Learning Decision Trees for Unbalanced Data

David A. Cieslak and Nitesh V. Chawla
{dcieslak,nchawla}@cse.nd.edu

University of Notre Dame, Notre Dame IN 46556, USA

Abstract. Learning from unbalanced datasets presents a convoluted problem in which traditional learning algorithms may perform poorly. The objective functions used for learning the classifiers typically tend to favor the larger, less important classes in such problems. This paper compares the performance of several popular decision tree splitting criteria – information gain, Gini measure, and DKM – and identifies a new skew insensitive measure in Hellinger distance. We outline the strengths of Hellinger distance in class imbalance, propose its application in forming decision trees, and perform a comprehensive comparative analysis between each decision tree construction method. In addition, we consider the performance of each tree within a powerful sampling wrapper framework to capture the interaction of the splitting metric and sampling. We evaluate over this wide range of datasets and determine which operate best under class imbalance.

1 Introduction

Data sets in which one class is particularly rare, but more important – termed unbalanced or unbalanced datasets – continue to be a pervasive problem in a large variety of supervised learning applications, ranging from telecommunications to finance to medicine to web categorization to biology. Typically sampling methods [1–5] are used for countering class imbalance.

Decision trees, particularly C4.5 [6], have been among the more popular algorithms that have been significantly helped by sampling methods for countering the high imbalance in class distributions [3, 4, 7]. In fact, the vast majority of papers in the ICML'03 Workshop on unbalanced Data included C4.5 as the base classifier. While it is understood that sampling generally improves decision tree induction, what is undetermined is the interaction between sampling and how those decision trees are formed. C4.5 [6] and CART [8] are two popular algorithms for decision tree induction; however, their corresponding splitting criteria — information gain and the Gini measure — are considered to be skew sensitive [9]. It is because of this specific sensitivity to class imbalance that use of sampling methods prior to decision tree induction has become a de facto standard in the literature. The sampling methods alter the original class distribution, driving the bias towards the minority or positive class[1]. Dietterich, Kearns, and Mansour [10] suggested an improved splitting criterion for a top down decision tree induction, now known as DKM. Various authors have implemented DKM as a

[1] Without loss of generality, we will assume that positive and minority class is the same.

decision tree splitting criterion and shown its improved performance on unbalanced datasets [9, 11, 12]. However, DKM has also been shown to be (weakly) skew-insensitive [9, 11].

We posit that class imbalance is also a characteristic of sparseness in feature space, in addition to the skewed class distributions. Thus, it becomes important to design a decision tree splitting criterion that captures the divergence in distributions without being dominated by the class priors. To that end, we consider the Hellinger distance [13, 14] as a decision tree splitting criterion, which we show to be skew-insensitive. We also demonstrate similarities between DKM and Hellinger distance, albeit Hellinger distance offers a stronger skew insensitivity. Finally, we consider the popular sampling methods and study their impact on the decision tree splitting criteria. *Does having a skew insensitive splitting criterion mitigate the need of sampling?*

Contributions: Our key contributions include the following: 1) Characterization of the Hellinger distance metric in data mining context as a skew insensitive decision tree splitting metric. 2) Analytical demonstration of the utility of proposed formulation of Hellinger distance using isometrics graphs. 3) A theoretical comparison between Hellinger distance and DKM. 4) A decision tree algorithm called HDDT incorporating the Hellinger distance as the tree splitting criterion. 5) Comparison of the effect of sampling on the decision tree splitting methods. We have used a total of 19 datasets, from UCI and real-world domains, with varying properties and skew in this study. We have also used statistical measures suggested by Demsar [15] to robustly compare the classifiers across multiple datasets. Note that we only used unpruned decision trees for all our experiments, irrespective of the splitting criterion used, as the previous work has pointed to the limitations of pruning for unbalanced datasets [16, 17].

2 Hellinger Distance

Hellinger distance is a measure of distributional divergence [13, 14]. Let (Θ, λ) denote a measurable space with P and Q as two continuous distributions with respect to the parameter λ. Let p and q be the densities of the measures P and Q with respect to λ. The definition of Hellinger distance can be given as:

$$d_H(P,Q) = \sqrt{\int_\Omega (\sqrt{P} - \sqrt{Q})^2 d\lambda} \tag{1}$$

This is equivalent to:

$$d_H(P,Q) = \sqrt{2(1 - \int_\Omega \sqrt{PQ} d\lambda)} \tag{2}$$

where $\int_\Omega \sqrt{pq} d\lambda$ is the Hellinger integral. Note the Hellinger distance does not depend on the choice of the dominating parameter λ. It can also be defined for a countable space Φ, as $d_H(P,Q) = \sqrt{\Sigma_{\phi \in \Phi} (\sqrt{P(\phi)} - \sqrt{Q(\phi)})^2}$. The Hellinger

distance carries the following properties: 1) $d_H(P,Q)$ is in $[0,\sqrt{2}]$. 2) d_H is symmetric and non-negative, implying $d_H(P,Q) = d_H(Q,P)$. Moreover, squared Hellinger distance is the lower bound of KL divergence.

In this paper, the P and Q in Equations 1 & 2 are assumed to be the normalized frequencies of feature values across classes. This allows us to capture the notion of "affinity" between the probability measures on a finite event space. If P = Q, then distance = 0 (maximal affinity) and if P and Q are completely disjoint then distance = $\sqrt{(2)}$ (zero affinity). This dictates the decision tree splitting criterion for separability between classes. We want to select a feature that carries the minimal affinity between the classes. Thus, the Hellinger distance can be used to capture the propensity of a feature to separate class distributions.

For application as a decision tree splitting criterion, we assume a countable space, so we discretize all continuous features into p partitions or bins. Assuming a two-class problem (class + and class −), let X_+ be class + and X_- be class −. Then, we are essentially interested in calculating the "distance" in the normalized frequencies aggregated over all the partitions of the two class distributions X_+ and X_-. The Hellinger distance between X_+ and X_- is:

$$d_H(X_+, X_-) = \sqrt{\sum_{j=1}^{p} \left(\sqrt{\frac{|X_{+j}|}{|X_+|}} - \sqrt{\frac{|X_{-j}|}{|X_-|}} \right)^2} \tag{3}$$

We postulate that this formulation is strongly skew insensitive as the prior does not influence the distance calculation. It essentially captures the divergence between the feature value distributions given the two different classes. There is no factor of class prior. We will show the effectiveness of this enumeration isometric plots.

