ITERATIVE DENOISING USING JENSEN-RENYI DIVERGENCES WITH AN APPLICATION TO UNSUPERVISED DOCUMENT CATEGORIZATION

*Damianos Karakos**, *Sanjeev Khudanpur**, *Jason Eisner** *and Carey E. Priebe*†

* Center for Language and Speech Processing
† Department of Applied Mathematics and Statistics
Johns Hopkins University, MD
{*damianos, khudanpur, eisner, cep*}*@jhu.edu*

ABSTRACT

Iterative denoising trees were used by Karakos *et al.* [1] for unsupervised hierarchical clustering. The tree construction involves projecting the data onto low-dimensional spaces, as a means of smoothing their empirical distributions, as well as splitting each node based on an information-theoretic maximization objective. In this paper, we improve upon the work of [1] in two ways: (i) the amount of computation spent searching for a good projection at each node now adapts to the intrinsic dimensionality of the data observed at that node; (ii) the objective at each node is to find a split which maximizes a generalized form of mutual information, the Jensen-Rényi divergence; this is followed by an iterative Naïve Bayes classification. The single parameter α of the Jensen-Rényi divergence is chosen based on the "strapping" methodology [2], which learns a meta-classifer on a related task. Compared with the sequential Information Bottleneck method [3], our procedure produces state-of-the-art results on an unsupervised categorization task of documents from the "20 Newsgroups" dataset.

Index Terms— Unsupervised learning, clustering methods, information theory, text processing

1. INTRODUCTION

Decision trees are best known as tools for performing *supervised* classification. The strategy, roughly speaking, is to recursively divide the space of observations to obtain regions, each with a preponderance of a single label. During training, this procedure amounts to splitting the set of observations into subsets, each corresponding to a node in the tree. The splitting of each node is guided by an optimization objective that depends on knowing the labels of the training observations.

In [1], the authors considered the problem of *unsupervised* classification (clustering) using *integrated sensing and processing decision trees* (ISPDTs). More precisely, ISPDTs were used in [1] to cluster *distributions*[1] —for example, in document clustering, one can regard each document as an empirical distribution over words.[2] Since in unsupervised classification there are no training labels, the optimization objective that guides the growing of an ISPDT has to depend only on the statistics computed from the data points. ISPDTs, also known as iterative denoising trees, differ from regular regression trees in that some form of *dimensionality reduction* (or projection) takes place at each internal node, before splitting [4]. This per-node projection allows feature extraction and *smoothing* of the empirical distributions of the data that end up in each node, independently of other nodes. The smoothed distributions are subsequently used in the computation of the objective functions.

The optimization objectives that were evaluated in [1] for guiding the split of each node were maximization of

- **Mutual information:** This was computed using the overall statistics of the data points in the node, as well as the statistics of the data points of the two clusters, in the form of a weighted sum of Kullback-Leibler (KL) divergences (see expression (1) in [1]). Chou's algorithm [5], which is a version of K-means that has the KL divergence in the role of the "distance" between data points, was used to find a locally optimum split.
- **Log-probability of error:** This was approximated by the Chernoff information [6], which gives an upper bound to the exponent of the probability of error in a binary hypothesis testing problem. It is computed using Rényi's divergence [7] (see expression (2) in [1]), and its maximization results in clusters whose centroids are as far as possible (under this divergence "distance").

Under a fairly weak assumption that the data points are realizations of stationary, ergodic, and finite-memory Markov chains, the above information-theoretic objectives correspond to optimal decision rules, as the lengths of the realizations go to infinity.

