Inter-node Hellinger Distance based Decision Tree

Pritom Saha Akash[1], **Md. Eusha Kadir**[1], **Amin Ahsan Ali**[2] and **Mohammad Shoyaib**[1]
[1]Institute of Information Technology, University of Dhaka, Bangladesh
[2]Department of Computer Science & Engineering, Independent University, Bangladesh
{bsse0604 & bsse0708}@iit.du.ac.bd, aminali@iub.edu.bd, shoyaib@du.ac.bd

Abstract

This paper introduces a new splitting criterion called Inter-node Hellinger Distance (iHD) and a weighted version of it (iHDw) for constructing decision trees. iHD measures the distance between the parent and each of the child nodes in a split using Hellinger distance. We prove that this ensures the mutual exclusiveness between the child nodes. The weight term in iHDw is concerned with the purity of individual child node considering the class imbalance problem. The combination of the distance and weight term in iHDw thus favors a partition where child nodes are purer and mutually exclusive, and skew insensitive. We perform an experiment over twenty balanced and twenty imbalanced datasets. The results show that decision trees based on iHD win against six other state-of-the-art methods on at least 14 balanced and 10 imbalanced datasets. We also observe that adding the weight to iHD improves the performance of decision trees on imbalanced datasets. Moreover, according to the result of the Friedman test, this improvement is statistically significant compared to other methods.

1 Introduction

Three of the major tasks in machine learning are Feature Extraction, Feature Selection, and Classification [Iqbal *et al.*, 2017; Sharmin *et al.*, 2019; Kotsiantis *et al.*, 2007]. In this study, we only focus on the classification task. One of the simplest and easily interpretable classification methods (also known as classifiers) is decision tree (DT) [Quinlan, 1986]. It is a tree-like representation of possible outcomes to a problem. Learning of a DT is a greedy approach. Nodes in a DT can be categorized into two types: decision nodes and leaf nodes. At each decision node, a locally best feature is selected to split the data into child nodes. This process is repeated until a leaf node is reached where the further splitting is not possible. The best feature is selected based on a splitting criterion which measures the goodness of a split. One of the most popular splitting criteria is Information Gain (IG) [Quinlan, 1986; Breiman *et al.*, 1984] which is an impurity based splitting criterion (i.e., entropy and gini). DTs based on IG perform quite well for balanced datasets where the class distribution is uniform. However, as class prior probability is used to calculate the impurity of a node, in an imbalanced dataset, IG becomes biased towards the majority class which is also called skew sensitivity [Drummond and Holte, 2000].

To improve the performance of standard DTs, several splitting criteria are proposed to construct DTs in Distinct Class based Splitting Measure (DCSM) [Chandra *et al.*, 2010], Hellinger Distance Decision Tree (HDDT) [Cieslak and Chawla, 2008] and Class Confidence Proportion Decision Tree (CCPDT) [Liu *et al.*, 2010]. Besides these, to deal with class imbalance problem in Lazy DT construction, two skew insensitive split criteria based on Hellinger distance and K-L divergence are proposed in [Su and Cao, 2019]. Since Lazy DTs use the test instance to make splitting decisions, in this paper, we omit it from our discussion.

In DCSM, the number of distinct classes in a partition is incorporated. Trees generated with DCSM is smaller in size, however, DCSM is still skew sensitive because of its use of class prior probability. HDDT and CCPDT propose new splitting criteria to address the class imbalance problem. However, the Hellinger distance based criterion proposed in HDDT can perform poorly when training samples are more balanced [Cieslak and Chawla, 2008]. At the same time, HDDT fails more often to differentiate between two different splits (specifically for multiclass problems) which is illustrated in the following example:

Assume, there are 80 samples, 40 of class A, 20 of class B, 10 of class C and rest of class D. Two splits (split X and Y) are compared where each split has 50 observations on the left child and the rest on the right child. Split X channels all the samples of class A and class D into the left child, and the rest to the right child while Split Y places all the samples of class A, 5 each of classes C and D into the left child, and the rest to the right child. It is easily observable that, Split X is more exclusive than Split Y. However, HDDT cannot differentiate between these two splits and provides the same measure.

On the other hand, instead of using class probability, CCPDT calculates the splitting criteria like entropy and gini using a new measure called Class Confidence Proportion (CCP) which is skew insensitive. However, it uses HDDT to break ties while splits based on two different features provide the same split measure. Hence, CCPDT exhibits the same limitation as HDDT.

To address the limitations of the above methods, we propose a new splitting criterion called Inter-node Hellinger Distance (iHD) which is skew insensitive. We then propose iHDw which adds a weight to iHD to make sure child nodes are purer without forgoing skew insensitivity. Both iHD and iHDw exhibit exclusivity preference property defined in [Taylor and Silverman, 1993]. Rigorous experiments over a large number of datasets and statistical tests are performed to show the superiority of iHD and iHDw for the construction of DTs.

The rest of the paper is organized in the following manner. In Section 2, several related node splitting criteria for DTs are discussed. The new node splitting criteria are presented in Section 3. In Section 4, datasets, performance measures, and experimental results are described. Finally, Section 5 concludes the paper.

2 Related Work

In this section, we discuss several split criteria for DT related to our proposed measure.

