Compact Ensemble Trees for Imbalanced Data

Yubin Park and Joydeep Ghosh

Department of Electrical and Computer Engineering
The University of Texas at Austin, Austin, TX-78712, USA

Abstract. This paper introduces a novel splitting criterion parametrized by a scalar 'α' to build a class-imbalance resistant ensemble of decision trees. The proposed splitting criterion generalizes information gain in C4.5, and its extended form encompasses Gini(CART) and DKM splitting criteria as well. Each decision tree in the ensemble is based on a different splitting criterion enforced by a distinct α. The resultant ensemble, when compared with other ensemble methods, exhibits improved performance over a variety of imbalanced datasets even with small numbers of trees.

1 Introduction

Imbalanced datasets are pervasive in real-world applications, including fraud detection, risk management, text classification, medical diagnosis etc. Despite their frequent occurrence and huge impact in day to day applications, many standard machine learning algorithms fail to address this problem properly since they assume either balanced class distributions or equal misclassification costs [10]. There have been various approaches proposed to deal with imbalanced classes, including: over/undersampling [13], [17], SMOTE (synthetic minority oversampling technique), cost-sensitive [15], modified kernel-based, and active learning methods [1], [8].

Several authors have tried to theoretically address the nature of the class imbalance problem [3], [11], [18]. Their results suggest that the degree of imbalance is not the only factor hindering the learning process [10]. Rather, the difficulties reside with various other factors such as overlapping classes, lack of representative data, small disjuncts etc, that get amplified when the distribution of classes is imbalanced. In this paper, we approach the imbalanced learning problem by combining multiple decision trees. If these different "base" classifiers can focus on different features of the data and handle complex objectives collaboratively, then an ensemble of such trees can perform better for datasets with class imbalance.

Breiman had observed that the most challenging classification problem is how to increase *simplicity* and *understanding* without losing *accuracy* [16]. Also, it has been shown that a small variety of strong learning algorithms are typically more effective than using a large number of *dumbed-down* models [9]. So a second goal is to build robust, imbalance-resistant ensembles using only a few classifiers.

While many ensemble trees induce diversity using random selection of features or data points, this paper proposes a novel splitting criterion parametrized by a scalar α. By varying α, we get dissimilar decision trees in the ensemble. This new approach results in ensembles that are reasonably simple yet accurate over a range of class imbalances.

We briefly summarize the main contributions of this paper here:

1. We introduce a new decision tree algorithm using α-divergence. A generalized tree induction formula is proposed, which includes Gini, DKM, and C4.5 splitting criteria as special cases.
2. We propose a systematic ensemble algorithm using a set of α-Trees covering a range of α. The ensemble shows consistent performance across a range of imbalance degrees. The number of classifiers needed in the method is far less than Random Forest or other ensembles for a comparable performance level.

Related Work Several approaches try to tackle the imbalanced learning problem by oversampling or generating synthetic data points in order to balance the class distributions [14]. An alternative is to employ cost-sensitive methods that impose different misclassification costs. Even though these methods have shown good results, their performance depends on heuristics that need to be tuned to the degree of imbalance.

Ensemble methods generally outperform single classifiers [5], and decision trees are popular choices for the base classifiers in an ensemble [2]. In recent years, the Random Forest has been modified to incorporate sampling techniques and cost matrices to handle class-imbalance [6]. Though this modified Random Forest shows superior performance over other imbalance-resistant classifiers, its complexity increases too.

Some of the earlier works such as [4] by L. Breiman investigate various splitting criteria - Gini impurity, Shannon entropy and twoing in detail. Dietterich et. al. [7] showed that the performance of a tree can be influenced by its splitting criteria and proposed a criterion called DKM which results in lower error bounds based on the Weak Hypothesis Assumption. Karakos et. al. proposed Jensen-Rényi divergence parametrized by a scalar α as a splitting criterion [12], but the determination of the"best" α was based on heuristics.