In this paper, we extend the ISPDT work of [1] in a number of ways:

- When searching for a good low-dimensional projection at each node, we adaptively decide how many projections to try, in contrast to the fixed-size search space of [1]. Nodes that contain data with high variability (e.g., at the root of the ISPDT) tend to result in larger search spaces than nodes that contain, say, 1-2 tight clusters (typically nodes close to the leaves). This approach reduces the computational complexity of the procedure without adversely affecting the clustering performance.
- We replace mutual information with a more flexible objective function, namely, the Jensen-Rényi divergence [8] (also called α-Jensen difference [9]), which is parameterized by a positive quantity, α. Jensen-Rényi divergences have many desirable properties (continuity, non-negativity, convexity, etc.) and have been applied successfully to image registration problems [8]. The parameter α is tuned in an *unsupervised* manner, through "strap-

[1] Clustering of distributions is useful when building statistical models from sparse data. Combining (clustering) models which are built from seemingly different data fragments (but which are realizations of a common underlying phenomenon) can significantly improve inference. This has been demonstrated in a variety of fields, such as speech recognition, machine translation, computer vision, etc.

[2] Formally, each data point in [1] corresponded to the empirical distribution of a finite sample of a random process.

ping" [2]. Specifically, using a randomly-generated training data set, whose labels do not overlap with the labels of the test set, we build a meta-classifier which learns how to distinguish good from bad clusterings. This meta-classifer is then applied to the collection of classifications that result from various values of α.

Experimental results from a benchmark task of document categorization from the "20 Newsgroups" corpus [10] show that ISPDTs, combined with Jensen-Rényi divergences and "strapping", are competitive with, and in most cases outperform, the sequential information bottleneck procedure [3], which is considered the state-of-the-art in unsupervised document categorization.

The paper is organized as follows. Section 2 contains mathematical preliminaries, as well as a formal definition of the clustering problem. The methodology for growing ISPDTs, using the Jensen-Rényi divergence as the optimization objective, is outlined in Section 3. Finally, the experimental procedure, and results from document categorization, are described in Section 4.

2. PRELIMINARIES AND PROBLEM FORMULATION

For the rest of the paper, we assume that we are given a set of data sequences, $\mathcal{A} = \{X^n(1), \ldots, X^n(N)\}$, where n is the length of each sequence. (The assumption of equal length n is not crucial, since all sequences can be "padded" appropriately). The "hidden" labels associated with these sequences are $Y(1), \ldots, Y(N)$, belonging to some set $\mathcal{Y}$ of known cardinality. Each sequence X^n is generated by some unknown random process $P_{X|Y}$, uniquely determined by its label Y. Assuming that each element of X^n lies in a discrete and finite set $\mathcal{X}$, its empirical distribution (or *type* [11]) is defined as the pmf $\hat{P}_{X^n}(x) = n^{-1}\sum_{i=1}^{n} \mathbf{1}(X_i = x)$, i.e., it results from counting the number of occurences of each symbol[3] x of $\mathcal{X}$ in X^n, and is an approximation of the true process [4]. We now have the following

Problem Formulation: Given $L = |\mathcal{Y}|$, we want to find a partition $A_1, \ldots, A_L$ of the data set $\mathcal{A}$, such that, with as high probability as possible, if $X^n(i), X^n(j)$ are in the same A_k, then $Y(i) = Y(j)$. This is equivalent to finding a function $g : \mathcal{X}^n \to \mathcal{Y}$ such that, with high probability, $g(X^n) = Y$. We avoid the model selection problem by assuming that L is given, and we construct an ISPDT with L leaves. The classification error is computed as in the supervised case, after we perform a post-hoc assignment of labels to the leaves of the ISPDT such that the error is as small as possible.

As mentioned in the introduction and in [1], the greedy construction of ISPDTs requires an optimization objective at each node. In this paper, the objective function which guides the splitting of each ISPDT node is a generalized form of mutual information, the Jensen-Rényi divergence:

$$I_\alpha(X;Y) = H_\alpha(P_X) - \sum_y P_Y(y) H_\alpha(P_{X|Y}(\cdot|y)), \qquad (1)$$

where $H_\alpha(P)$ is the Rényi entropy of order α of pmf P, defined as

$$H_\alpha(P) = \frac{1}{1-\alpha}\log\left(\sum_{x\in\mathcal{X}} P(x)^\alpha\right), \ \alpha \geq 0, \ \alpha \neq 1. \qquad (2)$$