2.1 Information Gain

Information Gain (IG) calculates how much "information" a feature gives about the class, and measures the decrease of impurity in a collection of examples after splitting into child nodes. IG for splitting a data, (X, y) with attribute A and threshold, T is calculated using (1).

$$Gain(A, T : X, y) = Imp(y|X) - \sum_{i=1}^{V} \frac{|X_i|}{|X|} Imp(y|X_i) \quad (1)$$

where V is the number of partitions and Imp is the impurity measure. Widely used impurity metrics are Entropy [Quinlan, 1986] and Gini Index [Breiman *et al.*, 1984] and calculated using (2) and (3) respectively.

$$Entropy(y|X) = -\sum_{j=1}^{k} p(y_j|X) \log p(y_j|X) \quad (2)$$

$$Gini(y|X) = 1 - \sum_{j=1}^{k} p(y_j|X)^2 \quad (3)$$

where k is the number of classes. When data are balanced, IG gives a reasonably good splitting boundary. However, when there is an imbalanced distribution of classes in a dataset, IG becomes biased towards the majority classes [Drummond and Holte, 2000]. Another drawback of IG is that it favors attributes with a large number of distinct values. To reduce this bias, a new criterion called Gain Ratio (GR) [Quinlan, 1993] was proposed by taking account of the size of a split while choosing an attribute. GR defines the size of a split, g as (4).

$$g = -\sum_{i=1}^{V} \frac{|X_i|}{|X|} \log \frac{|X_i|}{|X|} \quad (4)$$

GR is just the ratio between IG and g, defined in (5).

$$GainRatio(A, T : X, y) = \frac{Gain(A, T : X, y)}{g} \quad (5)$$

2.2 DCSM

Chandra et. al. proposed a new splitting criterion called Distinct Class based Split Measure (DCSM) that emphasizes the number of distinct classes in a partition [Chandra *et al.*, 2010]. For a given attribute x_j and V number of partitions, the measure $M(x_j)$ is defined as (6).

$$M(x_j) = \sum_{v=1}^{V} \left[\frac{N^{(v)}}{N^{(u)}} * D(v) \exp(D(v)) * \sum_{k=1}^{C} [a_{\omega_k}^{(v)} * \exp(\delta^{(v)}(1 - (a_{\omega_k}^{(v)})^2))] \right] \quad (6)$$

where C is the number of distinct classes in the dataset, u is the splitting node (parent) representing x_j and v represents a partition (child node). $D(v)$ denotes the number of distinct classes in a partition v, $\delta^{(v)}$ is the ratio of the number of distinct classes in the partition v to that of u, i.e., $\frac{D(v)}{D(u)}$ and $a_{\omega_k}^{(v)}$ is the probability of class ω_k in the partition v.

The first term $D(v) * \exp(D(v))$ deals with the number of distinct classes in a partition. It increases when the number of distinct classes in a partition increases causing purer partitions to be preferred. The second term is $a_{\omega_k}^{(v)} * \exp(\delta^{(v)}(1 - (a_{\omega_k}^{(v)})^2))$. Here, impurity decreases when $\delta^{(v)}$ decreases. On the other hand, $(1 - (a_{\omega_k}^{(v)})^2))$ decreases when there are more examples of a class compared to the total number of examples in a partition. Hence, the DCSM is intended to reduce the impurity of each partition when it is minimized.

The main difference between DCSM and other splitting criteria is that DCSM introduces the concept of distinct classes. The limitation of DCSM is same as IG. It cannot deal with imbalanced class distribution and thus, is biased towards the majority classes.

2.3 HDDT

Hellinger Distance Decision Trees (HDDT) uses Hellinger distance as the splitting criterion to solve the problem of class imbalance [Cieslak and Chawla, 2008; Cieslak *et al.*, 2012]. The details of Hellinger distance is presented in Section 3. In HDDT, Hellinger distance (d_H) is used as a split criterion to construct a DT. Assume a two-class problem (class $+$ and class $-$) and, X_+ and X_- are the set of classes $+$ and $-$ respectively. Then, d_H between the distributions, X_+ and X_- is calculated as (7).

$$D_H(X_+||X_-) = \sqrt{\sum_{j=1}^{p} \left(\sqrt{\frac{|X_{+j}|}{|X_+|}} - \sqrt{\frac{|X_{-j}|}{|X_-|}} \right)^2} \quad (7)$$

Here, instead of using class probability, normalized frequencies aggregated over all the p partitions across classes are used. HDDT is strongly considered to be skew insensitive because of not using prior probability in the distance calculation. However, the split criterion defined in (7) only works on the binary classification problem. For the multi-class classification problem, a technique named Multi-Class HDDT [Hoens *et al.*, 2012] is proposed. They decompose the multi-class problem into multiple binary class problems

by using the similar of One-Versus-All (OVA) decomposition technique. For each binary class problem, they calculate D_H and the maximum is taken as the split measure for an attribute. However, as HDDT tries to make pure leaves by capturing deviation between class conditionals which results in smaller coverage, thus can perform poorly for more balanced class distribution [Cieslak and Chawla, 2008].

2.4 CCPDT

Another decision tree algorithm named Class Confidence Proportion Decision Tree (CCPDT) is proposed in [Liu *et al.*, 2010] where they introduce a new measure named Class Confidence Proportion (CCP) calculated as (8).

$$CCP(X \to y) = \frac{CC(X \to y)}{CC(X \to y) + CC(X \to \neg y)} \tag{8}$$

where Class Confidence (CC) is defined in (9).

$$CC(X \to y) = \frac{Support(X \cup y)}{Support(y)} \tag{9}$$

CCP is insensitive to the skewness of class distribution because of not focusing on class priors. In CCPDT, CCP is used to replace $p(y|X)$ in Entropy/Gini to calculate IG. Whenever there is a tie between two attributes in IG value, the Hellinger distance (same as in HDDT) is used to break the tie. For which, CCPDT has the same limitation as HDDT of performing poorly for more balanced datasets.

