UNSUPERVISED CLASSIFICATION VIA DECISION TREES: AN INFORMATION-THEORETIC PERSPECTIVE

Damianos Karakos, Sanjeev Khudanpur
Jason Eisner
Center for Language and Speech Processing
Johns Hopkins University
{*damianos, sanjeev, eisner*}*@jhu.edu*

Carey E. Priebe
Dept. of Applied Mathematics and Statistics
Johns Hopkins University
cep@jhu.edu

ABSTRACT

Integrated Sensing and Processing Decision Trees (ISPDTs) were introduced in [1] as a tool for supervised classification of high-dimensional data. In this paper, we consider the problem of *unsupervised* classification, through a recursive construction of ISPDTs, where at each internal node the data (i) are split into clusters, and (ii) are transformed independently of other clusters, guided by some optimization objective. We show that the maximization of information-theoretic quantities such as mutual information and α-divergences is theoretically justified for growing ISPDTs, assuming that each data point is generated by a finite-memory random process given the class label. Furthermore, we present heuristics that perform the maximization in a greedy manner, and we demonstrate their effectiveness with empirical results from multispectral imaging.

1. INTRODUCTION

In unsupervised classification, no statistics of the data jointly with their class-labels are known, so the goal is to group the objects into *clusters* based only on their observable features, such that each cluster contains objects that share some important properties. In some cases, there may be a notion of a "true" class-label of each object that has simply not been provided; it may then be appropriate to view the class-label of each object as a *latent variable*, and to evaluate the performance of a clustering scheme by a *post hoc* assignment of the class-labels to (a subset of) objects in each resulting cluster. In other cases, there may be no natural notion of "true" class-labels; the efficacy of the clustering scheme is often measured in such cases by the economy in *description length* attained by a two-step description of the objects by first describing the attributes common to the clusters and then describing the differential attributes of each object within the cluster. k-Means Clustering and Mixture Modeling using the Expectation Maximization (EM) Algorithm [2, 3] are examples of techniques used for unsupervised classification. Furthermore, a common approach in classification is to map the "sparse" high-dimensional attributes of objects into a "dense" low-dimensional space, and carry out the clustering in this new space. One example of such techniques is Multidimensional Scaling, which maps a set of abstract objects, with given pairwise "distances," to points in a Euclidean space in such a way that all pairwise distances are nearly preserved. This allows the use of clustering algorithms which are known to be efficient in Euclidean space, e.g., model-based clustering [3].

In this paper, we investigate the problem of unsupervised classification using Integrating Sensing and Processing Decision Trees (ISPDTs) [1]. ISPDTs (also called Iterative Denoising Trees) grow in a greedy manner, successively transforming and splitting each node according to some local goodness criterion. They are provably optimal (i.e., they achieve the Bayes-optimal misclassification rate) in some bandwidth or complexity-constrained situations [1, 4]. Moreover, they model adaptive sensors by providing different "looks" at a scene, after a number (but not all) of features or data has been processed. In short, the following steps are performed in an ISPDT:

- Beginning with the whole data collection at the root, each node represents a subset of the data of its parent. The data in each node are transformed through a projection into a lower dimensional space, and partitioned into two clusters, according to an optimization criterion (for example, maximization of the minimum distance between points in different clusters, or maximization of the distance between cluster centroids). In the case where labeled (training) data are present, the projection and clustering may be tuned to maximize the separation between the classes. In our setting, we do not have any labeled data. The two clusters then end up at the two children of the node.
- Under some conditions, a node may not be split further (i.e., it becomes a leaf node). All the data points at each leaf are considered to belong to a single class; that is, classification is done only at the leaves of the ISPDT.

In the following, we will present two information-theoretic criteria for transforming and clustering data in an ISPDT. These two criteria correspond to optimal decision rules in an asymptotic sense (that is, as the number of dimensions of each data point goes to infinity).

The paper is organized as follows. We begin by formulating our problem in Section 2. We show in Section 3 how mutual information and α-divergences may be used as criteria for unsupervised classification via ISPDTs, and we present heuristics for achieving our objective in Section 4. In Section 5 we present experimental results on a classification task in hyperspectral imaging. Finally, concluding remarks appear in Section 6.