As $\alpha \to 1$, $H_\alpha(P)$ converges (non-obviously) to the usual Shannon entropy $H(P) = -\sum_x P(x)\log P(x)$, and it follows that $I_\alpha(X;Y)$ converges to the usual mutual information $I(X;Y)$. We assume that $\alpha \in (0,1]$ in this paper, in order to guarantee that $H_\alpha(P)$ is a concave function of P (as mentioned in [7], H_α is neither convex nor concave when $\alpha > 1$). This implies that $I_\alpha(X;Y)$ is non-negative (and it is equal to zero if and only if X and Y are independent). The conditional case can be treated similarly to the regular mutual information [6]: $I_\alpha(X;Y|Z) = \sum_z P_Z(z) I_\alpha(X;Y|Z = z)$, where $I_\alpha(X;Y|Z=z)$ is the conditional analogue of (1).

2.1. Motivation for using Jensen-Rényi Divergences

We briefly present our motivation behind using these generalized forms of mutual information in the optimization objective.

- *Rényi entropy is less sensitive to sparseness*
  As can be seen from Figure 4 of [8], Rényi entropy remains strictly larger than Shannon's entropy when its argument approaches the edges of the probability simplex. This means that the effects of extreme sparsity in the data (resulting, for instance, in underestimation of Shannon's entropy, or, in overestimation of regular mutual information) can be dampened, if Shannon's entropy is replaced with Rényi entropy and, consequently, the regular mutual information is replaced with the Jensen-Rényi divergence.
- *Jensen-Rényi divergence introduces a degree of freedom*
  A degree of freedom is added through α of the Jensen-Rényi divergence. Learning to distinguish good classifications from bad ones (which arise when α ranges over the interval $(0,1]$) can be done using an unsupervised procedure, such as "strapping" [2]. Note that there are multiple definitions of mutual information of order α in the literature (see, for example, [7, 9, 12, 13, 8]); we plan to empirically compare them as objective functions for unsupervised classification in future work.

3. GREEDY CONSTRUCTION OF ISPDTS

The main steps for growing ISPDTs are described in [1]. We summarize the algorithm here, including important differences in our current implementation.

Beginning with a trivial 1-node ISPDT (i.e., the root of the tree, which contains all the data points) we repeatedly choose a single leaf node to split, until we have the desired L leaves. In the present paper, we always split whichever leaf yields the maximum increase in the average (per node) Jensen-Rényi divergence (1) between data and children identities (in the case $\alpha = 1$, this average mutual information is equal to equation (1) of [1]). If we were to split a leaf t into children t_0, t_1, the resulting increase would be given by $N(t)/N$ times

$$S(t, t_0, t_1) \triangleq H_\alpha(\hat{P}(t)) - \frac{N(t_0)}{N(t)} H_\alpha(\hat{P}(t_0)) - \frac{N(t_1)}{N(t)} H_\alpha(\hat{P}(t_1)) \qquad (3)$$

[3] We use the term "symbol" loosely; in general, it could be a pair, triple, etc. of symbols from $\mathcal{X}$, or some finite-length sequence from $\mathcal{X}^*$; such approximations are needed in order to capture higher-order effects of the underlying process. In the document categorization application that we consider in this paper, it suffices to deal with the simplest case where the $\hat{P}$ is a pmf over words.

[4] For stationary processes, each $\hat{P}_{X^n}$ converges almost surely to the data-generating distribution $P_{X|Y}$ as $n \to \infty$. Since $\hat{P}_{X^n}$ is a sufficient statistic for estimating the true distribution (and, hence, the hidden label Y), each data point X^n is represented in the ISPDT by (a projected version of) the vector $\hat{P}_{X^n}$. The projection, among other things, plays the role of *smoothing* $\hat{P}_{X^n}$, i.e., of assigning non-zero probabilities to all of its elements. This is crucial when $\hat{P}_{X^n}$ is a very sparse vector, as is the case of text documents, where $|\mathcal{X}|$ is typically of the order of a few thousands; the projection reduces dramatically the vocabulary size.

where $N(j)$ is the number of data points in a node j, while $\hat{P}(j)$ is the centroid of the empirical distributions of those data points. We seek the leaf t and the split t_0, t_1 that maximizes $N(t)/N \times S(t, t_0, t_1)$ by generating many candidate splits, as follows:

- We take the data at t (a collection of high-dimensional empirical distributions) and compute $J(J-1)/2$ different PCA projections onto a 2-dimensional probability simplex. Each projection is determined by a pair of distinct eigenvectors (principal components) from among the first J eigenvectors. J is computed automatically (separately for each t), by locating the "knee" on the scree plot[5]. As mentioned earlier, [1] used the same J across all nodes.
- For each one of the projections computed at the previous step, we seek a binary split of node t into two clusters t_0, t_1 in order to maximize the Jensen-Rényi divergence $S(t, t_0, t_1)$ of (3). Finding the optimum split is a combinatorially hard problem, so we resort to several runs of the exchange algorithm [14]. Each run of the exchange algorithm is initialized with a different random split, and uses $S(t, t_0, t_1)$ as the optimization criterion to choose exchanges.

From all runs of the exchange algorithm over all projections of all leaf nodes t (a triple loop), we choose the split t, t_0, t_1 with the highest $S(t, t_0, t_1)$. Note that we are comparing S values that are smoothed by different projections.

Finally, to further refine the boundary between the chosen t_0 and t_1, we perform an iterative Naïve Bayes classification using the *unsmoothed, high-dimensional* empirical distributions. We initialize the procedure with t_0 and t_1 as initially computed. At each iteration, new cluster centroids are computed, based on the most confident 90% of the data set (i.e., the data points which are as far from the boundary between the two clusters as possible, in terms of likelihood ratio under a Naïve Bayes model). The same procedure was followed in [1] as well.

4. EXPERIMENTS FROM TEXT CATEGORIZATION

We demonstrate the usefulness of the iterative denoising procedure, combined with information-theoretic optimization criteria, with experiments on a benchmark task of unsupervised document categorization from the 20 Newsgroups corpus [10]. This corpus consists of roughly 20,000 news articles, evenly divided among 20 UseNet discussion groups. Document categorization is the task of deciding whether a piece of text belongs to any of a set of prespecified categories. It is a generic text processing task useful in indexing documents for later retrieval, as a stage in natural language processing systems, for content analysis, and in many other roles [15]. We compare ISPDTs with 3 other unsupervised clustering techniques: (i) the sequential Information Bottleneck (sIB) [3] (the Matlab code is available on the web [16]), (ii) the EM-based Gaussian mixtures clustering R package *mclust* [17], and (iii) K-means; we ran K-means with 10 random initializations for the cluster centroids and we averaged the results. Both *mclust* and K-means were fed with a low-dimensional PCA projection of the data; the number of dimensions was determined by locating the "knee" on the scree plot.

To compare our results with the Information Bottleneck method (which has been shown to be very effective for document categorization, matching the classification accuracy of supervised methods), we use the same exact documents as the ones used in [3]: three *Binary* data sets, three *Multi5* data sets, and three *Multi10* data sets, each containing 500 documents[6]. We pre-process the documents as follows (for all methods):

- Excluding the subject line, the header of each abstract is removed.
- Stop-words (such as *a, the, is,* etc.) are removed, and stemming is performed (e.g., common suffixes such as -ing, -er, -ed, etc., are removed). Also, all numbers are collapsed to one symbol, and non-alphanumeric sequences are converted to whitespace. Moreover, as suggested in [18] as an effective method for reducing the dimensionality of the feature space (number of distinct words), all words which occur less than t times in the corpus are removed (a similar thresholding was done in [3], as well). For the IB experiments, we use $t = 2$ (as was done in [3]), while for the ISPDT experiments we use $t = 3$; these choices result in the best performance for each method, respectively.
- For each word in the vocabulary, the term frequency (tf) in each document, and the *inverse document frequency* (idf—the number of documents divided by the number of documents containing the term) are computed. Then, each document is represented by a vector v, whose elements correspond uniquely to the words in the vocabulary, and have values according to the Okapi [19] formula

$$v(w) = \frac{tf(w)}{tf(w) + 0.5 + 1.5\ \left(|d|/|\bar{d}|\right)} \sqrt{\log(idf(w))} \quad (4)$$