Different from the above approaches, we propose new splitting criteria for DTs which provide better results for both balanced and imbalanced datasets.

3 Proposed Method

In this section, we propose two new splitting criteria named Inter-node Hellinger Distance (iHD) and wighted iHD (iHDw) for constructing DT classifiers.

3.1 Inter-node Hellinger Distance

We use squared Hellinger distance (D_H^2) to measure the dissimilarity between the class probability distributions of the parent and each of the child nodes in a split. The distance is intended to be maximized so that the instances in the parent node are divided into mutually exclusive regions. Note that, the Hellinger distance is a divergence measure which is a member of α divergence family [Cichocki and Amari, 2010]. For two discrete probability mass functions, $\mathbf{P} = (p_1, p_2, ..., p_k)$ and $\mathbf{Q} = (q_1, q_2, ..., q_k)$, the α divergence is defined in (10).

$$D_\alpha(\mathbf{P}||\mathbf{Q}) = \frac{1}{\alpha(1-\alpha)} \sum_{j=1}^{k} \Big(\alpha p_j + (1-\alpha) q_j - p_j^\alpha q_j^{1-\alpha} \Big) \tag{10}$$

where $\alpha \notin \{0, 1\}$. For $\alpha = \frac{1}{2}$, we obtain D_H from (11).

$$\begin{aligned} D_{\frac{1}{2}}(\mathbf{P}||\mathbf{Q}) = 4D_H^2(\mathbf{P}||\mathbf{Q}) &= 2\sum_{j=1}^{k} \Big(\sqrt{p_j} - \sqrt{q_j} \Big)^2 \\ &= 4\Big(1 - \sum_{j=1}^{k} \sqrt{p_j q_j} \Big) \end{aligned} \tag{11}$$

Hellinger distance has the following basic properties :

- D_H is symmetric ($D_H(\mathbf{P}||\mathbf{Q}) = D_H(\mathbf{Q}||\mathbf{P})$) and non-negative.
- D_H is in [0,1]. It takes its maximum value when $\sum_{j=1}^{k} p_j q_j = 0$ and minimum value when $p_j = q_j, \forall j$.
- $D_H(\mathbf{P}||\mathbf{Q})$ is convex with respect to both P and Q.
- D_H satisfies the triangle inequality, $D_H(\mathbf{P}||\mathbf{Z}) \leq D_H(\mathbf{P}||\mathbf{Q}) + D_H(\mathbf{Q}||\mathbf{Z})$.

Suppose, there are N number of samples in a node distributed over k classes. For a binary split, N samples are divided into left (L) and right (R) child nodes which are N_L and N_R respectively. The class probability distribution for the parent node is $\mathbf{P}(p_1, ..., p_k)$ and, for the left and right child nodes are $\mathbf{P}_L(p_{L1}, ..., p_{Lk})$ and $\mathbf{P}_R(p_{R1}, ..., p_{Rk})$ respectively. D_H^2 between the class probability distribution of the parent and left child $D_H^2(\mathbf{P}_L||\mathbf{P})$ and, the parent and right child $D_H^2(\mathbf{P}_R||\mathbf{P})$ are calculated using (12).

$$D_H^2(\mathbf{P}_t||\mathbf{P}) = 1 - \sum_{j=1}^{k} \sqrt{p_{tj} p_j} \quad \forall t \in \{L, R\} \tag{12}$$

where k is the number of classes. From these distances, the proposed splitting criterion Inter-node Hellinger Distance (iHD) is defined as (13).

$$iHD = \rho_L D_H^2(\mathbf{P}_L||\mathbf{P}) + \rho_R D_H^2(\mathbf{P}_R||\mathbf{P}) \tag{13}$$

where $\rho_L = \frac{N_L}{N}$ and $\rho_R = \frac{N_R}{N}$.

Note that, the difference between the use of Hellinger distance in iHD and HDDT is that, in iHD, the distance between the class probability distributions of the parent and child nodes are measured rather than the distance between the class pairs over all partitions. As a consequence of the triangle inequality of D_H, maximizing the distance between the class probability distributions of the parent and the child nodes also maximizes the distance between the distributions of the two child nodes.

Properties of iHD

iHD has the exclusivity preference property (proved in Theorem 1) which is expected for a good splitting criterion [Taylor and Silverman, 1993; Shih, 1999]. This property is defined by the following two conditions:

1. Firstly, for a certain value of $\rho_L \rho_R$, the criterion has the maximum value when $\sum_{j=1}^{k} p_{Lj} p_{Rj} = 0$ which indicates that the two child nodes are mutually exclusive.
2. Secondly, regardless of $\rho_L \rho_R$, it obtains its minimum value when the class probability distributions of child nodes are identical which can be defined as $p_{Lj} = p_{Rj} = p_j, \forall j$.