This paper applies a novel splitting criterion(α-divergence) to ensemble methods to solve the class imbalance problem with a small number of base trees. Decision trees based on distinct α values possess different properties, which in turn increases diversity in the ensemble.

2 Preliminaries

α-Divergence Decision tree algorithms try to determine the best split based on a certain criterion. However, the "best" split usually depends on the characteristics of the problem. For example, for some datasets we might want 'low precision'-'high recall' results and for some the other way around. For this to be resolved it's better to have a criterion that can be adapted by easy manipulation.

Our metric, α-divergence, which generalizes KL-divergence [19], easily achieves this feat.

$$D_\alpha(p||q) = \frac{\int_x \alpha p(x) + (1-\alpha)q(x) - p(x)^\alpha q(x)^{1-\alpha} dx}{\alpha(1-\alpha)} \tag{1}$$

where p, q are any two probability distributions and α is a real number. Some special cases are:

$$D_{\frac{1}{2}}(p||q) = 2\int_x (\sqrt{p(x)} - \sqrt{q(x)})^2 dx \tag{2}$$

$$\lim_{\alpha\to 1} D_\alpha(p||q) = KL(p||q) \tag{3}$$

$$D_2(p||q) = \frac{1}{2}\int_x \frac{(p(x)-q(x))^2}{q(x)} dx \tag{4}$$

Equation (2) is Hellinger distance, and equation (3) is KL-divergence. α-Divergence is always positive and is 0 if and only if $p = q$. This enables α-divergence to be used as a (dis)similarity measure between two distributions.

Splitting Criterion using α-Divergence The splitting criterion function of C4.5 can be written using α-divergence as :

$$I(X;Y) = \lim_{\alpha\to 1} D_\alpha(p(x,y)||p(x)p(y)) = KL(p(x,y)||p(x)p(y)) \tag{5}$$

where $p(x, y)$ is a joint distribution of a feature X and the class label Y, and $p(x)$ and $p(y)$ are marginal distributions. To maintain consistency with the C4.5 algorithm, a new splitting criterion function is proposed as follows:

Definition 1. *α-Divergence splitting criterion is $D_\alpha(p(x,y)||p(x)p(y))$, where $0 < \alpha < 2$.*

Note that $\alpha = 1$ gives the information gain in C4.5.

Using this splitting criterion, a splitting feature is selected, which gives the maximum α-divergence splitting criterion.

Constructing an α-Tree Using the proposed decision criterion the decision tree induction follows in **algorithm 1**. Let us call this new tree as ***α*-Tree**. In algorithm 1 'Classify' can be either 'majority voting' or 'probability approximation' depending on the purpose of the problem. This paper uses 'probability approximation' as 'majority voting' might cause overfitting for the imbalanced data. The effect of varying α will be discussed in Section 4.2.

Algorithm 1 Grow Single α-Tree

Input: Training Data (features $X_1, X_2, ..., X_n$ and class Y), $\alpha \in (0, 2)$
Output: α-Tree
Select the best feature X^*, which gives the maximum **α-divergence criterion**
if (no such X^*) or (number of data points < cut-off size) **then**
  return Classify(Training Data)
else
  partition the training data into m subsets, based on the value of X^*
  for for $i = 1$ to m **do**
    ith child = Grow Single α-Tree (ith partitioned data, α)
  end for
end if

