formulated as follows:

$$\begin{aligned}H_{\mathrm{D}}(j)=&\frac{\mathrm{tpr}_\theta^{\mathrm{L}}+\mathrm{tpr}_{\overline{\theta}}^{\mathrm{L}}}{2}\exp\left(\sqrt{\mathrm{tpr}_\theta^{\mathrm{L}}}-\sqrt{\mathrm{tpr}_{\overline{\theta}}^{\mathrm{L}}}\right)^2\\&\cdot D^{\mathrm{L}}\exp\left[\left(\frac{\mathrm{tpr}_\theta^{\mathrm{L}}}{S^{\mathrm{L}}}\right)^2+\mathrm{DR}^{\mathrm{L}}\right]\\&+\frac{\mathrm{tpr}_\theta^{\mathrm{R}}+\mathrm{tpr}_{\overline{\theta}}^{\mathrm{R}}}{2}\exp\left(\sqrt{\mathrm{tpr}_\theta^{\mathrm{R}}}-\sqrt{\mathrm{tpr}_{\overline{\theta}}^{\mathrm{R}}}\right)^2\\&\cdot(1-D^{\mathrm{L}})\exp\left[\left(\frac{\mathrm{tpr}_\theta^{\mathrm{R}}}{C-S^{\mathrm{L}}}\right)^2+\mathrm{DR}^{\mathrm{R}}\right].\end{aligned}\tag{10}$$

From Eq. (10), we can see that OAHD can be transformed as a pattern that contains only tpr and the constant values ($D^{\mathrm{L}}, S^{\mathrm{L}}$, and DR^v), without the prior probability of classes (Akash et al., 2019). Thus, OAHD is insensitive to the skewness of the class distribution of the data set (Vilalta and Oblinger, 2000; Flach, 2003).

Proposition 2 OAHD is nonnegative. It will obtain the lowest value, i.e., $H_{\mathrm{D}}=0$, when most of the samples of all classes are divided into the same child node, and it will seek a purer node.

Proof Through the calculation using Eq. (7), both critical parts are larger than 0, so H_{D} is nonnegative and the lowest value of H_{D} is equal to zero. Assuming a data set with three classes ($k=1,2,3$), for a feature j, we suppose that the split threshold divides 60% samples of each class to the left node and 40% to the right node. Eq. (11) shows the specific calculation:

$$\begin{aligned}H_{\mathrm{D}}(j)=&\sum_v\frac{\omega_\theta^v}{\omega_\theta^u}\exp\left(\sqrt{\frac{N_\theta^v}{N_\theta^u}}-\sqrt{\frac{N_{\overline{\theta}}^v}{N_{\overline{\theta}}^u}}\right)^2\\&\cdot\frac{\beta^u-\beta^v}{\beta^u}\exp\left(\sum_{k=1}^{C}W_k^2\right)\\=&\frac{3}{5}\exp\left(\sqrt{\frac{3}{5}}-\sqrt{\frac{3}{5}}\right)^2\frac{3-3}{3}\exp\left[\sum_{k=1}^{C}\left(\frac{1}{3}\right)^2\right]\\&+\frac{2}{5}\exp\left(\sqrt{\frac{2}{5}}-\sqrt{\frac{2}{5}}\right)^2\frac{3-3}{3}\exp\left[\sum_{k=1}^{C}\left(\frac{1}{3}\right)^2\right]\\=&0.\end{aligned}\tag{11}$$

As we can see from Eq. (11), when most samples of all classes are divided into the same child node, $(\beta^u-\beta^v)/\beta^u$ equals zero, so H_{D} achieves the lowest value of zero. In this situation, this node is disordered as impure.