2.1 Comparing isometrics

Vilalta & Oblinger [18] proposed the use of isometric lines to define the bias of an evaluation metric by plotting contours for a given metric over the range of possible values. They presented a case study on information gain. While they did not produce isometrics under class skew, they note that "A highly skewed distribution may lead to the conclusion that two metrics yield similar generalization effects, when in fact a significant difference could be detected under equal class distribution. [18]" Subsequently, Flach [9] connected the isometric plots to ROC analysis, demonstrating the effects of true and false positives on several common evaluation metrics: accuracy, precision, and f-measure. In addition, he also presented isometrics for three major decision tree splitting criteria: entropy (used in information gain) [6], Gini index [8], and DKM [10]. Flach also established the effect on class skew on the shape of these isometrics [9].

We adopted the formulation of Flach in this paper, where the isometric plots show the contour lines in 2D ROC space, representative of the performance of different decision tree splitting criteria with respect to their estimated true and false positive rates, conditioned on the skew ratio ($c = \frac{neg}{pos}$). A decision

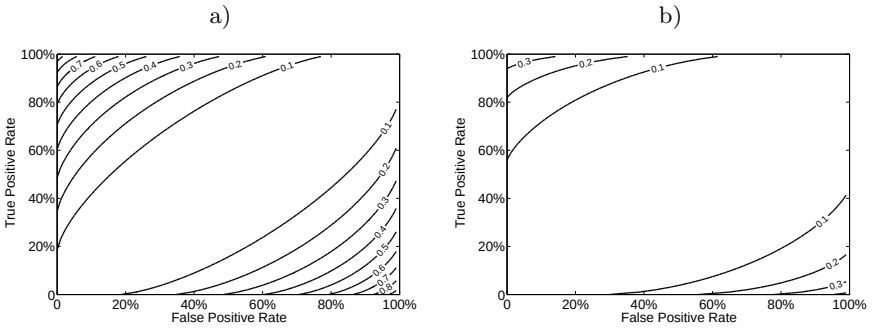

Fig. 1. Information gain isometrics for (+:-)=(1:1) in (a) and (+:-)=(1:10) in (b).

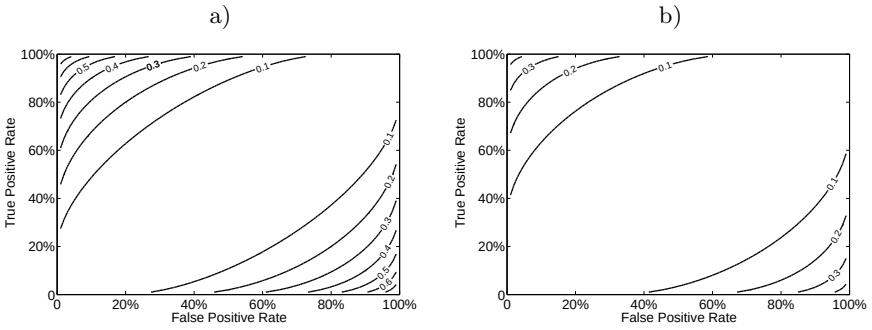

Fig. 2. DKM isometrics for (+:-)=(1:1) in (a) and (+:-)=(1:10) in (b).

tree split, for a binary class problem, can be defined by a confusion matrix as follows. A parent node will have POS positive examples and NEG negative examples. Assuming a binary split, one child will carry the true and false positive instances, and the other child will carry the true and false negative instances. The different decision tree splitting criteria, as considered in this paper, can then be modeled after this impurity (distribution of positives and negatives). Thus, in the isometric plots, each contour represents the combinations of true positives and false negatives that will generate a particular value for a given decision tree splitting criterion. For example, the 0.1 contour in Figure 1 (a) indicates that the value of information gain is 0.1 at (fpr, tpr) of approximately $(0\%, 20\%)$, $(20\%, 60\%)$, $(80\%, 100\%)$, $(20\%, 0\%)$, $(60\%, 20\%)$, $(100\%, 80\%)$, and all other combinations along the contour. In Figures 1 (a) & (b), information gain is observed as contours formed in ROC space under a $(+ : -)$ skew of $(1 : 1)$ and $(1 : 10)$, respectively. As the skewness increases, the isometrics become flatter and information gain will operate more poorly as a splitting criterion. Vilalta & Oblinger [18] and Flach [9] observed similar trends. Additionally, Flach [9] notes that DKM is (weakly) skew-insensitive. It is affected like information gain (and therefore C4.5) and Gini (and therefore CART) which are highly skew dependent,

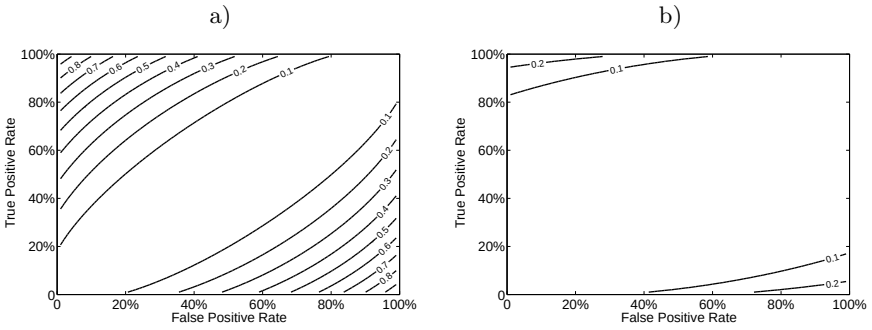

Fig. 3. Gini isometrics for (+:-)=(1:1) in (a) and (+:-)=(1:10) in (b).