3 Properties of α-Divergence Criterion

Properties of α-Divergence If both $p(x, y)$ and $p(x)p(y)$ are properly defined probability distributions, then the above α-divergence becomes:

$$D_\alpha(p(x, y)||p(x)p(y)) = E_X[D_\alpha(p(y|x)||p(y))]. \tag{6}$$

Consider two Bernoulli distributions, $p(x)$ and $q(x)$ having the probability of success θ_p, θ_q respectively, where $0 < \theta_p, \theta_q < 1/2$. Then the α-divergence from $p(x)$ to $q(x)$ and its 3rd order Taylor expansion w.r.t. θ_p is:

$$D_\alpha(p||q) = \frac{1 - {\theta_p}^\alpha {\theta_q}^{1-\alpha} - (1-\theta_p)^\alpha (1-\theta_q)^{1-\alpha}}{\alpha(1-\alpha)} \tag{7}$$

$$\approx A(\theta_p - \theta_q)^2 + B(\alpha - 2)(\theta_p - \theta_q)^3 \tag{8}$$

where $A = \frac{1}{2}(\frac{1}{\theta_q} + \frac{1}{1-\theta_q})$, $B = \frac{1}{6}(\frac{1}{{\theta_q}^2} - \frac{1}{(1-\theta_q)^2})$ and $A, B > 0$. Then, given $0 < \alpha < 2$ and $\theta_p > \theta_q$, the 3rd order term in equation (8) is negative. So by increasing α the divergence from p to q increases. On the other hand if $\theta_p < \theta_q$ the 3rd order term in equation (8) is positive and increasing α decreases the divergence. This observation motivates proposition 1 below. Later we describe proposition 2 and its corollary 1.

Proposition 1. *Assume that we are given Bernoulli distributions $p(x)$, $q(x)$ as above and $\alpha \in (0, 2)$. Given $\theta_q < 1/2$, $\exists\ \epsilon > 0$ s.t. $D_\alpha(p||q)$ is a monotonic 'increasing' function of α where $\theta_p \in (\theta_q, \theta_q + \epsilon)$, and $\exists\ \epsilon' > 0$ s.t. $D_\alpha(p||q)$ is a monotonic 'decreasing' function of α where $\theta_p \in (\theta_q - \epsilon', \theta_q)$. (Proof. This follows from equation (8).)*

Proposition 2. *$D_\alpha(p||q)$ is convex w.r.t. θ_p. (Proof. 2nd derivative of equation (7) w.r.t θ_p is positive.)*

Corollary 1. *Given binary distributions, $p(x)$, $q(x)$, $r(x)$, where $0 < \theta_p < \theta_q < \theta_r < 1$, $D_\alpha(q||p) < D_\alpha(r||p)$ and $D_\alpha(q||r) < D_\alpha(p||r)$. (Proof. Since $D_\alpha(s(x)||t(x)) \geq 0$ and is equal if and only if $s(x) = t(x)$, using proposition 2, corollary 1 directly follows.)*

Effect of varying α Coming back to our original problem, let us assume that we have a binary classification problem whose positive class ratio is θ_c where $0 < \theta_c \ll 1/2$ (imbalanced class). After a split, the training data is divided into two subsets: one with higher ($> \theta_c$) and the other with lower ($< \theta_c$) positive class ratio. Let us call the subset with higher positive class ratio as *positive*, and the other as *negative subset*. Without loss of generality, suppose we have binary features $X_1, X_2, ..., X_n$ and $p(y = 1|x_i = 0) < p(y = 1) < p(y = 1|x_i = 1)$ and $p(x_i) \approx p(x_j)$ for any i, j. From equation (6) the α-divergence criterion becomes:

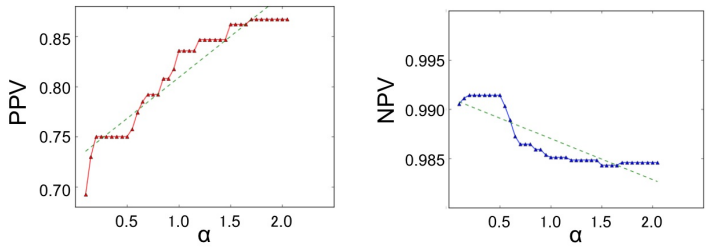
$$p(x_i = 1)D_\alpha(p(y|x_i = 1)||p(y)) + p(x_i = 0)D_\alpha(p(y|x_i = 0)||p(y)) \tag{9}$$

where $1 \leq i \leq n$. From 'proposition 1' we observe that increase in α increases $D_\alpha(p(y|x_i = 1)||p(y))$ and decreases $D_\alpha(p(y|x_i = 0)||p(y))$ (lower-bounded by 0).