  where $|d|$ is the length of document d, $|\bar{d}|$ is the average length of the documents in the collection, and *tf, idf* are the term frequency and inverse document frequency respectively. The denominator in (4) tries to limit the influence of words which appear too many times in a document—usually two or three occurences are enough to signify the importance of a term. In addition, the *idf* component discounts words which are very frequent in the whole collection, and hence do not offer any gain in classification.
- As suggested in the information retrieval literature, each tf-idf vector is normalized so that its norm is equal to one. An L_2 (Euclidean) norm of 1 is used when projecting into the lower-dimensional space through PCA, since PCA is suitable for preserving L_2 distances. But an L_1 norm of 1 is used for the log-likelihood ratio lists (the extra step that re-assigns the data points to the clusters based on the *high-dimensional* information), so that each data point becomes a probability vector.

4.1. Selecting α

Each of the nine datasets is clustered with ten different values of α, namely $\alpha = 0.1, 0.2, \ldots, 1.0$.

It would be customary to choose α based on some *supervised* held-out data. We use a refinement of this idea, based on the "strapping" method of [2], where the supervised held-out data are not used to determine the best α directly, but to learn what good ISPDT clusterings "look like", i.e., on other properties of the clusters. Due to lack of space, we present only a summary of this procedure; the details will appear in a subsequent publication.

[5] To reduce computation time, we attempt to keep J small. At the root of the ISPDT, we take $J = \min(L, 5)$. At other nodes, we choose J such that the mean-squared error introduced by zeroing out the remaining eigenvectors is no greater than what it was at the root.

[6] The *Binary* data sets contain documents from *talk.politics.mideast, talk.politics.misc*, the *Multi5* data sets contain documents from *comp.graphics, rec.motorcycles, rec.sport.baseball, sci.space, talk.politics.mideast*, and the *Multi10* data sets contain documents from *alt.atheism, comp.sys.mac.hardware, misc.forsale, rec.autos, rec.sport.hockey, sci.crypt, sci.electronics, sci.med, sci.space, talk.politics.guns*. In all of these subsets, the documents are evenly divided between the newsgroups.

For each test condition (e.g., *Binary*), we perform the following steps:

(i) We choose 5 held-out supervised datasets, by randomly choosing documents whose labels (newsgroups) do *not* appear in the test set. These classification problems are similar to the test condition but do not give direct information about it. (ii) We cluster each *labeled* data set using unsupervised ISPDTs, for $\alpha = 0.1, 0.2, \ldots, 1.0$. (iii) We evaluate each of these 50 clusterings using the true labels. (iv) We use a supervised ranking SVM [20] to learn how to predict the true ranking of the 50 clusterings based on a number of features: the average cosine of the angle (in the tf-idf space) between each data point and the centroid of its assigned cluster, the average Rényi divergence between the empirical distribution of each data point and the empirical centroid of its assigned cluster, as well as other plausible features which combine α with various measures of "goodness" of clusterings in general (specific details will appear in the full version of the paper). (iv) We apply the SVM model to rank the 10 unsupervised clusterings of the test data, based on the same features, and we pick the highest-ranked clustering.

Set	ISPDT	[3]	sIB	Mclust	K-means
Binary	**5.7%**	8.8%	7.3%	35.9%	37.6%
Multi5	**9.5%**	8.4%	9.5%	22.7%	26.9%
Multi10	**38.5%**	33.0%	40.3%	42.0%	45.5%

Table 1. Average classification errors for the text categorization task of the 20 Newsgroups corpus. ISPDT results in **bold** are at least as good as the corresponding sIB results (4th column).

4.2. Results

Table 1 shows the classification error results, averaged over the 3 data sets of each type (Binary, Multi5, Multi10). The column "[3]" shows the average of the errors reported in [3], while column "sIB" shows the result after applying the *sIB* code on the term-document frequency matrix that was created through our pre-processing. (Note that our pre-processing involves *stemming*, as well as possibly a different stop list from the one in [3][7].)