2. PROBLEM FORMULATION

Let $\mathcal{A} = \{X^n(1), \ldots, X^n(M)\}$ be a collection of n-dimensional data objects (sequences) that we wish to classify. Each object $X^n(j)$ has a hidden label $Y(j)$, drawn from a finite set $\mathcal{Y}$ (of possibly known cardinality), and $(X_i(j), Y(j))$ are jointly distributed

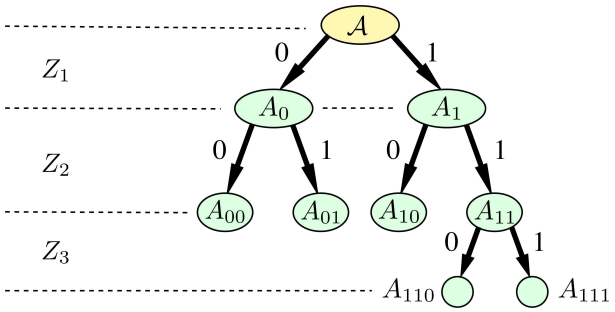

Fig. 1. An ISPDT partitions the data at each node into two sets, according to some optimization criterion. The variable Z_i corresponds to the i-th level of the tree; the vector $\boldsymbol{Z} = g(X^n)$ represents the path from the root to the leaf where X^n is placed.

according to $p_Y \cdot p_{X|Y}$. For simplicity, we can assume that

$$p_{X^n|Y}(x^n|y) = \prod_{i=1}^{n} p_{X|Y}(x_i|y),$$

although similar techniques can be applied to other stochastic processes with memory (e.g., Markov chains).

We now have the following

Problem Formulation: Find a partition $A_1, \ldots, A_m$ of $\mathcal{A}$, such that, with high probability, $X^n(i), X^n(j) \in A_k$, iff $Y(i) = Y(j)$.

For solving the problem, we will use an ISPDT with two different information-theoretic optimization objectives at each node: (i) maximization of a weighted sum of KL-divergences, or (ii) maximization of an α-divergence score. Asymptotically, as $n \to \infty$, these two objectives turn out to correspond to maximization of mutual information and minimization of probability of classification error, respectively.

3. INFORMATION-THEORETIC ASPECTS OF ISPDTS

As we mentioned earlier, ISPDTs are built recursively, through a greedy procedure, such that:

- The object features extracted at each node are *not* necessarily the same as the features extracted at the parent nodes.
- The partitioning at each node is done according to an optimization criterion; this criterion depends on the objects in that node *only*, and not on the splitting of other nodes.

Any denoising tree is associated with a function

$$g : \mathcal{X}^n \to \{0,1\}^*,$$

which takes as input a data sequence, and returns a bit vector that describes the unique leaf to which the object is placed. For example, in Figure 1, all data sequences X^n in leaf A_{110} satisfy $g(X^n) = 110$. As can be easily established, there is a $1-1$ relationship between a denoising tree and a function g (modulo differences in branch labels). Classification is performed only at the leaves through a function $h : \mathcal{L} \to \mathcal{Y}$, where $\mathcal{L} \subseteq \{0,1\}^*$ corresponds to the set of leaves. In the following, we will use the notation $\boldsymbol{Z} = g(X^n)$.[1]

[1]Boldface quantities represent vectors; their dimensionality is determined by the context.

We now explore two approaches for building an ISPDT (or, equivalently, for determining the function g). They both rely on information-theoretic quantities: the mutual information functional, and the α-divergence.

3.1. Greedy Maximization of Mutual Information

Here, our goal is to maximize the mutual information $I(Y; g(X^n))$ with respect to $g(\cdot)$. Fano's inequality [5]

$$I(Y; g(X^n)) \geq (1 - P_e)H(Y) - 1,$$

suggests that, for any classifier implied by h, the probability of error P_e cannot be small if $I(Y; g(X^n))$ is small; this provides the motivation for maximization of the latter.

Through the chain rule of mutual information [5] we have:

$$I(Y; \boldsymbol{Z}) = I(Y; Z_1) + I(Y; Z_2|Z_1) + \ldots + I(Y; Z_m|Z_1, Z_2, \ldots, Z_{m-1}),$$

where m is the maximum leaf depth in an ISPDT (we can take $m = M$, without loss of generality). Each one of the terms above corresponds to node splits of a particular level; for instance, $I(Y; Z_1)$ corresponds to the split at the root, while $I(Y; Z_j|Z_1, \ldots, Z_{j-1})$ corresponds to the splits at level j. Now, to maximize $I(Y; \boldsymbol{Z})$ in a greedy manner, it suffices to maximize iteratively each of the above terms. I.e.,