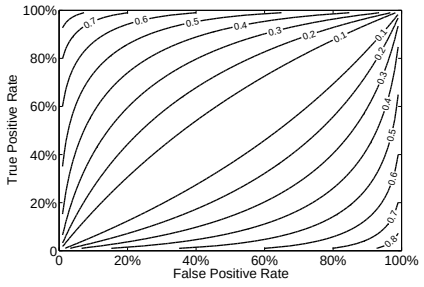

Fig. 4. Hellinger distance isometric for any (+:-)

but not to the same degree. Additionally, its contours do not "twist" – there is some a for which each contour intersects $(0, a)$ and $(1 - a, 0)$ – under skew. Gini is by far the most skew sensitive metric of this group. We only considered two class proportions of $(1 : 1)$ and $(1 : 10)$ to highlight the impact of even a marginal class skew. We point the interested reader to the paper by Flach for a more elaborate analysis of class skew using isometrics on these three metrics [9].

On the other hand, an important observation may be drawn from an isometric of Hellinger distance. First, using Flach's model of relative impurity, we derive the following for Hellinger distance: $\sqrt{(\sqrt{tpr} - \sqrt{fpr})^2 + (\sqrt{1 - tpr} - \sqrt{1 - fpr})^2}$. Figure 4 contains Hellinger distance contours. The Hellinger distance isometrics will not deviate from the contours with varying class skew (c), as there is no factor of c in the relative impurity formulation. This result follows from the previous section and the independence of the Hellinger distance to the parameter λ, which in our case is the respective class priors. The isometric contours for Hellinger distance are unaffected by an increase in the class skew rate.

2.2 Comparing DKM and Hellinger distance

We posit that DKM and Hellinger distance have similar properties, albeit Hellinger distance has stronger skew insensitivity than DKM. We consider both DKM and

Hellinger distance within the canonical two class, binary split problem. $P(L)$ and $P(R)$ designate the weight of examples falling under the left and right branches respectively. $P(+)$ and $P(-)$ represent the probability of belonging to class $+$ and $-$. In these terms, we may state DKM as follows.

$$d_{DKM} = 2\sqrt{P(+)P(-)} - 2P(L)\sqrt{P(L|+)P(L|-)} - 2P(R)\sqrt{P(R|+)P(R|-)} \quad (4)$$

Applying these terms to Equation 3, the same terms, Hellinger distance maybe be stated as follows.

$$d_H = \sqrt{\left(\sqrt{P(L|+)} - \sqrt{P(L|-)}\right)^2 + \left(\sqrt{P(R|+)} - \sqrt{P(R|-)}\right)^2} \quad (5)$$

$$d_H = \sqrt{2 - 2\sqrt{P(L|+)P(L|-)} - 2\sqrt{P(R|+)P(R|-)}} \quad (6)$$

Representing the equations in this form demonstrate a clear similarity between Hellinger and DKM. Both are capturing the divergence in conditionals at a split point, albeit with some differences. DKM places the notion of "branch strength" in terms of $P(L)$ and $P(R)$ for each of the corresponding left and right branch conditionals. Moreover, DKM also takes into account the class priors as $2\sqrt{P(+)P(-)}$, which can also be considered as the upper bound for pure splits. On the other hand, Hellinger is upper bounded by $\sqrt{2}$, and does not take into account the notion of class skew in the calculation. It is simply capturing the deviation between $P(x|y)$ at a split node, without factoring in the relative class distributions at the parent node, which DKM does. This also highlights why DKM may be less skew insensitive than Hellinger distance.

Hellinger distance aims for "more pure" leaves as it aims for partitioning the space by capturing the deviations in the class conditionals at the node. However, this can result in smaller coverage, which may be damaging for more balanced class rates, but could prove helpful for highly unbalanced datasets as it tries to form purer leaves that are minority class specific. Nevertheless, it depends on the relative distribution of features with respect to the classes. DKM, on the other hand, may not be as greedy and stop the splitting for the sake of larger coverage.

We demonstrate this property using value surfaces in Figure 5, which display the full range of possible split values for both metrics. Figure 5(a) shows the Hellinger distance throughout all possible class skew ratios, while Figure 5(b),(c), & (d) display DKM values for the $(+:-)$ class ratios of (1:1), (1:10), & (1:100), respectively. As skew increases, the DKM surface flattens and potentially reduces the set of usable values, as it gets dominated by the skew factor in the favor of majority class. We do note that in a vast number of data sets such a scenario may not arise, and Hellinger and DKM may converge to similar performances. Nevertheless, it is important to consider this difference as we may want to grow a completely unpruned tree for unbalanced datasets, and at lower nodes in the tree the class skew may get to the very extreme. This is obviously conditioned on the property of the data, but theoretically it is possible. At this point, Hellinger

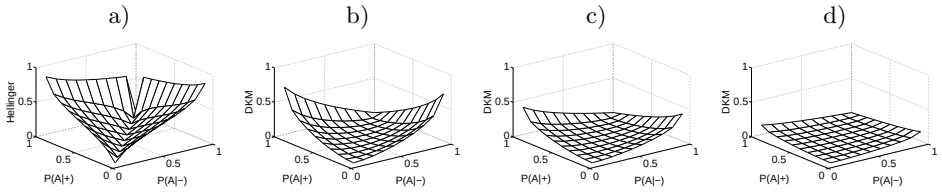

Fig. 5. Full value surfaces for the total range of Hellinger distances and DKM. (a) The Hellinger distance remains unchanged over all possible class skews. The range of DKM values vary with class skew: we note the (+:-) ratios of (b) (1:1), (c) (1:10), & (d) (1:100).

distance may prove more amenable. Thus, we suggest use of Hellinger over DKM given its stronger skew insensitivity, albeit with the caveat that at the general case both Hellinger and DKM will converge to similar performances. But, we want to be prepared for the worst case.

2.3 HDDT: Hellinger Distance Decision Tree

The following algorithm outlines the approach to incorporating Hellinger distance in learning decision trees. We will refer to Hellinger distance and Hellinger distance based decision trees as HDDT for the rest of the paper. In our algorithm, $T_{y=i}$ indicates the subset of training set T that has class i, $T_{x_k=j}$ specifies the subset with value j for feature k, and $T_{x_k=j,y=i}$ identifies the subset with class i and has value j for feature k .