$$(9) \approx p(x_i = 1)D_\alpha(p(y|x_i = 1)||p(y)) + const. \tag{10}$$

From 'corollary 1', increasing α shifts our focus to high $p(y = 1|x_i = 1)$. In other words, increasing α results in the splitting feature having higher $p(y = 1|x_i = 1)$, *positive predictive value* (PPV) or *precision*. On the other hand reducing α results in lower $D_\alpha(p(y|x_i = 1)||p(y))$ and higher $D_\alpha(p(y|x_i = 0)||p(y))$. As a result, reducing α gives higher $p(y = 0|x_i = 0)$, *negative predictive value* (NPV) for the splitting features.

The effect of varying α appears clearly with an experiment using real datasets. For each α value in the range of $(0, 2)$, α-Tree was built based on 'sick' dataset from UCI thyroid dataset. α-Trees were grown until '3rd level depth', as fully-grown trees deviate from the above property. Note that this analysis is based on a single split, not on a fully grown tree. As the tree grows, a subset of data on each node might not follow the imbalanced data assumption. Moreover, the performance of a fully grown tree is affected by not only 'α', but also other heuristics like 'cut-off size'. 5-fold cross validation is used to measure each performance. Averaged PPV and NPV over 5-cv are plotted in Figure 1.



(a) PPV vs. α (b) NPV vs. α

Fig. 1: Effect of varying α. Dotted lines are linearly regressed lines.

By varying the value of α we can control the selection of splitting features. This is a crucial factor in increasing 'diversity' among decision trees. The greedy nature of decision trees means that even a small change in α may result in a substantially different tree.

Note that the above analysis is based on Taylor expansion of α-divergence that holds true when $p(y|x_i) \approx p(y)$, which is the case when datasets are imbalanced. This property may not hold if $p(y|x_i)$ differs a lot from $p(y)$.

Connection to DKM and CART The family of α-divergence naturally includes C4.5's splitting criterion. But the connection to DKM and CART is not that obvious. To see the relation between α-divergence and the splitting functions of DKM and CART, we extend the definition of α-divergence (equation (6)) as follows:

Definition 2. *Extended α-divergence is defined as $E_X[D_\alpha(p(y|x)||q(y))]$ where $q(y)$ is any arbitrary probability distribution.*

Definition 2 is defined by replacing $p(y)$ with any arbitrary distribution $q(y)$, which serves as a reference distribution. The connection to DKM and CART is summarized in the following two propositions:

Proposition 3. *Given a binary classification problem, if $\alpha = 2$ and $q(y) = (\frac{1}{2}, \frac{1}{2})$ then the extended α-divergence splitting criterion gives the same splitting feature as the Gini impurity criterion in CART. (Proof. See Appendix A.)*

Proposition 4. *Given a binary classification problem, if $\alpha = \frac{1}{2}$ and $q(y) = p(\bar{y}|x)$ then the extended α-divergence splitting criterion gives the same splitting feature as the DKM criterion. (Proof. See Appendix A.)*

CART implicitly assumes a balanced reference distribution while DKM adaptively changes its reference distribution for each feature. This explains why CART generally performs poorly on imbalanced datasets and DKM provides a more skew-insensitive decision tree.

4 Bootstrap Ensemble of α-Trees

In this section, we propose the algorithm for creating an ensemble of α-Trees. The **BEAT** (**B**ootstrap **E**nsemble of **A**lpha **T**rees) algorithm for an ensemble of k trees is illustrated in Algorithm 2. Observe that the parameters (a, b) for Beta distribution and the number of trees are design choices. The parameters (a, b) can be chosen using a validation set.