Then, if $(\beta^u-\beta^v)/\beta^u\neq 0$, when the difference between $\sqrt{N_\theta^v/N_\theta^u}$ and $\sqrt{N_{\overline{\theta}}^v/N_{\overline{\theta}}^u}$ becomes greater, H_{D} will grow, so does the part of $\exp\left(\sum_{k=1}^{C}W_k^2\right)$. For example, considering a data set with three classes, there are two split thresholds A and B. For the split threshold A, 90% samples of each class are divided into the right child node, and the remaining samples are classified to the left child node. For the split threshold B, only 90% samples are selected from one class for the right child node, and 90% samples from the other classes are categorized to the left child node. Under these two split threshold rules, we can find that the node purity obtained by the split threshold B is better than that obtained by the split threshold A, and that a greater value can be obtained with the split threshold B. Thereby, the lowest value of OAHD is zero, and it will seek a purer node.

4 Evaluation

4.1 Imbalanced data sets

In this subsection, the performance of OAHD was evaluated using 20 public real-word multiclass imbalanced data sets, which were collected from two well-known public sources, namely, the KEEL and UCI repositories. Details of the multiclass imbalanced data sets are shown in Table 1. The imbalance ratio (IR) denotes the ratio of the number of samples of the largest majority class to the number of samples of the smallest minority class. The OAHD decision tree was compared with five unpruned decision tree classifiers, which included CART, C4.5, DCSM, internode Hellinger distance (iHD), and iHDw. Furthermore, to guarantee a fair comparison, all experiments were conducted using five-fold cross-validation and with 20 independent runs.

4.2 Performance measures

Precision and F-measure are widely discussed single-class metrics in imbalance problems (He and Garcia, 2009; Nekooeimehr and Lai-Yuen, 2016), and it is a suitable way to use them to evaluate the performance of the algorithm. In the next experiment,

Table 1 Description of the imbalanced data sets, including the name, size, number of attributes, number of classes, class distribution, and imbalance ratio (IR)

Number	Name	Size	Number of attributes	Number of classes	Class distribution	IR
1	ESL12vs3vs456vs7vs89	488	4	5	62/14/351/38/23	25.1
2	Heart	270	13	2	150/120	1.3
3	Liver	345	6	2	145/200	1.4
4	Wine	178	13	3	71/59/48	1.5
5	Glass	214	10	6	76/70/29/17/13/9	8.5
6	Automobile12vs345vs6	205	71	3	25/153/27	6.1
7	ERA	1000	4	9	142/181/172/88/158/18/92/31/118	10.0
8	ERA1vs2345vs7vs8vs9	1000	4	5	771/88/92/31/18	42.8
9	Yeast52	982	8	5	463/25/35/429/30	18.5
10	Plates-faults1	1941	27	7	158/190/391/72/55/402/673	12.2
11	Plates-faults3	1941	27	5	158/863/793/72/55	15.7
12	abalone8discre	2148	10	8	126/203/267/487/634/259/115/57	11.1
13	abalone10discre	2297	10	10	57/115/259/391/634/487/126/103/67/58	11.1
14	page-blocks	559	10	4	329/115/87/28	11.8
15	pendigits	1100	16	10	115/114/114/106/114/106/105/115/105/106	1.1
16	housing5	506	13	5	36/123/239/77/31	7.7
17	vertebral-column	310	6	3	60/150/100	2.5
18	vehicle-mc	846	18	3	199/429/218	2.2
19	vowel5	990	10	5	180/90/360/270/90	4.0
20	vowel7	990	10	8	90/90/90/90/180/90/180/180	2.0

Precision and F-measure were used only to evaluate the smallest minority class, and they are shown in Eqs. (12) and (14), respectively. There is another popular evaluation metric, namely, the multiclass area under the receiver operating characteristic (ROC) curve (MAUC), which has been extended by the area under the ROC curve (AUC) (Hanley and McNeil, 1982; Bradley, 1997; Ali et al., 2019) to evaluate the multiclass imbalanced classification problem. MAUC is calculated in Eq. (15).