Theorem 1. *iHD has the exclusivity preference property.*

Proof. From (13), iHD can be written as:

$$\begin{aligned} iHD &= \rho_L \Big(1 - \sum_{j=1}^{k} \sqrt{p_{Lj} p_j} \Big) + \rho_R \Big(1 - \sum_{j=1}^{k} \sqrt{p_{Rj} p_j} \Big) \\ &= 1 - \sum_{j=1}^{k} \sqrt{p_j} (\rho_L \sqrt{p_{Lj}} + \rho_R \sqrt{p_{Rj}}) \end{aligned} \tag{14}$$

For a certain value of $\rho_L\rho_R$, (14) is maximum when the value from the summation over all classes is minimum. In other words, (14) is maximum when we get the minimum value separately for each class in the summation. Let, for j^{th} class, $a_j = p_{Lj}$ and $b_j = p_{Rj}$ for a fixed $\rho_L = 1 - \rho_R \neq 0$. Thus, $p_j = \rho_L a_j + \rho_R b_j$ which is a nonzero constant. Now, the j^{th} term in the summation can be expressed by a function of a_j as follows:

$$f(a_j) = \sqrt{p_j}(\rho_L\sqrt{a_j} + \sqrt{\rho_R}\sqrt{p_j - \rho_L a_j})$$

The second derivative of $f(a_j)$ is:

$$f^{''}(a_j) = -\frac{\rho_L\sqrt{p_j}}{4a_j^{\frac{3}{2}}} - \frac{a_j^2\sqrt{\rho_R}\sqrt{p_j}}{4(p_j - \rho_L a_j)^{\frac{3}{2}}}$$

Here, $f^{''}(a_j) < 0$ in the interval of $0 \leq a_j \leq \frac{p_j}{\rho_L}$, thus is a concave function. Hence, $f(a_j)$ has the minimum value at one of the extreme points of the interval which is either $a_j = 0$ or $a_j = \frac{p_j}{\rho_L}$ (equivalent to $b_j = 0$). Now, for regardless of $\rho_L\rho_R$, when $p_{Lj} = p_{Rj} = p_j, \forall j$, (14) becomes:

$$iHD = 1 - \sum_{j=1}^{k} p_j(\rho_L + \rho_R) = 0$$

Therefore, iHD is minimum when $p_{Lj} = p_{Rj} = p_j, \forall j$. □

From the exclusivity preference property of iHD, we can say that it is minimum when $D_H^2(\mathbf{P}_L||\mathbf{P}) = D_H^2(\mathbf{P}_R||\mathbf{P}) = 0$ which means the parent and child nodes have the same probability distribution. And iHD gets its maximum when the samples at the parent node are distributed among the child nodes as disjoint subsets of classes.

Moreover, iHD is skew insensitive in the sense that maximizing iHD in a split focuses on generating mutually exclusive child nodes whatever class distribution there is in the parent node.

3.2 Weighted Inter-node Hellinger Distance

To obtain purer child nodes in a split we also consider a weight, w for each partitioned node which is defined as (15).

$$w_t = 1 - \prod_{j=1}^{k} \frac{N_{tj}}{N_j} = 1 - \prod_{j=1}^{k} \frac{\rho_t p_{tj}}{p_j} \quad \forall t \in \{L, R\} \quad (15)$$

where N_j and N_{t_j} are the number of samples in a class j at the parent and the child node t respectively. Here, instead of using class probability, the proportion of instances of each class from the parent node placed in the child node t is used to calculate w_t. For which, w_t is not dependent on the prior probability of a class, thus, is not biased towards the majority classes. It is easy to say from (15) that for any value of ρ_t, w_t will give maximum value of 1 when for any class j, $p_{tj} = 0$. And, when the difference between the proportion of samples of classes of the parent node ($\frac{N_{tj}}{N_j}$) increases in a child node t, w_t increases, thus favors a purer partition. On the other hand, w_t gives the minimum value of 0 when all the samples of a parent node come to a single child node t.

The distance measure and the weight (defined in (12) and (15) respectively) for each partition are combined to formulate the final proposed splitting criterion named weighted Inter-node Hellinger Distance (iHDw) as (16).

$$iHDw = \rho_L D_H^2(\mathbf{P}_L||\mathbf{P})w_L + \rho_R D_H^2(\mathbf{P}_R||\mathbf{P})w_R \quad (16)$$

The weighted sum by the proportion of samples ($\rho_t, i \in \{L, R\}$) is taken to evaluate the contribution from each partition and to favor partitions with similar sizes. As the Hellinger distance D_H and the weight w are both non-negative, the proposed splitting criterion is also non-negative.

Properties of iHDw

Now, we prove that the splitting criterion iHDw also preserves the exclusivity preference property.

Theorem 2. *iHDw has the exclusivity preference property.*

Proof. As Theorem 1 states that iHD has the exclusivity preference property, it is enough to show that, after incorporating the weights to iHD, iHDw also provides the maximum value when the child nodes are mutually exclusive and minimum value when the parent and child nodes have identical class probability distributions.

Here, w_t from (15) gives maximum value of 1 when for any class j, $p_Lj = 0$. Hence, for $\sum_{j=1}^{k} p_{Lj}p_{Rj} = 0$, it is straightforward to say that $w_L = w_R = 1$, thus fulfills the first condition of the exclusivity preference property.

Regardless of the value of $w_L * w_R$, iHDw gets its minimum value of 0 as $D_H^2(\mathbf{P}_L||\mathbf{P}) = D_H^2(\mathbf{P}_R||\mathbf{P}) = 0$ for $p_{Lj} = p_{Rj} = p_j, \forall j$ (second property of D_H) which fulfills the second condition of the exclusivity preference property. □

Therefore, the combination of the two terms prefers a split with child nodes purer and mutually exclusive. We also show the behavior of optimal splits using iHDw in Theorem 3.