BEAT uses Bootstrapping when making its base classifiers. Like other Bagging methods, BEAT exhibits better performance as the number of trees in BEAT increases. The test errors of BEAT and Bagged-C4.5 ($C4.5^B$) are shown in Figure 2 (a). The experiment is performed based on 'glass' dataset from UCI repository. The 'headlamps' class in the dataset is set as positive class, and the

Algorithm 2 Bootstrap Ensemble of α-Trees (**BEAT**)

Input: Training Data D (features $X_1, X_2, ..., X_n$ and class Y) and parameters (a, b).
for for $i = 1$ to k **do**
  Sample $\alpha/2 \sim Beta(a, b)$.
  Sample D_i from D with replacement (Bootstrapping).
  Build an α-Tree C_i from D_i using **algorithm 1**.
end for
for for each test record $t \in T$ **do**
  $C^*(t) = Avg(C_1(t), C_2(t), ..., C_k(t))$
end for

other classes are set as negative class (13.5% positive class ratio). 5×2 cross validation is used. The performance of BEAT is comparable with $C4.5^B$.

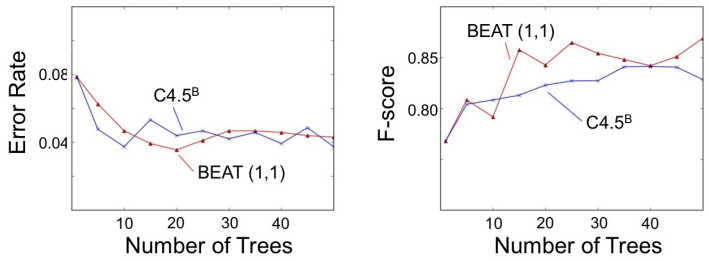
As the value of α affect the performance of α-Tree, the parameters (a, b), which determine the distribution of α, change the performance of BEAT. Misclassification rate generally doesn't capture the performance on imbalanced datasets. Although the misclassification rate of BEAT doesn't vary much from $C4.5^B$, the improvement can be seen apparently in 'precision' and 'recall', which are crucial when dealing with imbalanced datasets. This property is based on the observation in Section 3, but the exact relationship between the parameters and the performance is usually data-dependent. Figure 2 (b) shows the averaged 'f-score' result based on the same 'glass' dataset. Unlike the error rate result, the f-scores of BEAT and $C4.5^B$ show clear distinction. Moreover the resultant average ROC curves of BEAT (Figure 2 (c), (d)) changed as the parameters (a, b) change. The ability to capture different ROC curves allows great flexibility on choosing different decision thresholds.

Experimental Evaluation All the datasets used in this paper are from the UCI Repository. Datasets with multiple classes are converted into 2-class problems. 5×2 cross validation instead of 10-fold cross validation is used due to highly imbalanced data. Aside from the stated modifications each dataset is used "as is".

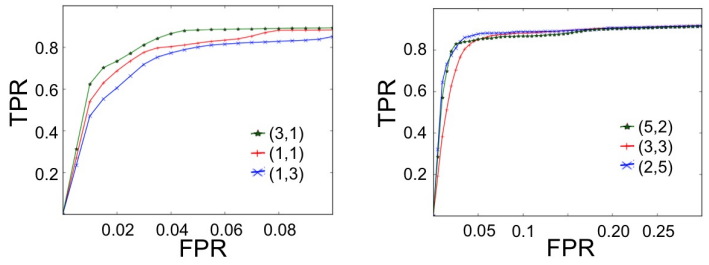
A comparative evaluation of BEAT with C4.5, $C4.5^B$, and Balanced Random Forest (BRF) [6], was performed. All trees are binary/fully grown and 30 base trees are used. No prunning is applied, and for features having more than 2 categories, dummy coding scheme is used to build a binary tree. Table 1 reports the average f-score over 5×2 cross validation. For BRF, the number of random attributes are fixed to 2 ($m = 2$). BRF generally needs more number of base classifiers to perform stably.