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}}, \tag{12}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}}, \tag{13}$$

$$\text{F-measure} = \frac{(1+\beta)^2\text{Recall}\cdot\text{Precision}}{\beta^2\text{Recall}+\text{Precision}}, \tag{14}$$

$$\text{MAUC} = \frac{2}{|C|(|C|+1)}\sum_{i<j}\frac{A(C_i|C_j)+A(C_j|C_i)}{2}, \tag{15}$$

where TP is the number of true positive samples (actually positive class, and classified as positive class), FP represents the number of false positive samples (actually negative class, but classified as positive class), FN denotes the number of false negative samples (actually positive class, but classified as negative class), and parameter β is usually set to one. Here, $|C|$ indicates the number of classes, and $A(C_i|C_j)$ and $A(C_j|C_i)$ are the AUC values between classes C_i and C_j in the two-class imbalance problem; however, for the multiclass imbalance problem, $A(C_i|C_j)$ may not equal $A(C_j|C_i)$.

To further evaluate the statistical differences between OAHD and the comparative methods on multiple imbalanced data sets, we conducted the Friedman test (Friedman, 1937, 1940) with the corresponding post-hoc test, i.e., Nemenyi test (Nemenyi, 1963). The Friedman statistic χ_{F}^2 is calculated as follows:

$$\chi_{\text{F}}^2 = \frac{12N}{k(k+1)}\left[\sum_{i=1}^{k}R_i^2 - \frac{k(k+1)^2}{4}\right], \tag{16}$$

where k is the number of compared classifiers and R_i is the mean rank of the i^{th} classifier on N data sets.

To compare the different classifiers upon multiple data sets, the Iman F-statistic (Iman and Davenport, 1980) is calculated from χ_{F}^2 as

$$F_{\text{F}} = \frac{(N-1)\chi_{\text{F}}^2}{N(k-1)-\chi_{\text{F}}^2}. \tag{17}$$

If the null hypothesis is rejected (the performances of all the classifiers are similar), the post-hoc Nemenyi test is adopted to find the classifier with performance significantly better than those of

the others. In the Nemenyi test, since the critical difference (CD) is an important value as defined in Eq. (18), when the difference between the mean rankings of the two classifiers is larger than CD, the performance of one classifier is significantly better than those of the others. The calculation of CD is given below:

$$\mathrm{CD} = q_\delta \sqrt{\frac{k(k+1)}{6N}}, \tag{18}$$

where q_δ is the crucial value for the two-tailed Nemenyi test for k classifiers at the δ significance level. In this study, by referring to the table, we know that at the 0.05 significance level, q_δ is equal to 2.850.

4.3 Experimental results

To show the effectiveness of OAHD, we provided the experimental results of the metrics Precision, F-measure, and MAUC. Tables 2–4 show the mean values of the three metrics of OAHD and the five other unpruned decision trees (CART, C4.5, DCSM, iHD, and iHDw); the best mean values on each data set are in bold. From Tables 2–4, it can be observed that the proposed OAHD outperformed the five other unpruned decision trees for these imbalanced data sets. For the metrics Precision, F-measure, and MAUC, OAHD obtained the best results in 11 out of the 20 multiclass imbalanced data sets.

The results are further summarized in Table 5, which shows the mean rankings of the six decision trees in terms Precision, F-measure, and MAUC on the 20 multiclass imbalanced data sets. For each data set, the algorithm with the best performance obtains the ranking of "1," and the ranking of "6" indicates the algorithm with the worst performance. When the rankings obtained by some algorithms are at the same level, the ranking will be equally assigned. For instance, when two algorithms are tied for the ranking of "1," the ranking result with "1.5" is allocated for each of them. As listed in Table 5, we can find that OAHD obtained the mean rankings of 1.775, 1.85, and 1.6 in terms Precision, F-measure, and MAUC, respectively, which were all the best mean rankings among the six decision trees.