Theorem 3. *iHDw on its optimal split channels the classes to two disjoint subsets s and $s^c \subset \{1, 2, ..., k\}$ where s minimizes $\mid \rho_L - \frac{1}{2} \mid$.*

Proof. From (16) we rewrite iHDw as

$$\begin{aligned} iHDw &= \rho_L(1 - \sum_{j\in s} \frac{p_j}{\sqrt{\rho_L}}) + (1 - \rho_L)(1 - \sum_{j\in s^c} \frac{p_j}{\sqrt{1-\rho_L}}) \\ &= \rho_L(1 - \sqrt{\rho_L}) + (1 - \rho_L)(1 - \sqrt{1 - \rho_L}) \quad (17) \end{aligned}$$

where $w_L = w_R = 1$. Let denote (17) as a function of $\boldsymbol{\rho}_L$:

$$f(\boldsymbol{\rho}_L) = 1 - \rho_L^{\frac{3}{2}} - (1 - \rho_L)^{\frac{3}{2}}$$

Second derivative of $f(\boldsymbol{\rho}_L)$ is:

$$f''(\boldsymbol{\rho}_L) = -\frac{3}{4}\rho_L^{-\frac{1}{2}} - \frac{3}{4}(1 - \rho_L)^{-\frac{1}{2}}$$

Hence, $f(\boldsymbol{\rho}_L)$ is a concave function and is symmetric w.r.t $p_L = \frac{1}{2}$, thus $f(\boldsymbol{\rho}_L)$ takes its maximum at $\rho_L = \frac{1}{2}$. □

Algorithm 1 outlines the procedure of learning a binary DT using the proposed split criterion iHDw.

Algorithm 1 Grow binary tree

Input: Training set $\mathcal{D}$

if stopping criteria meet **then**
  Stop growing and create a $leaf$.
  return $leaf$
else
  Create a $node$
  Find a feature f^* and split point s^* maximizing (16)
  Split $\mathcal{D}$ into $\mathcal{D}_\mathcal{L}$ ($f^* < s^*$) and $\mathcal{D}_\mathcal{R}$ ($f^* \geq s^*$)
  node.left ← Grow binary tree (with $\mathcal{D}_\mathcal{L}$)
  node.right ← Grow binary tree (with $\mathcal{D}_\mathcal{R}$)
  return $node$
end if

f_1	26	44	34	42	32	24	40	36	22	28	38	30
f_2	12	20	16	22	28	24	26	32	30	18	34	14
Class	B	B	B	B	A	B	B	B	A	A	A	B

Table 1: Sample dataset for the example.

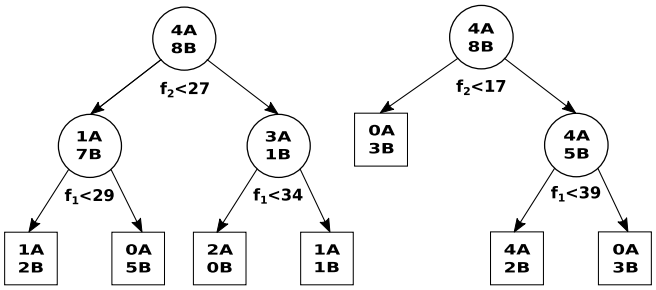

Figure 1: Tree based on iHD

Figure 2: Tree based on iHDw

3.3 An Illustrative Example

Let consider a sample dataset with two classes (Class "A" and "B") having two features (f_1 and f_2) shown in Table 1. The number of samples for class "A" and "B" are 4 and 8 respectively, thus having imbalance class ratio. In Figures 1 and 2, two DTs constructed using iHD and iHDw are shown respectively. We observe that at the first split iHD and iHDw choose two different split points using feature f_2 and unlike iHD, iHDw produces a pure child. In the final tree for iHDw, we find that none of the samples from the minority class is misclassified. On the other hand, although the final tree for iHD produces two pure leaf nodes, the minority class can be misclassified in the other leaf nodes. Moreover, the DT based on iHDw has fewer impure leaf nodes than that of iHD.

4 Experiments and Results

In this section, we first describe the datasets used in the experiments followed by the description of the performance measures. We then present the results obtained from the experiments with proper discussion.

4.1 Description of Datasets

Table 2 shows 40 datasets chosen from various areas like biology, medicine and finance. Datasets are grouped into two sets: Balanced datasets and Imbalanced datasets. These datasets are collected from two well-known public sources

Balanced Datasets				Imbalanced Datasets				
Datasets	#Insts	#Ftrs	#Cls	Datasets	#Insts	#Ftrs	#Cls	IR
appendicitis	106	7	2	balance	625	4	3	5.88
australian	690	14	2	dermatology	366	34	3	5.55
breast	569	32	2	ecoli-0-1_vs_2-3-5	244	7	2	9.17
cardio	2126	21	10	ecoli-0-1-4-6_vs_5	280	6	2	13
digits	5620	64	10	ecoli-0-1-4-7_vs_2-3-5-6	336	7	2	10.59
german	1000	20	2	ecoli-0-6-7_vs_5	220	6	2	10
liver	351	33	2	ecoli2	336	7	2	5.46
lung	203	3312	5	haberman	306	3	2	2.78
lymphography	148	18	4	hayes-roth	132	4	3	1.7
musk	345	7	2	new-thyroid	215	5	3	4.84
segment	476	166	2	new-thyroid1	215	5	2	5.14
semeion	2310	19	7	pageblocks	548	10	5	164
sonar	1593	256	10	paw02a-600-5-70-BI	600	2	2	5
spambase	208	60	2	penbased	1100	16	10	1.95
steel	4601	57	2	vehicle3	846	18	2	2.99
transfusion	748	4	2	winequality-red-4	1599	11	2	29.17
vowel	528	10	11	wisconsin	683	9	2	1.86
waveform	5000	21	3	yeast-0-2-5-6_vs_3-7-8-9	1004	8	2	9.14
wine	178	13	3	yeast-0-3-5-9_vs_7-8	506	8	2	9.12
yeast	1484	8	10	yeast-2_vs_4	514	8	2	9.08