- First, find the split at the root which maximizes $I(Y; Z_1)$.
- Given the split at the root, find the splits which maximize $I(Y; Z_2|Z_1 = 0)$ and $I(Y; Z_2|Z_1 = 1)$. These two quantities correspond to the two children of the root; the maximization of each one is done through appropriate splitting of the corresponding node/child.
- Iteratively, given the splits at the tree levels $1, \ldots, j-1$, find the splits at level j, such that $I(Y; Z_j|Z_1 = z_1, \ldots, Z_{j-1} = z_{j-1})$ is the maximum possible, for each binary string $(z_1, \ldots, z_{j-1})$. Moreover, a node with path $(z_1, \ldots, z_{j-1})$ is not split any further if $I(Y; Z_j|Z_1 = z_1, \ldots, Z_{j-1} = z_{j-1}) = 0$ (i.e., Y can be determined perfectly from $(z_1, \ldots, z_{j-1})$).

Note that the above procedure is *not* guaranteed to find the maximum of $I(Y; g(X^n))$ with respect to g; a non-greedy procedure could possibly yield a higher value.

3.2. Greedy Minimization of Probability of Error

Here, we assume that each class label is represented by a unique binary sequence that corresponds to a path from the root to a leaf in an ISPDT. In other words, there exists a 1-1 function $L : \mathcal{Y} \to \{0,1\}^*$. Then, a sequence X^n is erroneously classified iff $g(X^n) \neq L(Y)$, where, as before, Y is the true class label of X^n. In other words, there is an error if (at least) one bit of $g(X^n)$ is wrong.

Obviously, in order to minimize the overall error, we have to transform and split the data at each node such that the two sets of each partition do not contain any common classes. Note that the distribution that generates the data of each set is a mixture of distributions corresponding to the classes in the set. Let P_0, P_1 be the two mixtures. Then, the optimum decision rule is the Maximum Aposteriori Probability (MAP). For large n, this can be translated to a KL-divergence decision rule: classify X^n in set A_0 if

$$D(\hat{P}_{X^n}||P_0) < D(\hat{P}_{X^n}||P_1),$$

where $\hat{P}_{X^n}$ is the empirical distribution of X^n, and $D(\cdot||\cdot)$ is the Kullback-Leibler distance between distributions [5]. Then, the exponent of the probability of error (at each node) is given by the Chernoff information [5]:

$$\begin{aligned} C(P_0, P_1) &= -\min_{0\le\alpha\le 1} \log\left(\sum_{x^n} P_0^{\alpha}(x^n) P_1^{\alpha}(x^n)\right) \\ &= \max_{0\le\alpha\le 1} (1-\alpha) D_{\alpha}(P_0||P_1), \end{aligned}$$

where $D_\alpha(P||Q)$ is the α-divergence between distributions P, Q. Moreover, the overall probability of error P_e of the ISPDT is upper-bounded by the sum of the probabilities of error at each node. Hence, the exponent of P_e is the *minimum* of all the exponents. Finally, different class label encodings yield different trees (and hence, different P_e). Therefore, finding the tree that has the maximum probability of error exponent (minimum probability of error, for sufficiently large n) entails computing the following:

$$\begin{aligned} \hat{T} &= \arg\max_{\text{Tree } T} \min_{\text{Internal node } j \text{ in } T} \max_{\alpha_j, P_0(j), P_1(j)} \\ &\quad (1-\alpha_j) D_{\alpha_j}(P_0(j)||P_1(j)), \end{aligned}$$

where $P_0(j), P_1(j)$ are the mixture distributions that result from splitting node j.

In the following, we will see heuristics for building Iterative Denoising Trees that try to optimize the above quantities.

4. HEURISTICS FOR GROWING ISPDTS

As we mentioned above, the conditional distribution that generates the data sequences (given the hidden labels) is unknown. Hence, it is impossible to compute the above information-theoretic quantities precisely. However, for sufficiently large n, where the law of large numbers starts to have an effect, we have the following (the proof appears in [6]).

- The mutual information $I(Y; \boldsymbol{Z})$ can be approximated by

$$\sum_{\text{Internal node } j} \frac{N_0(j)}{M} D(\hat{P}_0(j)||\hat{P}(j)) + \frac{N_1(j)}{M} D(\hat{P}_1(j)||\hat{P}(j)) \quad (1)$$

  where $\hat{P}_0(j), \hat{P}_1(j)$ are the empirical distributions of the data that follow the left or right branch of node j, $\hat{P}(j)$ is the overall empirical distribution of the data in node j, $N_0(j), N_1(j)$ are the number of data points that follow the left or right branch, and M is the total number of data points.