4.4 Analysis of results

To further evaluate the statistical significance of OAHD compared to the five other unpruned decision trees, the Friedman test and the post-hoc Nemenyi test were applied in our experiments; the results are shown in Table 6. The Friedman test was used to compare the performances of OAHD and the five unpruned decision trees for all the data sets. The null hypothesis of the Friedman test is to observe the differences that occasionally occur in the performance

Table 2 Precision of the proposed OAHD and the five decision trees for the 20 imbalanced data sets

Number	Precision					
	CART	C4.5	DCSM	iHD	iHDw	OAHD
1	0.532 14	**0.721 43**	0.542 86	0.521 43	0.507 14	0.607 14
2	0.704 17	0.669 58	**0.719 17**	0.703 17	0.701 25	0.705 67
3	0.585 86	0.563 10	0.572 07	0.575 17	0.586 55	**0.612 41**
4	0.910 42	0.915 63	0.908 33	0.926 04	**0.927 08**	0.916 42
5	0.411 11	0.688 89	0.411 11	0.655 56	0.655 56	**0.716 67**
6	0.802 00	0.220 00	0.820 00	0.812 00	0.806 00	**0.840 00**
7	**0.777 78**	**0.777 78**	**0.777 78**	**0.777 78**	**0.777 78**	**0.777 78**
8	**0.777 78**	0.766 67	**0.777 78**	**0.777 78**	**0.777 78**	**0.777 78**
9	0.372 00	0.406 00	0.354 00	0.308 00	0.298 00	**0.452 00**
10	0.720 91	0.431 82	0.717 27	0.780 91	0.780 91	**0.814 55**
11	0.650 91	0.305 45	0.619 11	0.610 09	0.611 82	**0.706 36**
12	0.448 25	**0.516 67**	0.451 60	0.450 26	0.447 37	0.452 63
13	0.438 60	**0.556 14**	0.438 60	0.478 07	0.477 19	0.536 84
14	0.921 43	0.791 07	0.916 07	0.967 86	0.967 86	**0.978 57**
15	0.888 10	0.933 33	0.912 86	0.910 95	0.910 95	**0.935 71**
16	0.698 39	0.690 32	0.695 16	0.737 10	0.725 81	**0.803 23**
17	0.562 50	0.590 83	0.559 17	0.602 50	0.603 33	**0.632 50**
18	0.868 03	0.848 24	0.867 67	0.863 33	0.864 32	**0.887 19**
19	0.725 56	0.592 22	0.708 33	**0.796 11**	0.794 44	0.768 33
20	**0.933 33**	0.898 89	0.915 00	0.886 67	0.882 78	0.898 33

Best results are in bold

Table 3 F-measure of the proposed OAHD and the five decision trees for the 20 imbalanced data sets

Number	F-measure					
	CART	C4.5	DCSM	iHD	iHDw	OAHD
1	0.505 30	**0.655 61**	0.519 13	0.527 51	0.516 01	0.567 48
2	0.723 26	0.695 15	**0.731 30**	0.714 88	0.715 83	0.724 09
3	0.577 47	0.574 57	0.583 52	0.571 72	0.576 85	**0.592 08**
4	0.893 04	0.919 18	0.892 39	0.931 18	**0.931 23**	0.919 97
5	0.461 15	0.593 11	0.450 84	0.640 51	0.643 54	**0.712 75**
6	0.779 02	0.342 09	0.770 05	0.841 82	0.818 60	**0.881 63**
7	**0.700 00**	0.662 82	**0.700 00**	**0.700 00**	**0.700 00**	**0.700 00**
8	**0.700 00**	0.696 67	**0.700 00**	**0.700 00**	**0.700 00**	**0.700 00**
9	0.418 81	**0.528 59**	0.400 76	0.354 33	0.337 16	0.509 03
10	0.702 79	0.452 78	0.695 39	**0.755 38**	0.753 38	0.745 75
11	0.674 37	0.401 17	0.639 41	0.639 21	0.629 27	**0.718 38**
12	0.480 87	**0.515 25**	0.470 95	0.495 10	0.486 87	0.495 27
13	0.456 63	0.491 67	0.470 74	0.467 98	0.463 92	**0.496 88**
14	0.942 79	0.857 19	0.939 66	0.956 55	0.955 70	**0.974 97**
15	0.901 60	0.936 70	0.909 53	0.927 09	0.927 31	**0.938 30**
16	0.757 11	0.729 26	0.755 84	0.750 49	0.741 55	**0.792 73**
17	0.562 06	0.579 52	0.550 98	0.602 69	**0.607 41**	0.603 68
18	0.858 73	0.847 00	0.857 66	0.857 42	0.856 96	**0.865 68**
19	0.737 90	0.592 81	0.719 93	**0.799 36**	0.796 36	0.781 71
20	**0.933 34**	0.898 23	0.920 20	0.897 27	0.894 59	0.906 68