Table 2: Summary of the datasets. #Insts, #Ftrs and #Cls denote the number of instances, features and classes respectively.

called UCI Machine Learning Repository [Dua and Graff, 2017] and KEEL Imbalanced Data Sets [Alcalá-Fdez *et al.*, 2011]. Imbalance Ratio (IR) between the samples of majority and minority classes is also shown in Table 2. The higher value of IR indicates the dataset is highly imbalanced.

4.2 Performances Measures

For each dataset, we build eight unpruned DT classifiers based on iHD, iHDw, information gain (using both Entropy and Gini), Gain Ratio (GR) and, the splitting criteria proposed in DCSM, HDDT and CCPDT respectively. For balanced datasets accuracy (%) is used as the performance measure. Since accuracy as an evaluation measure is inappropriate for class imbalance problem [Chawla *et al.*, 2004], we use the area under the ROC curve (AUC) [Hand and Till, 2001] for imbalanced datasets. We conduct 10-fold cross-validation on each dataset to get the unbiased result.

The method proposed in [Demšar, 2006] is used to compare the performance of DT classifiers based on iHD and iHDw with other six methods. For comparing different classifiers over multiple datasets, Demsar proposed the use of Iman's F statistic [Iman and Davenport, 1980] using Friedman's χ^2_F statistic [Friedman, 1937; Friedman, 1940]. χ^2_F statistic is calculated as follows:

$$\chi^2_F = \frac{12N}{k(k+1)}\Big[\sum_i R_i^2 - \frac{k(k+1)^2}{4}\Big] \tag{18}$$

Iman's F statistic is then calculated from χ^2_F as (19).

$$F_F = \frac{(N-1)\chi^2_F}{N(k-1)-\chi^2_F} \tag{19}$$

where k is the number of compared classifiers and R_i is the average rank of i^{th} classifier on N datasets. After rejecting the null hypothesis that all the classifiers are equivalent, a post-hoc test called Nemenyi test [Nemenyi, 1963] is used to determine the performance of which classifier is significantly better than the others. Based on the Nemenyi test, the performance of a classifier can be said significantly different than others if the difference between their corresponding average ranks is larger than a critical difference (CD) which is

Datasets	Entropy	Gini	GR	DCSM	HDDT	CCPDT	iHD	iHDw
appendicitis	79.17	78.33	79.17	**80.83**	80.83	76.67	79.17	**80.83**
australian	81.86	82.57	81.00	82.14	80.29	82.57	**82.86**	82.71
breast	92.24	92.24	93.10	**93.97**	93.31	93.10	**93.97**	93.62
cardio	81.79	81.24	80.73	81.51	80.78	81.97	**82.89**	82.71
digits	87.24	85.08	85.14	84.70	86.81	**88.76**	87.68	87.51
german	67.50	66.90	64.20	69.00	68.30	68.90	**69.90**	69.80
liver	62.00	58.57	62.00	62.00	63.14	**64.86**	61.43	64.00
lung	90.45	86.82	**91.36**	88.18	85.45	82.27	87.73	87.27
lymphography	75.00	69.44	77.78	81.11	**83.89**	76.67	76.11	77.78
musk	80.21	79.79	79.17	78.54	82.29	78.75	**85.63**	84.79
segment	96.62	96.93	95.84	96.88	96.14	96.93	**97.23**	96.97
semeion	75.55	76.34	**77.20**	69.15	70.24	73.90	**77.20**	76.95
sonar	78.64	77.27	72.27	76.36	77.27	75.91	76.82	**80.45**
spambase	92.02	92.04	91.67	92.30	92.73	92.02	**92.89**	92.69
steel	**74.85**	74.34	68.08	71.06	71.72	70.91	74.60	74.55
transfusion	69.33	69.60	71.07	70.93	71.47	**72.40**	70.93	71.47
vowel	81.92	80.91	79.60	80.71	81.82	**83.03**	82.83	83.03
waveform	76.27	75.33	72.08	74.49	74.79	75.31	**77.25**	76.77
wine	92.11	90.00	92.63	91.58	**96.32**	91.05	94.21	94.21
yeast	51.57	**52.61**	51.05	50.52	50.46	50.13	51.44	52.16
Avg. Rank	4.70	5.55	5.88	5.15	4.55	4.93	2.80	2.45
W/T/L (iHD)	14/1/5	18/0/2	14/2/4	14/2/4	14/0/6	15/0/5	—	—
W/T/L (iHDw)	18/0/2	19/0/1	17/1/2	16/1/3	15/2/3	16/1/3	7/1/12	—
Fr.T (iHD)		✓	✓	✓		✓	Base	
Fr.T (iHDw)	✓	✓	✓	✓	✓	✓		Base

Table 3: Accuracy (%) of the unpruned DTs based on different split criteria on balanced datasets.