5 Concluding Remarks

In this paper, we presented the BEAT approach incorporating a novel decision criterion parametrized by α. Experimental results show that BEAT is stronger



(a) Error rate vs. Number of Trees (b) F-score vs. Number of Trees



(c) ROC curves (d) ROC curves

Fig. 2: Properties of BEAT. ROC curves are measured using 30 α-Trees.

and more robust for imbalanced data, compared to other tree based ensemble methods.

Even though our algorithm gives consistent results for various cases, a joint optimization hasn't been achieved yet with respect to the parameters (a, b) and the number of trees. Moreover, the extended α-divergence criterion needs to be further investigated as well.

Acknowledgements This research was supported by SK Telecom, IC^2 Institute, NHARP and NSF IIS-1016614. We are also grateful to Abhimanu Kumar for some useful discussions.

References

1. Akbani, R., Kwek, S., Japkowicz, N.: Applying support vector machines to imbalanced datasets. In: Proceedings of the 15th European Conference on Machine Learning (2004)
2. Banfield, R.E., Hall, L.O., Bowyer, K.W., Kegelmeyer, W.P.: A comparison of decision tree ensemble creation techniques. In: IEEE Transactions on Pattern Analysis and Machine Intelligence (2006)

Table 1: F-score results on real datasets. Parameters (a, b) of BEAT are indicated in the parenthesis. The test errors are shown in the parenthesis below f-scores. Although BRF has the highest f-score in 'wpbc' dataset, BRF records the highest error rate among ensemble trees. $C4.5^B$ and BEAT show comparable test errors, but BEAT outperforms in f-scores.

Dataset	C4.5	$C4.5^B$	BRF	BEAT(1,1)	BEAT(1,3)	BEAT(3,1)
glass	0.756	0.846	nan	**0.868**	0.84	0.865
	(6.0%)	(3.7%)	(4.3%)	(3.7%)	(4.2%)	(3.6%)
allhypo	0.901	0.956	0.644	0.958	**0.96**	0.956
(Thyroid)	(1.5%)	(0.69%)	(7.7%)	(0.67%)	(0.63%)	(0.69%)
allhyper	0.57	0.711	0.434	0.715	0.689	**0.728**
(Thyroid)	(2.2%)	(1.5%)	(6.6%)	(1.4%)	(1.5%)	(1.4%)
sick	0.826	0.871	0.580	0.866	**0.876**	0.866
(Thyroid)	(2.1%)	(1.5%)	(8.5%)	(1.6%)	(1.5%)	(1.6%)
allrep	0.74	0.869	0.413	0.876	0.870	**0.884**
(Thyroid)	(1.6%)	(0.83%)	(8.8%)	(0.84%)	(0.82%)	(0.76%)
wpbc	0.293	0.387	*0.434*	**0.425**	0.355	0.315
(Breast Cancer)	(33.7%)	(22.3%)	*(31.7%)*	(22.3%)	(22%)	(23.6%)
page blocks	0.791	0.860	0.746	0.863	**0.865**	0.858
	(4.3%)	(2.7%)	(6.8%)	(2.7%)	(2.7%)	(2.9%)