Algorithm 1 *Calc_Hellinger*

Input: Training set T, Feature f

1: **for** each value v of f **do**

2: $\quad Hellinger+ = (\sqrt{|T_{x_f=v,y=+}|/|T_{y=+}|} - \sqrt{|T_{x_f=v,y=-}|/|T_{y=-}|})^2$

3: **end for**

4: **return** $\sqrt{Hellinger}$

In the case that a given feature is continuous, a slight variant to Algorithm 1 is used in which *Calc_Hellinger* sorts based on the feature value, finds all meaningful splits, calculates the binary Hellinger distance at each split, and returns the highest distance. This is identical to the methodology used by C4.5. With this practical distance calculator, Algorithm 2 outlines the procedure for inducing HDDT trees.

Algorithm 2 $HDDT$

Input: Training set T, Cut-off size C

if $|T| < C$ **then**
  return
end if
for each feature f of T **do**
  $H_f = Calc\_Hellinger(T, f)$
end for
$b = \max(H)$
for each value **v** of b **do**
  $HDDT(T_{x_b=v}, C)$
end for

We do not consider any pruning with HDDT and smoothed the leaf frequencies by the Laplace estimate. This was primarily motivated by the observations of Provost & Domingos [16]. We likewise considered only the unpruned decision trees for C4.5, CART, and DKM, and smoothed the leaf frequencies by the Laplace estimate.

3 Sampling Methods

Treatment of class imbalance by under- and/or over-sampling, including variants of the same, has resulted in improvement in true positives without significantly increasing false positives [3, 19]. However, we believe it is important to understand the interaction of the sampling methods with different decision tree splitting metrics with different skew sensitivities. This study examines combining two samplings methods: random undersampling and SMOTE [5]. While seemingly primitive, randomly removing majority class examples has been shown to improve performance in class imbalance problems. Some training information is lost, but this is counterbalanced by the improvement in minority class accuracy and rank-order. SMOTE is an advanced oversampling method which generates synthetic examples at random intervals between known positive examples.

Elkan discusses the interaction of cost and class imbalance [20], proposing a simple method to calculate optimal sampling levels. However, our evaluation occurs without explicit costs. In this case, Elkan's calculation simply indicates sampling the classes to a balanced proportion. In addition, this approach leaves open much to interpretation: should the negative class be undersampled, the positive class be oversampled, or should a combination be used to reach the balance point? To address this, we search a larger sampling space (which includes several potential balance points) via wrapper to determine optimal class proportions [19]. Testing for each pair of undersampling and SMOTE percentages will result in an intractable search space. The wrapper framework first explores the amounts of undersampling that result in an improvement in performance over the baseline, where baseline is defined as the decision tree classifier learned on

the original distribution of data. Subsequently, once the majority class is undersampled to the point where the performance does not deteriorate anymore, the wrapper searches for the appropriate levels of SMOTE. This strategy removes the "excess" negative examples, thereby reducing the size of the training dataset and making learning time more tractable. Then SMOTE adds synthetic positive examples and generalizes performance of the classifier over the positive class. AUROC is the primary metric for considering performance in unbalanced datasets, so it will be used both as a wrapper objective function and the final performance metric. We point the reader to the paper by Chawla et al. [19] for further details on the wrapper framework.

We perform experiments using each base decision tree classifier in combination with the sampling wrapper. We note that both undersampling and SMOTE contain elements of randomization. Therefore, we first constructed an exhaustive sets of sampled datasets at different amounts of undersampling and different amounts of SMOTE. We let the wrapper search only on these prior constructed undersampled datasets and SMOTE datasets to determine the appropriate levels of sampling for different splitting criteria. For example, each splitting metric considers the removal of exactly the same majority class examples in the first comparison to the baseline. Of course, each splitting criterion may converge to different amounts of undersampling. But this ensures uniformity of results and that the potential performance differences stem from the bias of the decision tree metrics themselves, rather than possible variance due to randomness in the applied sampling methods.

4 Experimental Evaluation

In this section, we provide experimental results to determine performance compares the characteristics of HDDT, C4.5, CART, and DKM and the combination of each with the sampling wrapper. We use a variety of unbalanced, binary-class, mixed feature type real-world and UCI datasets. Such a wide variety should comprehensively outline the strengths and weaknesses of using more skew insensitive metrics such as DKM or Hellinger versus information gain and Gini. We used the 5x2-fold cross-validation (cv) over 10-fold cv as that is more appropriate for unbalanced data sets, as the latter can result in an elevated Type 1 error [21], which is particularly punishing for unbalanced datasets because of the trade-off between false positives and false negatives. Demsar [15] also encourages use of 5x2 cross-validation for statistical comparisons among classifiers across datasets. We statistically evaluate and compare classifiers using the Holm procedure of the Friedman test – a procedure to determine the statistical significance of performance rankings across multiple datasets [15].

4.1 Datasets

Table 1 describes the characteristics of the datasets used in our experiments. We have a number of real-world and UCI datasets. We will briefly describe

Table 1. All the datasets used in this paper.

No.	Dataset	Examples	Features	MinClass %
1	Boundary	3,505	174	4%
2	Breast-W	569	32	37%
3	Calmodoulin	18,916	131	5%
4	E-State	5,322	12	12%
5	Forest Cover	38,500	10	7.1%
6	FourClass	862	2	36%
7	German.Numer	1,000	24	30%
8	Letter	20,000	16	19%
9	Mammography	11,183	6	2%
10	Oil	937	49	4%
11	Page	5,473	10	10%
12	Pendigits	10,992	16	10%
13	Phoneme	5,404	5	21%
14	PhosS	11,411	479	5%
15	Pima	768	8	35%
16	Satimage	6,435	36	10%
17	Segment	2,310	19	14%
18	Splice	1,000	60	4.8%
19	SVMGuide1	3,089	4	35%

the real-world datasets used in this paper. E-state contains electrotopological state descriptors for a series of compounds from the National Cancer Institute's Yeast AntiCancer drug screen. Mammography is highly unbalanced and records information on calcification in a mammogram. Oil dataset contains information about oil spills; it is relatively small and very noisy [22]. The Phoneme dataset originates from the ELENA project and is used to distinguish between nasal and oral sounds. Boundary, Calmodoulin, and PhosS are various biological datasets [23]. FourClass, German.Numer, Splice, and SVMGuide1 all are available from $LIBSVM$ [24]. The remaining datasets all originate from the UCI repository [25]. Some of these are originally multiple class datasets and were converted into 2-class problems by keeping the smallest class as minority and the rest as majority. The exception is Letter, for which each vowel became a member of the minority class, against all of the consonants as the majority class. Aside from stated modifications, each dataset is used "as is."