- The exponent of the probability of error of the ISPDT is approximated by

$$\min_{\text{Internal node } j} \max_{\alpha_j} (1-\alpha_j) D_{\alpha_j}(\hat{P}_0(j)||\hat{P}_1(j)), \quad (2)$$

  where $\hat{P}_0(j), \hat{P}_1(j)$ are as above.

Hence, in both cases, we need to estimate the empirical distributions $\hat{P}_0, \hat{P}_1$ at each node. But, in order to overcome the data sparseness problem due to finite n, we need to perform dimensionality reduction before we compute the empirical distributions. In our experiments, the dimensionality reduction is done through Principal Components Analysis (other projection techniques, e.g., Wavelet Packet Decomposition, can be used, too). Then, depending on the particular optimization objective (mutual information or probability of error) we perform the following heuristics at each node:

- **Mutual Information:** We perform a number of linear projections of the sparse empirical distributions of the data objects, to a lower-dimensional simplex (e.g., 3-dimensional). For each projection, we apply Chou's algorithm [7] (which is a variant of K-Means) and we partition the data into two clusters. The distributions $\hat{P}_0, \hat{P}_1$ in (1) are the centroids of these clusters, and the transformation/clustering which is chosen is the one which maximizes the score (1).
- **Probability of error:** As above, we perform a number of low-dimensional projections, and for each projection we perform a number of random splits of the data (e.g., using Chou's algorithm with random starting points). We use the resulting clusters as "seeds" for a maximum-likelihood approach, inspired by Yarowsky's decision lists [8], to further improve the partition: from the seeds, we compute the empirical distributions $\hat{P}_0, \hat{P}_1$ corresponding to each cluster. Then, for each data point, we compute the log-likelihood ratio with respect to $\hat{P}_0, \hat{P}_1$, and we sort the points. This sorted list of log-likelihood ratios will give us an indication of which data points can be discriminated more easily than others. By choosing a proportion of those objects with the highest (or, lowest negative) log-likelihood ratios, we create two new clusters and we re-estimate the empirical distributions; we then re-compute the log-likelihood ratios, and we repeat the whole procedure until convergence. This procedure tries to approximate an optimum decision rule, where the class probabilities are computed "on-the-fly". The resulting $\hat{P}_0, \hat{P}_1$ are then fed into (2), and the "maximizing" α is then computed by an exhaustive search through a discretization of the interval $(0, 1)$. Finally, among all the resultant clusterings (per projection and initial random splits), we choose the one which gives the highest value in (2).

For deciding when to *stop* growing the tree, we perform the following: (i) For the mutual information case, we split the node which yields the highest *increase* in total score (1), until a specified number of leaves has been reached, or the increase in total score is below a threshold τ. (ii) For the probability of error case, we split the node that yields the smallest *decrease* in score (2), until a specified number of leaves has been reached, or the score (2) is below a threshold ν.

5. EMPIRICAL RESULTS FROM MULTISPECTRAL IMAGING

To demonstrate the usefulness of the iterative denoising procedure with information-theoretic optimization criteria, we performed experiments with aerial image data. Each data point corresponds to a multidimensional pixel—each dimension represents a particular frequency band. Furthermore, the spectrum of each pixel is actually a *distribution* of energy over frequencies. Hence, with the appropriate normalization, the spectrum of a data point/pixel plays the role of its "empirical distribution".

The class labels of the pixels correspond to different types of vegetation: runway, pine, scrub and swamp. Raftery's EM-based clustering software *mclust* [3], applied on the original high-dimensional data, yields a 24% misclassification rate.

Using mutual information as the objective optimization score, and setting the target number of leaves to 4, the ISPDT first splits the root into two sets, one of which is pure, contains 100% of