Best results are in bold

Table 4 MAUC of the proposed OAHD and the five decision trees for the 20 imbalanced data sets

Number	MAUC					
	CART	C4.5	DCSM	iHD	iHDw	OAHD
1	0.833 75	0.841 53	0.840 42	0.840 81	0.840 84	**0.843 43**
2	**0.802 71**	0.779 60	0.787 18	0.786 63	0.784 89	0.790 75
3	0.648 13	0.630 91	0.637 91	0.639 93	0.642 05	**0.648 26**
4	0.934 49	0.940 99	0.938 66	0.941 99	0.943 96	**0.944 73**
5	0.809 81	0.811 31	0.810 23	0.814 96	0.817 71	**0.818 41**
6	**0.905 52**	0.828 06	0.871 88	0.882 58	0.886 39	0.892 06
7	**0.710 23**	0.709 76	0.709 92	0.710 00	0.710 05	0.710 08
8	0.760 96	**0.761 34**	0.761 21	0.761 15	0.761 11	0.761 09
9	0.698 26	0.702 75	0.701 74	0.705 37	0.707 09	**0.709 42**
10	0.880 13	0.874 97	0.876 52	0.881 04	0.883 76	**0.884 13**
11	**0.852 54**	0.844 26	0.844 72	0.844 56	0.844 08	0.846 79
12	0.680 53	0.676 45	0.677 71	0.678 28	0.678 47	**0.681 56**
13	0.668 15	0.668 86	0.668 66	0.668 96	0.669 16	**0.671 26**
14	0.962 73	0.956 87	0.958 39	0.960 54	0.961 59	**0.963 35**
15	0.940 62	0.940 39	0.939 64	0.939 62	0.939 56	**0.945 98**
16	0.856 90	0.846 60	0.849 80	**0.857 64**	0.849 20	0.856 93
17	0.854 26	0.843 78	0.847 47	0.848 17	0.848 98	**0.860 09**
18	0.940 75	0.940 58	**0.946 08**	0.942 82	0.940 25	0.944 57
19	0.900 69	0.900 22	0.902 93	0.907 67	**0.915 32**	0.911 44
20	**0.920 14**	0.907 17	0.910 00	0.911 42	0.911 87	0.919 36

Best results are in bold

for these decision trees. If F_{F} calculated using Eq. (17) is not less than the critical value, the difference between these two algorithms is statistically significant, and the null hypothesis will be rejected.

As listed in Table 6, we can see that the null hypothesis of the Friedman test was rejected. In the following, we detail the calculation of the null hypothesis of the Friedman test. In Eq. (16), for the metric Precision, we can calculate the Friedman statistic χ^2_{F} according to these mean rankings, i.e., $\chi^2_{\mathrm{F}}(\mathrm{P}) = \dfrac{12\times 20}{6\times 7}\left(\sum\limits_{i=1}^{6} R_i^{\mathrm{P}} - \dfrac{6\times 7^2}{4}\right) =$

Table 5 Mean ranking of the proposed OAHD and the five decision trees for the three metrics on the 20 imbalanced data sets

Algorithm	Mean ranking		
	Precision	F-measure	MAUC
CART	3.875	3.80	3.35
C4.5	4.175	4.40	4.90
DCSM	3.925	4.00	4.30
iHD	3.525	3.35	3.50
iHDw	3.725	3.60	3.35
OAHD	**1.775**	**1.85**	**1.60**

Best results are in bold

Table 6 Friedman test and Nemenyi test for the six methods on the 20 data sets, with OAHD as the base classifier

Algorithm	Precision	F-measure	MAUC
Friedman test	Reject	Reject	Reject
CART	✓	✓	✓
C4.5	✓	✓	✓
DCSM	✓	✓	✓
iHD	✓		✓
iHDw	✓	✓	✓
OAHD	Base	Base	Base