Datasets	Entropy	Gini	GR	DCSM	HDDT	CCPDT	iHD	iHDw
balance	65.82	66.04	66.70	65.75	65.85	66.50	66.24	**66.80**
dermatology	96.17	94.82	96.17	95.83	95.48	95.00	**96.36**	**96.36**
ecoli-0-1_vs_2-3-5	78.03	77.35	**82.80**	80.61	77.95	77.50	79.70	80.61
ecoli-0-1-4-6_vs_5	70.77	75.77	**81.73**	74.04	71.35	76.35	76.15	76.15
ecoli-0-1-4-7_vs_2-3-5-6	84.41	82.90	85.59	87.42	87.58	82.74	86.56	**89.89**
ecoli-0-6-7_vs_5	84.00	82.00	84.00	83.25	**84.75**	74.00	82.25	**84.75**
ecoli2	80.43	79.41	83.85	79.40	**83.93**	80.57	80.04	80.61
haberman	56.79	52.22	52.51	56.18	52.73	54.47	51.86	**56.88**
hayes-roth	91.48	91.48	90.42	91.48	91.48	91.48	**92.08**	**92.08**
new-thyroid	88.97	90.83	**93.70**	90.74	90.61	91.06	92.05	92.05
new-thyroid1	95.69	94.58	93.61	96.11	95.11	**96.39**	**96.39**	**96.39**
pageblocks	87.97	88.30	81.13	79.99	82.04	**89.83**	89.04	89.52
paw02a-600-5-70-BI	68.40	68.20	64.30	70.40	70.80	67.80	69.60	**71.10**
penbased	**94.99**	93.35	94.97	93.41	93.96	93.76	93.75	93.66
vehicle3	68.06	68.86	69.45	**73.32**	69.15	70.30	70.91	71.19
winequality-red-4	53.87	53.65	53.23	52.84	54.61	52.91	**54.74**	54.64
wisconsin	92.40	92.08	92.89	**94.38**	92.72	93.72	92.99	92.76
yeast-0-2-5-6_vs_3-7-8-9	75.48	74.26	74.21	**77.36**	76.47	71.03	74.31	75.75
yeast-0-3-5-9_vs_7-8	**64.85**	61.76	63.87	59.74	64.41	58.74	63.85	**64.85**
yeast-2_vs_4	84.68	80.41	**86.03**	78.32	84.15	82.38	84.26	85.09
Avg. Rank	4.80	6.40	4.28	4.98	4.43	5.05	3.80	2.28
W/T/L (iHD)	13/0/7	19/0/1	10/0/10	12/0/8	12/0/8	12/1/7	—	—
W/T/L (iHDw)	18/1/1	20/0/0	13/0/7	16/1/3	16/1/3	15/1/4	12/5/3	—
Fr.T (iHD)		✓					Base	
Fr.T (iHDw)	✓	✓	✓	✓	✓	✓		Base

Table 4: AUC (%) of the unpruned DTs based on different split criteria on imbalanced datasets.

calculated as:

$$CD = q_\alpha \sqrt{\frac{k(k+1)}{6N}} \tag{20}$$

where q_α is the critical value for the two-tailed Nemenyi test for k classifiers at α significance level.

4.3 Experimental Results

Tables 3 and 4 represent the comparison of eight classifiers on balanced and imbalanced datasets respectively. The results of Friedman test (Fr.T) with two "Base" classifiers (iHD and iHDw) are shown in the last two lines of the Tables 3 and 4. The " ✓" sign under a classifier indicates that the "Base" classifier significantly outperforms that classifier at 95% confidence level.

The average ranks based on the accuracy of the compared classifiers indicate that the iHDw is the best performing criterion compared to others on the balanced datasets. Friedman's χ_F^2 statistic from these average ranks is 35.77 according to (18). From which, we get Iman's F statistic as 6.52 (using (19)). With eight classifiers and 20 balanced datasets, the critical value, $F(7, 133)$ is 2.31 at 95% confidence level, thus rejects the null hypothesis (as $F_F > F_{\alpha=0.05}$) that, all the eight classifiers are equivalent. After conducting the post-hoc Nemenyi test, we see that iHD significantly outperforms other compared methods except for Entropy and HDDT while iHDw outperforms all the compared methods.

For imbalanced datasets, the average ranks based on AUC also indicate the superiority of iHDw against other seven classifiers. Friedman's χ_F^2 statistic is 32.42 followed by Iman's F statistic as $F_F = 5.72$. Since at 95% confidence level F_F is larger than the critical value, $F(7, 133)$, the performance of the eight classifiers are not equivalent on the imbalanced datasets. The post-hoc Nemenyi test states that the iHDw is the best performing classifier by outperforming Entropy, Gini, DCSM, GR, HDDT and CCPDT while iHD only outperforms Gini.

Between iHD and iHDw, iHD performs better than iHDw on balanced datasets. However, the counts of Win/Tie/Loss (**W/T/L**) of iHD and iHDw against other classifiers suggests that DTs based on iHDw is the better performing classifier in most of the cases. Therefore, from the above results, we can say that although iHD gives better performance than iHDw on balanced datasets, taking account of the comparisons with other methods, iHDw is considerably better performing classifier on both balanced and imbalanced datasets. As unpruned DTs are constructed, we also compare the node size and tree construction time of the eight classifiers. However, we do not find major differences between iHD and iHDw compared to the best performing existing criterion on node size and construction time. Moreover, DT using iHDw provides superior performance without requiring much additional time to construct the tree than iHD.