4.2 Experimental Results

Baseline Comparisons We first compare all the baseline decision tree algorithms. In Table 2, we report the average $AUROC$ over the 5x2 cv experimental framework. The relative ranking for each classifier is indicated parenthetically. Using the Holm procedure of the Friedman test [15] for comparing the ranking across all the 19 datasets and 4 classifiers, we determine HDDT and DKM are statistically significantly better than C4.5 and CART decision trees at 95% confidence interval. Thus, when applying decision trees to unbalanced data, selecting HDDT or DKM will typically yield a significant edge over C4.5 and CART. In general, DKM and HDDT converge towards similar trees and therefore the final performance. This is reflected by ties on 15 of the 19 datasets, with a marginal improvement in HDDT average ranks over DKM. Thus, these empirical obser-

Table 2. Baseline decision tree *AUROC* results with relative ranks in parentheses, with the average rank applied for all ties. HDDT achieves the best over-all ranking. We use the Friedman test to compare the ranks at 95% confidence interval as per the recommendation of Demsar [15] that it is more appropriate to compare classifiers' ranks when using multiple classifiers and multiple datasets. A ✓ in the bottom row indicates that HDDT statistically significantly improved over that classifier.

Dataset	C4.5	DKM	CART	HDDT
Boundary	0.554 ± 0.037 (4)	0.606 ± 0.044 (1)	0.558 ± 0.029 (3)	0.594 ± 0.039 (2)
Breast-w	0.948 ± 0.011 (3)	0.952 ± 0.010 (1.5)	0.937 ± 0.017 (4)	0.952 ± 0.010 (1.5)
Calmodoulin	0.668 ± 0.009 (3)	0.670 ± 0.012 (2)	0.621 ± 0.012 (4)	0.680 ± 0.010 (1)
E-State	0.554 ± 0.035 (3)	0.579 ± 0.014 (2)	0.547 ± 0.020 (4)	0.580 ± 0.014 (1)
Forest Cover	0.978 ± 0.002 (3)	0.982 ± 0.002 (1.5)	0.963 ± 0.004 (4)	0.982 ± 0.002 (1.5)
Fourclass	0.969 ± 0.011 (3)	0.975 ± 0.013 (1.5)	0.946 ± 0.023 (4)	0.975 ± 0.013 (1.5)
German.numer	0.705 ± 0.016 (1)	0.692 ± 0.028 (2.5)	0.629 ± 0.027 (4)	0.692 ± 0.028 (2.5)
Letter	0.990 ± 0.004 (2)	0.990 ± 0.004 (2)	0.962 ± 0.006 (4)	0.990 ± 0.004 (2)
Mammography	0.889 ± 0.008 (3)	0.912 ± 0.013 (1.5)	0.858 ± 0.017 (4)	0.912 ± 0.013 (1.5)
Oil	0.787 ± 0.074 (4)	0.799 ± 0.042 (2.5)	0.815 ± 0.052 (1)	0.799 ± 0.042 (2.5)
Page	0.971 ± 0.004 (3)	0.974 ± 0.005 (1.5)	0.964 ± 0.010 (4)	0.974 ± 0.005 (1.5)
Pendigits	0.985 ± 0.005 (3)	0.992 ± 0.002 (1.5)	0.976 ± 0.007 (4)	0.992 ± 0.002 (1.5)
Phoneme	0.892 ± 0.010 (3)	0.905 ± 0.006 (2)	0.887 ± 0.007 (4)	0.906 ± 0.005 (1)
PhosS	0.638 ± 0.025 (4)	0.677 ± 0.009 (1.5)	0.648 ± 0.017 (3)	0.677 ± 0.009 (1.5)
Pima	0.753 ± 0.013 (3)	0.760 ± 0.019 (1.5)	0.724 ± 0.019 (4)	0.760 ± 0.019 (1.5)
Satimage	0.906 ± 0.009 (3)	0.911 ± 0.008 (1.5)	0.862 ± 0.011 (4)	0.911 ± 0.007 (1.5)
Segment	0.982 ± 0.006 (3)	0.984 ± 0.007 (1.5)	0.977 ± 0.007 (4)	0.984 ± 0.007 (1.5)
Splice	0.954 ± 0.016 (1)	0.950 ± 0.014 (2.5)	0.806 ± 0.035 (4)	0.950 ± 0.014 (2.5)
SVMguide1	0.985 ± 0.005 (3.5)	0.989 ± 0.002 (1.5)	0.985 ± 0.002 (3.5)	0.989 ± 0.002 (1.5)
Avg. Rank	2.92	1.74	3.71	1.63
Friedman $\alpha = .05$	✓		✓	—

vations agree with the isometric analyses and discussion in the previous Section that as the splitting criterion becomes relatively more skew-insensitive, decision trees tend to perform more strongly on unbalanced data.

Interaction with Sampling We now consider the effect of sampling on each of the decision tree splitting criterion. We used a wrapper approach, as described in Section 3, to determine the potentially optimal levels of sampling for each of the decision tree algorithms. The wrapper optimized on AUROC. Note that the wrapper uses a separate validation framework to determine the sampling levels. Each decision tree algorithm used 5-fold cross-validation on the training set of the 5x2 cv to determine the optimal sampling levels by optimizing on AUROC. Once these were determined, the entire training set was resampled by that amount and evaluated on the corresponding 5x2 cv testing set. This approach is outlined in the paper by Chawla et al. [19]. The performances that we report are on the testing set of the 5x2 cv.