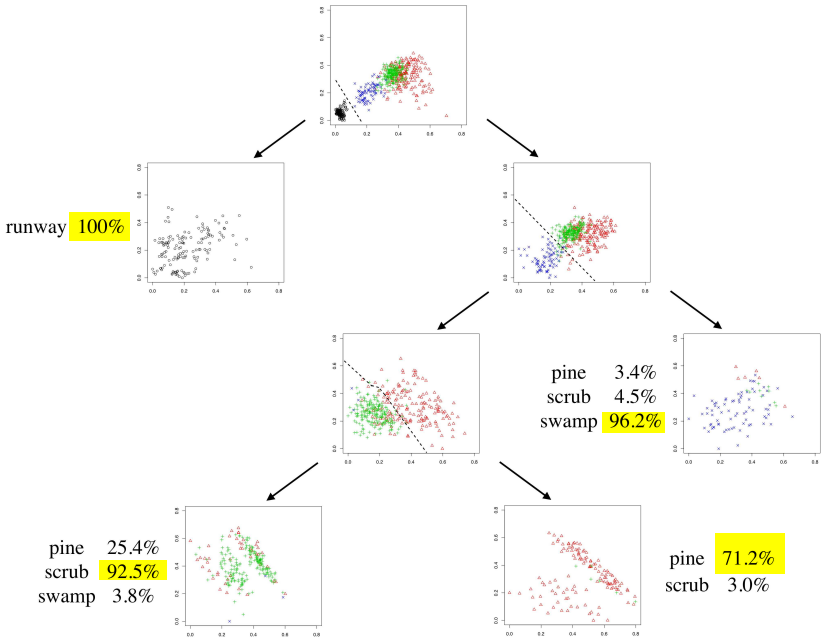

Fig. 2. The ISPDT, when the objective is maximization of mutual information. Depiction of labels is as follows: runway corresponds to circles, pine to triangles, scrub to crosses and swamp to x's. Each node is transformed differently (the corresponding 2 principal components are shown). Cluster boundaries are depicted with dashed lines.

the runway pixels and becomes a leaf. The other node is further split into two sets: one becomes a leaf and contains 96.2% of the swamp pixels (and less than 5% of the other classes), and the other is again divided into: 71.2% of pine and 92.5% of scrub. The total misclassification rate is 11.5%. Figure 2 shows the iterative denoising tree; each node shows the data points under the projection that maximizes the mutual information objective. Finally, using α-divergences as the objective optimization score, the ISPDT has the same structure as above, but slightly different leaf compositions: 100% of runway, 96% of scrub (but with 20% of pine), 87% of swamp, and 76% of scrub, respectively. The total misclassification rate is 11%.

In all cases, each node of the ISPDTs transforms the data *differently* from the other nodes, driven by a local optimization criterion. This transformation corresponds to feature extraction; different features are suppressed (or amplified) by each transformation (*corpus-dependent-feature-extraction* property [4]).

We have also performed experiments with other types of data; in particular, we obtained interesting results in text categorization, where the task is to cluster together documents that have some significant association (they are on the same topic, genre, etc). Preliminary results [6] have shown the ISPDTs are very successful in this task, since different transformations amplify the significance of different words in the documents, thus permitting reasonable discrimination.

6. CONCLUSIONS

In this paper, we presented two criteria for transforming and splitting nodes in an ISPDT for unsupervised classification. The splitting criterion at each node is either a maximization of a weighted sum of KL-divergences, or the maximization of an α-divergence. These criteria correspond, as the data dimensionality goes to infinity, to: (i) the maximization of mutual information between the class label and the path from the root to the leaf in the ISPDT, and (ii) the minimization of the probability of misclassification, respectively. We demonstrated the effectiveness of these techniques using real multispectral imaging data.

7. REFERENCES

[1] C.E.Priebe, D.J. Marchette, and D.M. Healy, "Integrated sensing and processing decision trees," *IEEE Trans. on Pat. Anal. and Mach. Intel.*, vol. 26, no. 6, pp. 699–708, June 2004.

[2] F. Jelinek, *Stat. Methods for Speech Recognition*, MIT Press, 1997.

[3] C. Fraley and A.Raftery, "Mclust: Software for model-based cluster analysis," *Jrnl on Class.*, vol. 16, pp. 297–306, 1999.

[4] C. E. Priebe et al., "Iterative denoising for cross-corpus discovery," in *Proc. 2004 International Symposium on Computational Statistics (COMPSTAT 2004)*, August 2004.

[5] T. Cover and J. Thomas, *Elements of Information Theory*, John Wiley and Sons, 1991.

[6] D. Karakos et al., "Information-theoretic aspects of iterative denoising," In preparation, December 2004.

[7] P. A. Chou, "Optimal partitioning for classification and regression trees," *IEEE Trans. on Pat. Anal. and Mach. Intel.*, vol. 13, no. 4, pp. 340–354, April 1991.

[8] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proc. 33rd Annual Meeting of the Assoc. for Comput. Ling.*, 1995, pp. 189–196.