5 Conclusion

In this paper, we propose two new splitting criteria for measuring the goodness of a split in a decision tree learning. The proposed splitting criteria favor mutually exclusive and purer partitions. Results over a large number of datasets provide the evidence that the decision trees constructed using proposed criteria are better than other six related splitting criteria on both balanced and imbalanced datasets. As future research direction, we will extend the work for tree-based ensemble classifiers and also want to investigate the effect of the proposed split criteria on pruning techniques.

Acknowledgments

This research is supported by the fellowship from ICT Division, Ministry of Posts, Telecommunications and Information Technology, Bangladesh. No - 56.00.0000.028.33.002.19.3; Dated 09.01.2019.

References

[Alcalá-Fdez *et al.*, 2011] Jesús Alcalá-Fdez, Alberto Fernández, Julián Luengo, Joaquín Derrac, and Salvador García. Keel data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. *Journal of Multiple-Valued Logic and Soft Computing*, 17(2-3):255–287, 2011.

[Breiman *et al.*, 1984] Leo Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees*. Wadsworth, 1984.

[Chandra *et al.*, 2010] B Chandra, Ravi Kothari, and Pallath Paul. A new node splitting measure for decision tree construction. *Pattern Recognition*, 43(8):2725–2731, 2010.

[Chawla *et al.*, 2004] Nitesh V Chawla, Nathalie Japkowicz, and Aleksander Kotcz. Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6, 2004.

[Cichocki and Amari, 2010] Andrzej Cichocki and Shunichi Amari. Families of alpha-beta-and gamma-divergences: Flexible and robust measures of similarities. *Entropy*, 12(6):1532–1568, 2010.

[Cieslak and Chawla, 2008] David A Cieslak and Nitesh V Chawla. Learning decision trees for unbalanced data. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 241–256. Springer, 2008.

[Cieslak *et al.*, 2012] David A Cieslak, T Ryan Hoens, Nitesh V Chawla, and W Philip Kegelmeyer. Hellinger distance decision trees are robust and skew-insensitive. *Data Mining and Knowledge Discovery*, 24(1):136–158, 2012.

[Demšar, 2006] Janez Demšar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine learning research*, 7(Jan):1–30, 2006.

[Drummond and Holte, 2000] Chris Drummond and Robert C Holte. Exploiting the cost (in) sensitivity of decision tree splitting criteria. In *ICML*, volume 1, pages 239–246, 2000.

[Dua and Graff, 2017] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[Friedman, 1937] Milton Friedman. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701, 1937.

[Friedman, 1940] Milton Friedman. A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, 11(1):86–92, 1940.

[Hand and Till, 2001] David J Hand and Robert J Till. A simple generalisation of the area under the roc curve for multiple class classification problems. *Machine learning*, 45(2):171–186, 2001.

[Hoens *et al.*, 2012] T Ryan Hoens, Qi Qian, Nitesh V Chawla, and Zhi-Hua Zhou. Building decision trees for the multi-class imbalance problem. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 122–134. Springer, 2012.

[Iman and Davenport, 1980] Ronald L Iman and James M Davenport. Approximations of the critical region of the fbietkan statistic. *Communications in Statistics-Theory and Methods*, 9(6):571–595, 1980.

[Iqbal *et al.*, 2017] Md Tauhid Bin Iqbal, Mohammad Shoyaib, Byungyong Ryu, M Abdullah-Al-Wadud, and Oksam Chae. Directional age-primitive pattern (dapp) for human age group recognition and age estimation. *IEEE Transactions on Information Forensics and Security*, 12(11):2505–2517, 2017.

[Kotsiantis *et al.*, 2007] Sotiris B Kotsiantis, I Zaharakis, and P Pintelas. Supervised machine learning: A review of classification techniques. *Emerging artificial intelligence applications in computer engineering*, 160:3–24, 2007.

[Liu *et al.*, 2010] Wei Liu, Sanjay Chawla, David A Cieslak, and Nitesh V Chawla. A robust decision tree algorithm for imbalanced data sets. In *Proceedings of the 2010 SIAM International Conference on Data Mining*, pages 766–777. SIAM, 2010.

[Nemenyi, 1963] Peter Nemenyi. *Distribution-free multiple comparisons*. PhD thesis, Princeton University, 1963.

[Quinlan, 1986] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[Quinlan, 1993] J. Ross Quinlan. *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers Inc., 1993.

[Sharmin *et al.*, 2019] Sadia Sharmin, Mohammad Shoyaib, Amin Ahsan Ali, Muhammad Asif Hossain Khan, and Oksam Chae. Simultaneous feature selection and discretization based on mutual information. *Pattern Recognition*, 91:162–174, 2019.

[Shih, 1999] Y-S Shih. Families of splitting criteria for classification trees. *Statistics and Computing*, 9(4):309–315, 1999.

[Su and Cao, 2019] Chong Su and Jie Cao. Improving lazy decision tree for imbalanced classification by using skew-insensitive criteria. *Applied Intelligence*, 49(3):1127–1145, 2019.

[Taylor and Silverman, 1993] Paul C Taylor and Bernard W Silverman. Block diagrams and splitting criteria for classification trees. *Statistics and Computing*, 3(4):147–161, 1993.