The results on the 5x2 cv are shown in Table 3. These results show a compelling trend. The benefits of DKM and HDDT over C4.5 are clearly eroded. CART still remains the worst performing classifier, and statistically significantly so. However, there are a couple of exceptions, such as Breast-w and Oil, in which CART and sampling produces the best classifier. We note that these datasets are very small (Breast-w being the smallest and Oil being the fourth smallest)

Table 3. *AUROC* values produced by each decision tree in combination with the sampling wrapper. Relative ranking is noted parenthetically. A ✓ in the bottom row indicates a 95% significant improvement over CART. There was no statistically significant difference among the other three decision tree classifiers – C4.5, HDDT, and DKM.

Dataset	C4.5	DKM	CART	HDDT
Boundary	0.616 ± 0.033 (1)	0.602 ± 0.023 (3)	0.582 ± 0.026 (4)	0.604 ± 0.029 (2)
Breast-w	0.953 ± 0.008 (3)	0.953 ± 0.008 (3)	0.955 ± 0.007 (1)	0.953 ± 0.008 (3)
Calmodoulin	0.676 ± 0.007 (1)	0.660 ± 0.011 (3.5)	0.660 ± 0.007 (3.5)	0.669 ± 0.010 (2)
E-State	0.580 ± 0.016 (1)	0.575 ± 0.013 (2.5)	0.560 ± 0.012 (4)	0.575 ± 0.013 (2.5)
Forest Cover	0.980 ± 0.002 (3)	0.983 ± 0.001 (1.5)	0.974 ± 0.001 (4)	0.983 ± 0.001 (1.5)
Fourclass	0.965 ± 0.012 (3)	0.971 ± 0.009 (1.5)	0.943 ± 0.010 (4)	0.971 ± 0.009 (1.5)
German.numer	0.690 ± 0.015 (1)	0.687 ± 0.015 (3)	0.668 ± 0.016 (4)	0.688 ± 0.014 (2)
Letter	0.989 ± 0.002 (2)	0.989 ± 0.003 (2)	0.977 ± 0.004 (4)	0.989 ± 0.003 (2)
Mammography	0.909 ± 0.011 (1)	0.906 ± 0.013 (2.5)	0.905 ± 0.008 (4)	0.906 ± 0.013 (2.5)
Oil	0.789 ± 0.029 (4)	0.803 ± 0.028 (3)	0.806 ± 0.041 (1)	0.804 ± 0.029 (2)
Page	0.978 ± 0.004 (1)	0.976 ± 0.004 (2.5)	0.970 ± 0.006 (4)	0.976 ± 0.004 (2.5)
Pendigits	0.987 ± 0.003 (3)	0.991 ± 0.002 (1.5)	0.982 ± 0.003 (4)	0.991 ± 0.002 (1.5)
Phoneme	0.894 ± 0.005 (3)	0.902 ± 0.006 (1.5)	0.890 ± 0.006 (4)	0.902 ± 0.006 (1.5)
PhosS	0.670 ± 0.013 (1)	0.666 ± 0.015 (2)	0.665 ± 0.019 (3.5)	0.665 ± 0.015 (3.5)
Pima	0.755 ± 0.013 (3)	0.759 ± 0.014 (1)	0.742 ± 0.014 (4)	0.758 ± 0.013 (2)
Satimage	0.904 ± 0.004 (3)	0.910 ± 0.005 (1.5)	0.887 ± 0.006 (4)	0.910 ± 0.005 (1.5)
Segment	0.982 ± 0.006 (3)	0.984 ± 0.007 (1.5)	0.980 ± 0.006 (4)	0.984 ± 0.007 (1.5)
Splice	0.942 ± 0.013 (1)	0.933 ± 0.010 (2)	0.829 ± 0.015 (4)	0.932 ± 0.009 (3)
SVMguide1	0.987 ± 0.002 (3)	0.988 ± 0.001 (1.5)	0.985 ± 0.001 (4)	0.988 ± 0.001 (1.5)
Avg. Rank	2.16	2.13	3.71	2.08
Friedman $\alpha = .05$	✓	✓	—	✓

and have the lowest feature-to-example ratio, suggestive of the curse of dimensionality. Moreover, Oil is also very noisy.

One question at this stage is: *how much do the different decision trees benefit from sampling when compared to their respective baseline performances?* Figure 6 depicts the percentage improvement in AUROC across all the datasets after applying sampling for each of the different decision tree splitting criteria. This figure shows a very compelling trend: C4.5 and CART are the biggest gainers from sampling, while DKM and HDDT, being skew insensitive, do not achieve significant gains from sampling. In fact, we note that DKM and HDDT often experience a reduction in performance when sampling is applied. 14 out of 19 datasets show a reduction in AUROC for HDDT, and 15 out of the 19 datasets show a reduction in AUROC for DKM. This also points out that the wrapper overfits on the sampling amounts over the validation set, and diminishes generalization capability of HDDT or DKM. Thus, using the natural distribution and letting one of the skew insensitive splitting criteria work the way through the data can be potentially more beneficial over using the computationally expensive step of sampling with DKM and HDDT.

Finally, we point out that the amounts of sampling determined for each of the decision trees varied. We elucidate that we had first generated the various levels of samplings and then let the wrapper search from that space for each decision tree metric. This ensured that the differences were not due to randomness in sampling, and were more intrinsic to the base decision tree splitting metric. Table 5 in the Appendix shows the different sampling levels. HDDT and DKM

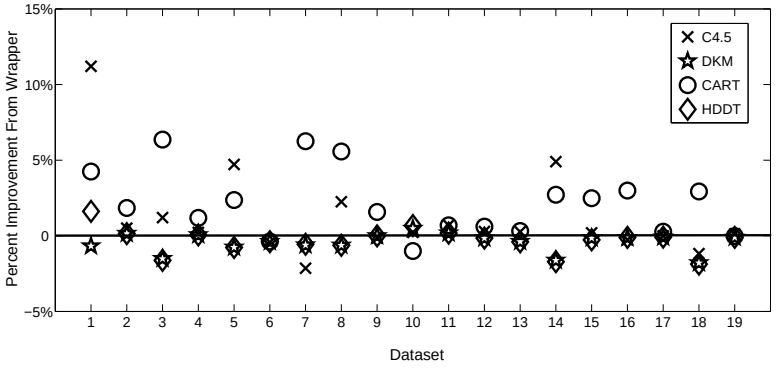

Fig. 6. Percent improvement in *AUROC* from sampling for each decision tree type, with relative rankings. We note that CART generally exhibits the highest improvement yielded from the wrapper. Position on the x-axis corresponds to the dataset number in Table 1.

continued to share similarities in the amounts of sampling level as well. C4.5 and CART generally required higher sampling levels than DKM and HDDT. While there were variations in the amounts of sampling, Friedman test's Holm procedure shows that there is no statistically significant difference in the ranks of levels of sampling for each decision tree.