3. Batista, G.E., Prati, R.C., Monard, M.C.: A study of the behavior of several methods for balancing machine learning training data. In: ACM SIGKDD Explorations Newsletter. vol. 6, pp. 20–29 (2004)
4. Breiman, L.: Technical note: Some properties of splitting criteria. In: Machine Learning. vol. 24, pp. 41–47 (1996)
5. Chawla, N.V.: Many are better than one: Improving probabilistic estimates from decision trees. In: Machine Learning Challenges. pp. 41–55 (2006)
6. Chen, C., Liaw, A., Breiman, L.: Using random forest to learn imbalanced data. Tech. rep., Dept. of Statistics, U.C. Berkeley (2004)
7. Dietterich, T., Kearns, M., Mansour, Y.: Applying the weak learning framework to understand and improve c4.5. In: Proceedings of the Thirteenth International Conference on Machine Learning. pp. 96–104 (1996)
8. Ertekin, S., Huang, J., Giles, C.L.: Learning on the border: Active learning in imbalanced data classification. In: Proceedings of the 30th annual international ACM SIGIR conference. pp. 823–824 (2007)
9. Gashler, M., Giraud-Carrier, C., Martinez, T.: Decision tree ensemble: Small heterogeneous is better than large homogeneous. In: The 7th International Conference on Machine Learning and Applications. pp. 900–905 (2008)
10. He, H., Garcia, E.A.: Learning from imbalanced data. In: IEEE Transactions on Knowledge and Data Engineering. vol. 21 (2009)
11. Japkowicz, N., Stephen, S.: The class imbalance problem: A systematic study. In: Intelligent Data Analysis. vol. 6, pp. 429–449 (2002)
12. Karakos, D., Eisner, J., Khudanpur, S., Priebe, C.E.: Cross-instance tuning of unsupervised document clustering algorithms. In: Proceedings of NAACL HLT. pp. 252–259 (2007)

13. Laurikkala, J.: Improving identification of difficult small classes by blancing class distribution. In: Proceedings of the 8th Conference of AI in Medicine in Europe: Artificial Intelligence Medicine. pp. 63–66 (2001)
14. Liu, A., Martin, C., Cour, B.L., Ghosh, J.: Effects of oversampling versus cost-sensitive learning for bayesian and svm classifiers. In: Annals of Information Systems. vol. 8, pp. 159–192 (2010)
15. McCarthy, K., Zarbar, B., weiss, G.: Does cost-sensitive learning beat sampling for classifying rare classes? In: Proceedings of International Workshop Utility-Based Data Mining. pp. 69–77 (2005)
16. Sharkey, A.J. (ed.): Combining Artificial Neural Nets: Ensemble and Modular Multi-Net Systems. Springer (1999)
17. Weiss, G., Provost, F.: The effect of class distribution on classifier learning: An empirical study. Tech. rep., Dept. of Computer Science, Rutgers University (2001)
18. Weiss, G., Provost, F.: Learning when training data are costly: The effect of class distribution on tree induction. In: Journal of Artificial Intelligence Research. vol. 19, pp. 315–254 (2003)
19. Zhu, H., Rohwer, R.: Information geometric measurements of generalization. Tech. Rep. 4350, Aston University (1995)

A CART, DKM and α-Divergence

Gini Assume a binary classification problem, $y \in \{0, 1\}$ and binary feature $x \in \{0, 1\}$. Since for choosing a best feature x the distribution of y is fixed, we can derive the following equation:

$$Gini = \sum_y p(y)(1 - p(y)) - \sum_x p(x) \sum_y p(y|x)(1 - p(y|x)). \tag{11}$$

$$= E_X[\frac{1}{2} - \sum_y p(y|x)(1 - p(y|x))] + const \tag{12}$$

$$\propto E_X[D_2(p(y|x)||q(y))] \tag{13}$$

where $q(y) = (\frac{1}{2}, \frac{1}{2})$. Equation (13) follows from equation (4). Linear relation between the Gini splitting formula and α-divergence completes the proof.

DKM Assuming the similar settings as in Appendix A, the splitting criterion function of DKM is:

$$DKM = \prod_y \sqrt{p(y)} - \prod_x p(x) \prod_y \sqrt{p(y|x)} \tag{14}$$

$$= E_X[\frac{1}{2} - \prod_y \sqrt{p(y|x)}] + const \tag{15}$$

$$\propto E_X[D_{\frac{1}{2}}(p(y|x)||q(y))] \tag{16}$$

where $q(y) = p(\bar{y}|x)$. Equation (16) follows from equation (2). Linear relation between the DKM formula and α-divergence completes the proof.