The symbol "✓" indicates that OAHD outperforms the compared algorithm

21.7357. Here, R_i^{P} denotes the mean ranking of the algorithms on the evaluation criterion Precision. The Iman F-statistic was calculated using Eq. (17), i.e., $F_{\mathrm{F}}(\mathrm{P}) = \dfrac{19 \times 21.7357}{20 \times 5 - 21.7357} = 5.2767$, and the critical value for the 0.05 significance level was 2.3102 with regard to the table, where $F_{\mathrm{F}}(\mathrm{P}) > 2.3102$; thus, the null hypothesis was rejected. Similar to Precision, using Eq. (16), we can calculate the Friedman statistic χ_{F}^2 of the metrics F-measure and MAUC as $\chi_{\mathrm{F}}^2(\mathrm{F}) = 22.3143$ and $\chi_{\mathrm{F}}^2(\mathrm{M}) = 35.7429$. Later, the Iman F-statistic of the metrics F-measure and MAUC can be obtained using Eq. (17), i.e., $F_{\mathrm{F}}(\mathrm{F}) = 5.4575$ and $F_{\mathrm{F}}(\mathrm{M}) = 10.5687$. Explicitly, we can see that the χ_{F}^2 values of both the F-measure and MAUC were greater than the critical value (2.3102); therefore, the null hypotheses were rejected.

After the null hypotheses of the Friedman test for the metrics Precision, F-measure, and MAUC were rejected, we conducted the post-hoc Nemenyi test to find which decision tree performed better. The results are listed in Table 6. The symbol "✓" indicates that OAHD outperforms the compared algorithm. Under the six classifiers and 20 imbalanced data sets, at the 0.05 significance level, q_δ was equal to 2.850. Thus, using Eq. (18), CD can be calculated as $2.850 \times \sqrt{\dfrac{6 \times 7}{6 \times 20}} = 1.6861$. For the metric Precision, the difference in the mean rankings between OAHD and CART was 2.1. Since the difference was greater than CD, the performance of OAHD was significantly better than that of the CART decision tree. Accordingly, we can calculate that the differences in the mean rankings between OAHD and C4.5, DCSM, iHD, and iHDw were 2.4, 2.15, 1.75, and 1.95, respectively. Because all of the differences were greater than CD, the performance of the OAHD decision tree was significantly better than those of the compared algorithms. For the metrics F-measure and MAUC, the differences in the mean rankings between OAHD and the five compared decision trees can be obtained using Eq. (18). The differences in the mean rankings between OAHD and the five unpruned decision trees were 1.95, 2.55, 2.15, 1.5, and 1.75, respectively, for the metric F-measure. Likewise, for the metric MAUC, the differences in the mean rankings between OAHD and the five unpruned decision trees were 1.75, 3.3, 2.7, 1.9, and 1.75, respectively. Evidently, the differences in the mean rankings were greater than CD, which means that OAHD was better than the other algorithms at the 0.05 significance level. The only exception was the metric Precision for iHD, where the difference in the mean rankings was less than CD.

From the above analysis of results, we can find that OAHD is significantly better than the compared algorithms, because the one-against-all decomposition scheme has been introduced to the process of computing the splitting criterion in OAHD, which is skew-insensitive. Furthermore, to guarantee the purity of nodes in the decision tree, OAHD accounts for the number of distinct classes and considers the class distribution of the multiclass imbalance problem; meanwhile, OAHD has a modified Gini index that fits the multiclass imbalance problem. Therefore, the performance of OAHD is greatly improved while processing the multiclass imbalance problem.