Finally, we consider a comparison of all eight potential classifiers: each baseline decision tree and its wrapper-enhanced counterpart. We re-ranked all the 8 classifiers across the 19 datasets. Note that exactly the same training and testing sets were used for all decision tree classifiers, albeit the training sets were modified by sampling when used with the wrapper. These comparative rankings across each dataset are presented in Table 4. There are some interesting observations from this table. The baseline HDDT still achieves the best rank. Wrapper has a positive effect on C4.5 and CART, as pointed out earlier, but a negative effect on both DKM and HDDT. Thus, there is merit to using skew insensitive metrics over sampling. The statistical significance test establishes that HDDT is better than C4.5, CART, and Wrapper + CART.

5 Conclusions

The primary focus of this paper is learning decision trees on unbalanced datasets. We first propose Hellinger distance as a decision tree splitting criterion. We then thoroughly compare four different decision tree splitting criteria with different reactions to the skew in data distribution. We also considered an evaluation of the effect of sampling and how it impacts the different metrics differently. We draw the following conclusions.

Hellinger distance and DKM share similar properties. The isometric in Section 2.1 show Hellinger distance to be skew-invariant while the isometric plot

Table 4. Comparative *AUROC* ranks across the entire set of tested classifiers. A ✓ in the bottom row indicates that using the Friedman test HDDT is statistically significantly better (at 95%) than the respective classifier.

	Baseline				Wrapper			
Dataset	C4.5	DKM	CART	HDDT	C4.5	DKM	CART	HDDT
Boundary	8	2	7	5	1	4	6	3
Breast-w	7	5.5	8	5.5	3	3	1	3
Cam	5	3	8	1	2	6.5	6.5	4
Covtype	6	3.5	8	3.5	5	1.5	7	1.5
Estate	7	3	8	1.5	1.5	4.5	6	4.5
Fourclass	5	1.5	7	1.5	6	3.5	8	3.5
German.numer	1	2.5	8	2.5	4	5	6.5	6.5
ism	7	1.5	8	1.5	3	4.5	6	4.5
Letter	2	2	8	2	4	4	7	4
Oil	8	5.5	1	5.5	7	4	2	3
Page	6	4.5	8	4.5	1	2.5	7	2.5
Pendigits	6	1.5	8	1.5	5	3.5	7	3.5
Phoneme	6	2	8	1	5	3.5	7	3.5
PhosS	8	1.5	7	1.5	3	4	5.5	5.5
Pima	6	1.5	8	1.5	5	3	7	4
Satimage	5	1.5	8	1.5	6	3.5	7	3.5
Segment	5.5	2.5	8	2.5	5.5	2.5	7	2.5
Splice	1	2.5	8	2.5	4	5	7	6
SVMguide1	7	1.5	7	1.5	5	3.5	7	3.5
Avg. Rank	5.61	2.58	7.42	2.5	4	3.76	6.13	3.79
Friedman $\alpha = .05$	✓		✓	—			✓	

for DKM varies with class skew ratio. However, in Section 2.2, we go on to demonstrate that although there are divergent components of both metrics.This carries over into our experimental results where we note frequent convergence to identical performance.

HDDT and DKM produce superior decision trees under class imbalance. Without using any sampling, both DKM and HDDT statistically significantly outperformed C4.5 and CART.

Sampling generally benefits C4.5 and CART, and hurts DKM and HDDT. We believe this is a compelling observation of this study. We can avoid the use of sampling when using more appropriate decision tree splitting criteria, as those remain superior even after considering sampling. In general, we can recommend the use of HDDT as a decision tree methodology given its skew insensitive properties and the best ranks (no statistical significance over DKM).

As part of future work, we are expanding this study to include balanced and multi-class datasets. We also want to explore the effect of pruning and what pruning methods may be more appropriate for DKM and HDDT. While the focus of this study has been largely on decision trees, we believe rule-based classifiers can also consider Hellinger distance to separate and conquer the instance space.