From these results we conclude: (i) ISPDTs, when combined with Jensen-Rényi divergences, Naïve Bayes classification and "strapping", are competitive with the sequential Information Bottleneck, which is considered the state-of-the-art in unsupervised document categorization; (ii) the widely-used *mclust* and K-means have consistently worse performance than ISPDTs. Moreover, the Jensen-Rényi results are on average better[8] than when $\alpha = 1$.

5. REFERENCES

[1] D. Karakos, S. Khudanpur, J. Eisner, and C.E.Priebe, "Unsupervised classification via decision trees: An information-theoretic perspective," in *Proc. 2005 International Conference on Acoustics, Speech and Signal Processing (ICASSP 2005)*, March 2005.

[2] J. Eisner and D. Karakos, "Bootstrapping without the boot," in *Proc. 2005 Conference on Human Language Technology / Empirical Methods in Natural Language Processing (HLT/EMNLP 2005)*, October 2005.

[3] N. Slonim, N. Friedman, and N. Tishby, "Unsupervised document classification using sequential information maximization," in *Proc. SIGIR'02, 25th ACM Int. Conf. on Research and Development of Inform. Retrieval*, 2002.

[4] 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.

[5] P. A. Chou, "Optimal partitioning for classification and regression trees," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 13, no. 4, pp. 340–354, April 1991.

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

[7] I. Csiszár, "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. on Information Theory*, vol. 41, no. 1, pp. 26–34, January 1995.

[8] Y. He, A. Ben Hamza, and H. Krim, "A generalized divergence measure for robust image registration," *IEEE Trans. on Signal Processing*, vol. 51, no. 5, pp. 1211–1220, May 2003.

[9] A.O. Hero, B. Ma, O. Michel, and J. Gorman, "Alpha-divergence for classification, indexing and retrieval," Technical Report CSPL-328, University of Michigan Ann Arbor, Communications and Signal Processing Laboratory, May 2001.

[10] K. Lang, "Learning to filter netnews," in *Proc. 13th Int. Conf. on Machine Learning*, 1995, pp. 331–339.

[11] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*, New York: Academic, 1981.

[12] A. Ben Hamza and H. Krim, "Jensen-Rényi divergence measure: Theoretical and computational perspectives," in *Proc. IEEE Int. Symp. on Information Theory*, Yokohama, Japan, June 2003.

[13] H. Neemuchwala, A. Hero, and P. Carson, "Feature coincidence trees for registration of ultrasound breast images," in *Proc. IEEE Int. Conf. on Image Processing*, Thesaloniki, Greece, October 2001.

[14] S. Martin, J. Liermann, and H. Ney, "Algorithms for bigram and trigram word clustering," *Speech Communication*, vol. 24, no. 3, pp. 171–192, 1998.

[15] David D. Lewis and Philip J. Hayes, "Guest editorial," *ACM Transactions on Information Systems*, vol. 12, no. 3, pp. 231, July 1994.

[16] N. Slonim, "IBA_1.0: Matlab code for information bottleneck clustering algorithms," Available from http://www.princeton.edu/∼nslonim/IB_Release1.0/IB_Release1_0.tar, 2003.

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

[18] Y. Yang and J. Pedersen, "A comparative study on feature selection in text categorization," in *Intl. Conf. on Machine Learning (ICML-97)*, 1997, pp. 412–420.

[19] S.E. Robertson *et al.*, "Okapi at TREC-3," in *The Third Text Retrieval Conference (TREC-3)*, 1995, pp. 109–126.

[20] T. Joachims, "Optimizing search engines using clickthrough data," in *ACM Conf. on Knowledge Discovery and Data Mining (KDD)*, 2002.

[21] N. Slonim, "Private communication," 2006.

[7] The non-stemmed vocabulary used in [3] is not available [21]; however, we tried to find a document representation that would give us the best possible sIB results, to allow a fair comparison between sIB and ISPDTs. For completeness, we also show the results from [3].

[8] The relative improvement achieved with the Jensen-Rényi divergence, compared to the regular mutual information, is 21.9%, 25.2% and 9.2% for the Binary, Multi5 and Multi10 datasets, respectively.