5 Conclusions

The major goal of this work is to construct a decision tree built upon the one-against-all-based Hellinger distance (OAHD) for addressing the multiclass imbalanced classification problem. Initially,

to enhance the performance of the decision tree in dealing with a multiclass imbalance problem, we design a new splitting criterion (i.e., OAHD) which is associated with the idea of the one-against-all scheme by extending the Hellinger distance to deal with the issue of multiclass imbalance. In OAHD, while considering the multiclass imbalance problem, the number of distinct classes and the class distribution are considered without a prior probability of the classes. Meanwhile, we modify the Gini index to fit the multiclass imbalance problem, which ensures the purity of the node in the decision tree. Furthermore, we theoretically prove that the proposed splitting criterion enables the decision tree with the property of skew-insensitivity and the ability to seek a purer node. Finally, OAHD is compared with five different unpruned decision trees upon 20 data sets. The experimental results show that the proposed splitting criterion is better than the five other splitting criteria. Moreover, the Friedman and Nemenyi tests are used to evaluate the performances of the six decision trees; the results demonstrate that this improvement is statistically significant.

In our future work, to improve the performance of OAHD, we intend to explore the effect of the pruning method on OAHD. Furthermore, we intend to make an extension of OAHD that is helpful for tree-based ensemble classifiers.

Contributors

Minggang DONG guided the research. Ming LIU designed the research and drafted the paper. Chao JING helped organize the paper. Minggang DONG and Chao JING revised and finalized the paper.

Compliance with ethics guidelines

Minggang DONG, Ming LIU, and Chao JING declare that they have no conflict of interest.

References

Abdi L, Hashemi S, 2016. To combat multi-class imbalanced problems by means of over-sampling techniques. *IEEE Trans Knowl Data Eng*, 28(1):238-251. https://doi.org/10.1109/TKDE.2015.2458858

Akash PS, Kadir ME, Ali AA, et al., 2019. Inter-node Hellinger distance based decision tree. Proc 28th Int Joint Conf on Artificial Intelligence, p.1967-1973. https://doi.org/10.24963/ijcai.2019/272

Alcala-Fdez J, Fernandez A, Luengo J, et al., 2011. KEEL data-mining software tool: data set repository, integration of algorithms and experimental analysis framework. *J Multi-valued Logic Soft Comput*, 17(2-3):255-287.

Ali H, Salleh MNM, Saedudin R, et al., 2019. Imbalance class problems in data mining: a review. *Indones J Elect Eng Comput Sci*, 14(3):1560-1571. https://doi.org/10.11591/ijeecs.v14.i3.pp1552-1563

Anand R, Mehrotra K, Mohan CK, et al., 1995. Efficient classification for multiclass problems using modular neural networks. *IEEE Trans Neur Netw*, 6(1):117-124. https://doi.org/10.1109/72.363444

Asuncion A, 2007. UCI Machine Learning Repository. University of California, Irvine, USA. https://archive.ics.uci.edu/ml/index.php

Boonchuay K, Sinapiromsaran K, Lursinsap C, 2017. Decision tree induction based on minority entropy for the class imbalance problem. *Patt Anal Appl*, 20(3):769-782. https://doi.org/10.1007/s10044-016-0533-3

Bradley AP, 1997. The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Patt Recogn*, 30(7):1145-1159. https://doi.org/10.1016/S0031-3203(96)00142-2

Breiman L, Friedman JH, Olshen RA, et al., 1984. Classification and regression trees. *Biometrics*, 40(3):874. https://doi.org/10.2307/2530946

Chandra B, Kothari R, Paul P, 2010. A new node splitting measure for decision tree construction. *Patt Recogn*, 43(8):2725-2731. https://doi.org/10.1016/j.patcog.2010.02.025

Cichocki A, Amari SI, 2010. Families of Alpha- Beta- and Gamma-divergences: flexible and robust measures of similarities. *Entropy*, 12(6):1532-1568. https://doi.org/10.3390/e12061532

Cieslak DA, Chawla NV, 2008. Learning decision trees for unbalanced data. In: Daelemans W, Goethals B, Morik K (Eds.), Machine Learning and Knowledge Discovery in Databases. Springer, Berlin, Germany, p.241-256. https://doi.org/10.1007/978-3-540-87479-9_34

Cieslak DA, Hoens TR, Chawla NV, et al., 2012. Hellinger distance decision trees are robust and skew-insensitive. *Data Min Knowl Discov*, 24(1):136-158. https://doi.org/10.1007/s10618-011-0222-1

Feng L, Wang HB, Jin B, et al., 2019. Learning a distance metric by balancing KL-divergence for imbalanced datasets. *IEEE Trans Syst Man Cybern Syst*, 49(12):2384-2395. https://doi.org/10.1109/TSMC.2018.2790914

Flach PA, 2003. The geometry of ROC space: understanding machine learning metrics through ROC isometrics. Proc 20th Int Conf on Machine Learning, p.194-201.