References

1. Japkowicz, N.: Class Imbalance Problem: Significance & Strategies. In: International Conference on Artificial Intelligence (ICAI). (2000) 111–117
2. Kubat, M., Matwin, S.: Addressing the Curse of Imbalanced Training Sets: One-Sided Selection. In: International Conference on Machine Learning (ICML). (1997) 179–186
3. Batista, G., Prati, R., Monard, M.: A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. SIGKDD Explorations **6**(1) (2004) 20–29
4. Van Hulse, J., Khoshgoftaar, T., Napolitano, A.: Experimental perspectives on learning from imbalanced data. In: ICML. (2007) 935–942
5. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: Synthetic Minority Over-sampling Technique. Journal of Artificial Intelligence Research **16** (2002) 321–357
6. Quinlan, J.R.: Induction of Decision Trees. Machine Learning **1** (1986) 81–106
7. Chawla, N.V., Japkowicz, N., Kołcz, A., eds.: Proceedings of the ICML'2003 Workshop on Learning from Imbalanced Data Sets II. (2003)
8. Breiman, L., Friedman, J., Stone, C.J., Olshen, R.: Classification and Regression Trees. Chapman & Hall (1984)
9. Flach, P.A.: The Geometry of ROC Space: Understanding Machine Learning Metrics through ROC Isometrics. In: ICML. (2003) 194–201
10. Dietterich, T., Kearns, M., Mansour, Y.: Applying the weak learning framework to understand and improve C4.5. In: Proc. 13th International Conference on Machine Learning, Morgan Kaufmann (1996) 96–104
11. Drummond, C., Holte, R.: Exploiting the cost (in)sensitivity of decision tree splitting criteria. In: ICML. (2000) 239–246
12. Zadrozny, B., Elkan, C.: Obtaining calibrated probability estimates from decision trees and naive Bayesian classifiers. In: Proc. 18th International Conf. on Machine Learning, Morgan Kaufmann, San Francisco, CA (2001) 609–616
13. Kailath, T.: The Divergence and Bhattacharyya Distance Measures in Signal Selection. IEEE Transactions on Communications **15**(1) (February 1967) 52–60
14. Rao, C.: A Review of Canonical Coordinates and an Alternative to Corresponence Analysis using Hellinger Distance. Questiio (Quaderns d'Estadistica i Investigacio Operativa) **19** (1995) 23–63
15. Demsar, J.: Statistical Comparisons of Classifiers over Multiple Data Sets. Journal of Machine Learning Research **7** (2006) 1–30
16. Provost, F., Domingos, P.: Tree Induction for Probability-Based Ranking. Machine Learning **52**(3) (September 2003) 199–215
17. Chawla, N.V.: C4.5 and Imbalanced Data Sets: Investigating the Effect of Sampling Method, Probabilistic Estimate, and Decision Tree Structure. In: ICML Workshop on Learning from Imbalanced Data Sets II. (2003)
18. Vilalta, R., Oblinger, D.: A Quantification of Distance-Bias Between Evaluation Metrics In Classification. In: ICML. (2000) 1087–1094
19. Chawla, N.V., Cieslak, D.A., Hall, L.O., Joshi, A.: Automatically countering imbalance and its empirical relationship to cost. Utility-Based Data Mining: A Special issue of the International Journal Data Mining and Knowledge Discovery (2008)
20. Elkan, C.: The Foundations of Cost-Sensitive Learning. In: International Joint Conference on Artificial Intelligence (IJCAI). (2001) 973–978

21. Dietterich, T.G.: Approximate statistical test for comparing supervised classification learning algorithms. Neural Computation **10**(7) (1998) 1895–1923
22. Kubat, M., Holte, R.C., Matwin, S.: Machine Learning for the Detection of Oil Spills in Satellite Radar Images. Machine Learning **30** (1998) 195–215
23. Radivojac, P., Chawla, N.V., Dunker, A.K., Obradovic, Z.: Classification and knowledge discovery in protein databases. Journal of Biomedical Informatics **37** (2004) 224–239
24. Chang, C., Lin, C.: LIBSVM: a library for support vector machines (2001) Software available at http://www.csie.ntu.edu.tw/∼cjlin/libsvm.
25. Asuncion, A., Newman, D.: UCI Machine Learning Repository (2007)

A Appendix: Wrapper selected sampling levels

Table 5. Optimized $(Undersample, SMOTE)$ levels for each respective splitting metric, along with relative ranking for sampling level among all classifiers. The noted undersample level reflects the percentage of negative class examples removed while the SMOTE level represents the percent of synthetic examples added to the training data relative to the original positive class size.

Dataset	C4.5	DKM	CART	HDDT
Boundary	(76,320) (1,1)	(44,160) (4,4)	(54,210) (3,3)	(59,230) (2,2)
Breast-w	(7,270) (4,4)	(26,350) (2.5,2.5)	(28,280) (1,1)	(26,350) (2.5,2.5)
Calmodoulin	(35,210) (2,1)	(29,90) (4,3)	(57,170) (1,2)	(33,40) (3,4)
E-State	(44,280) (2,1)	(41,120) (3.5,3.5)	(57,250) (1,2)	(41,120) (3.5,3.5)
Forest Cover	(11,440) (3,1.5)	(6,420) (3.5,3.5)	(13,440) (1,1.5)	(6,420) (3.5,3.5)
Fourclass	(14,270) (2.5,3)	(14,320) (2.5,1.5)	(14,100) (2.5,4)	(14,320) (2.5,1.5)
German.numer	(41,250) (1,4)	(39,280) (2.5,3)	(32,330) (4,1)	(39,300) (2.5,2)
Letter	(45,200) (2,4)	(38,210) (3.5,2.5)	(54,410) (1,1)	(38,210) (3.5,2.5)
Mammography	(32,370) (4,2)	(62,360) (1.5,3.5)	(58,420) (3,1)	(62,360) (1.5,3.5)
Oil	(44,330) (1,1)	(39,280) (3,3)	(38,180) (4,4)	(41,310) (2,2)
Page	(9,350) (4,4)	(19,370) (2,2.5)	(19,430) (2,1)	(19,370) (2,2.5)
Pendigits	(38,420) (1,1.5)	(33,320) (2.5,3.5)	(28,420) (4,1.5)	(33,320) (2.5,3.5)
Phoneme	(5,340) (2,4)	(4,370) (2.5,1.5)	(9,350) (1,3)	(4,370) (2.5,1.5)
PhosS	(64,180) (1,1)	(21,0) (3,3.5)	(32,50) (2,2)	(20,0) (4,3.5)
Pima	(15,180) (4,4)	(38,220) (2.5,3)	(40,360) (1,1)	(38,270) (2.5,2)
Satimage	(32,280) (2,2)	(22,240) (3.5,3.5)	(56,370) (1,1)	(22,240) (3.5,3.5)
Segment	(34,260) (1,1)	(23,140) (3.5,3.5)	(30,250) (2,2)	(23,140) (3.5,3.5)
Splice	(10,80) (4,3)	(12,100) (3,1.5)	(25,20) (1,4)	(13,100) (2,1.5)
SVMguide1	(18,300) (1,1)	(12,210) (2.5,2.5)	(5,140) (4,4)	(12,210) (2.5,2.5)
Avg. Undersample Rank	2.24	2.97	2.08	2.76
Avg. SMOTE Rank	2.32	2.89	2.11	2.68