Friedman M, 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *J Am Stat Assoc*, 32(200):675-701. https://doi.org/10.1080/01621459.1937.10503522

Friedman M, 1940. A comparison of alternative tests of significance for the problem of m rankings. *Ann Math Stat*, 11(1):86-92. https://doi.org/10.1214/aoms/1177731944

Hanley JA, McNeil BJ, 1982. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29-36. https://doi.org/10.1148/radiology.143.1.7063747

He HB, Garcia EA, 2009. Learning from imbalanced data. *IEEE Trans Knowl Data Eng*, 21(9):1263-1284. https://doi.org/10.1109/TKDE.2008.239

Iman RL, Davenport JM, 1980. Approximations of the critical region of the fbietkan statistic. *Commun Stat Theory Methods*, 9(6):571-595. https://doi.org/10.1080/03610928008827904

Kailath T, 1967. The divergence and Bhattacharyya distance measures in signal selection. *IEEE Trans Commun Technol*, 15(1):52-60. https://doi.org/10.1109/TCOM.1967.1089532

Kotsiantis SB, 2013. Decision trees: a recent overview. *Artif Intell Rev*, 39(4):261-283. https://doi.org/10.1007/s10462-011-9272-4

Liu W, Chawla S, Cieslak DA, et al., 2010. A robust decision tree algorithm for imbalanced data sets. Proc SIAM Int Conf on Data Mining, p.766-777. https://doi.org/10.1137/1.9781611972801.67

Nekooeimehr I, Lai-Yuen SK, 2016. Adaptive semi-unsupervised weighted oversampling (A-SUWO) for imbalanced datasets. *Expert Syst Appl*, 46:405-416. https://doi.org/10.1016/j.eswa.2015.10.031

Nemenyi P, 1963. Distribution-Free Multiple Comparisons. MS Thesis, Princeton University, Princeton, USA.

Osei-Bryson KM, 2014. Overview on decision tree induction. In: Osei-Bryson KM, Ngwenyama O (Eds.), Advances in Research Methods for Information Systems Research. Springer, Boston, USA, p.15-22. https://doi.org/10.1007/978-1-4614-9463-8_3

Quinlan JR, 1986. Induction of decision trees. *Mach Learn*, 1(1):81-106. https://doi.org/10.1007/BF00116251

Safavian SR, Landgrebe D, 1991. A survey of decision tree classifier methodology. *IEEE Trans Syst Man Cybern*, 21(3):660-674. https://doi.org/10.1109/21.97458

Sharmin S, Shoyaib M, Ali AA, et al., 2019. Simultaneous feature selection and discretization based on mutual information. *Patt Recogn*, 91:162-174. https://doi.org/10.1016/j.patcog.2019.02.016

Su C, Cao J, 2019. Improving lazy decision tree for imbalanced classification by using skew-insensitive criteria. *Appl Intell*, 49(3):1127-1145. https://doi.org/10.1007/s10489-018-1314-z

Vilalta R, Oblinger D, 2000. A quantification of distance bias between evaluation metrics in classification. Proc 17th Int Conf on Machine Learning, p.1087-1094.

Wan ZQ, Jiang C, Fahad M, et al., 2020. Robot-assisted pedestrian regulation based on deep reinforcement learning. *IEEE Trans Cybern*, 50(4):1669-1682. https://doi.org/10.1109/TCYB.2018.2878977

Wu XD, Kumar V, Quinlan JR, et al., 2008. Top 10 algorithms in data mining. *Knowl Inform Syst*, 14(1):1-37. https://doi.org/10.1007/s10115-007-0114-2