

methods. The remainder of the paper is organized as follows. We describe the related work in Section 2. The embedding method and uncertainty measure are discussed in Sections 3. Experiments are reported in Section 4, and we conclude the paper in Section 5.

2. RELATED WORK

In the active learning scenario [9, 12], unlabeled data are available and at each iteration an algorithm is able to choose an instance for a user/oracle to label. The objective is of learning the appropriate concept with certain accuracy while incurring the lowest cost. In most of real-world learning problems, the pool-based active learning framework is used, in which there is a large pool U of unlabeled data sampled from a distribution $P(x)$. In each step, the learner is allowed to query one unlabeled data $x \in U$ from the pool and get its label. The simplest and commonly used query strategy is uncertainty sampling, in which an active learner queries the instances about which it is least certain on how to label them. Suppose that we are going to build a predictive model $p_y = P(y|x) (y \in \{1, 2, \dots, m\})$: given the data x from the input space X , we can predict the conditional probability for label y . The key to measure how useful labeling a sample x is to measure the value of the information gained by requesting the unknown label y for each unlabeled sample $x \in U$. A generally used measure of information is to measure the uncertainty of an unknown label by entropy of the posterior class distribution, which is defined as:

$$\text{Entropy}(y; x) = - \sum_{y=1}^m p_y \log p_y \quad (1)$$

where $p_y = P(y = \textcircled{y}|x)$ and y ranges over all possible labels. There are some other measures proposed. Nader et al. find that using variance to measure uncertainty has very similar performance entropy in [10]. They show that selecting the unlabeled data that maximizes the entropy is equivalent to selecting the unlabeled data that maximizes the variance in some condition.

Most of the active learning approaches focus on binary classification. For multi-classification problem, each category is handled independently by a binary active learning algorithm in the traditional methods [13, 6]. These approaches largely ignore the relationship among multiple labels. Most of recent few approaches exploit the relationship of the labels using a "flat" label structure. Jain et al. in [7] presented an uncertainty measure that generalizes margin-based uncertainty to the multi-class case for active learning. In [11], Guojun et al. proposed a two-dimensional framework which considers the sample dimension and the inherent label correlation. Although modeling the labeling relationship of the labels using a "flat" label structure can boost the performance of active learning, it is arguable that exploiting label tree structure in the active learning schemas can further push the performance.

3. ACTIVE LEARNING EMBEDDING LABEL TREE

In this section, we propose to measure the uncertainty with the variance of the category prediction, and derive a method to embed the hierarchical tree of labels into a latent semantic space. All the category labels, both in the leaf

nodes and in the intermediate nodes as well as the training data, are first embedded into this space. Then the variance is computed by considering each label as a point in the space, and this variance is utilized as a measure of uncertainty.

3.1 Label Tree Embedding

We assume that our input consists of instances, represented as a set of vectors $x_1, x_2, \dots, x_N \in X$ of dimensionality d . In addition, these instances are accompanied by single topic labels $y \in \{1, 2, \dots, m\}$ that lie in the label tree T with m total topics. A label tree [1] is a tree $T = (V; E)$ with nodes V and edges E . Each node $v \in V$ is associated with a set of class label $l(v) \in \{1, 2, \dots, m\}$. It is required that the set labels $\{1, 2, \dots, m\}$ has a one-to-one mapping to the set of leaf nodes, and each non-root node's label set should be a subset of its parent's label set. A cost matrix $C \in \mathbb{R}^{m \times m}$ is defined, where $C_{\textcircled{y}, \textcircled{z}}$ is the distance of the labels between class \textcircled{y} and \textcircled{z} , and $C_{\textcircled{y}, \textcircled{y}} = 0$. The class distance matrix could be obtained from side-information from a category tree. In this paper, the distance between two labels is defined as the length of the shortest path between corresponding two nodes in the tree.

Given a label tree $T = (V; E)$, and the labeled data $(x_1; y_1), \dots, (x_N; y_N) (y_i \in \{1, 2, \dots, m\})$ of all the labels, we would like to embed both the labels and the data into a space, with the following criteria: 1), The embedded labels should be able to characterize the tree structure of these labels; and 2), the embedded labels should be representative of the data points of the accompanied category. Suppose the label $y = \textcircled{y}$, let $e_{\textcircled{y}} = [0; \dots; 1; \dots; 0]$ be the vector with a single 1 in the \textcircled{y} -th position and the others 0. This vector can be embedded with a linear transformation P :

$$z_{\textcircled{y}} = P e_{\textcircled{y}} \quad (2)$$

In the semantic space, each label $y = \textcircled{y}$ is represented as a prototype $z_{\textcircled{y}}$ after the embedding. Better results in this semantic space can be expected when the prototypes of similar categories are closer than those of dissimilar categories. We consider the distance between two embedded labels \textcircled{y} and \textcircled{z} in the semantic space, which is defined as $\|z_{\textcircled{y}} - z_{\textcircled{z}}\|$. It should both reflect the distance defined by the tree structure and the distance of the corresponding data with labels \textcircled{y} and \textcircled{z} . We use $t_{\textcircled{y}, \textcircled{z}}$ as an estimate of dissimilarity and aim to place the prototypes such that the distance $\|z_{\textcircled{y}} - z_{\textcircled{z}}\|$ reflects the cost specified in $t_{\textcircled{y}, \textcircled{z}}$. More formally, we set the distance error of the two in Eq.3 as:

$$\|z_{\textcircled{y}} - z_{\textcircled{z}}\| - t_{\textcircled{y}, \textcircled{z}} \quad (3)$$

where $C_{\textcircled{y}, \textcircled{z}}$, as we mentioned above, is the geographic distance. $t_{\textcircled{y}, \textcircled{z}}$ is the distance between the center of the data in the class \textcircled{y} and class \textcircled{z} defined as:

$$t_{\textcircled{y}, \textcircled{z}} = \left\| \frac{1}{N_{\textcircled{y}}} \sum_{y_i = \textcircled{y}} x_i - \frac{1}{N_{\textcircled{z}}} \sum_{y_j = \textcircled{z}} x_j \right\| \quad (4)$$

where $N_{\textcircled{y}}$ is the number of data points within category \textcircled{y} . Based on the embedding rule mentioned above, we want to minimize the error between the semantic distance $\|z_{\textcircled{y}} - z_{\textcircled{z}}\|$ and the estimate of dissimilarity $t_{\textcircled{y}, \textcircled{z}}$ for all the classes. Thus, the final objective can be written as Eq.5 and P can be obtained by minimizing the objective function:

$$P_{\text{mds}} = \arg \min_P \sum_{\textcircled{y}, \textcircled{z}} (\|z_{\textcircled{y}} - z_{\textcircled{z}}\| - t_{\textcircled{y}, \textcircled{z}})^2 \quad (5)$$

where $\mathcal{C}_i = \mu_1, \mu_2, \dots, \mu_m$. If the label distance C_{ij} defines squared Euclidean distance, minimizing Eq.5 is actually the Multidimensional Scaling problem [5], and can be solved via eigenvector decomposition. Define matrix $\mathbf{B} = \frac{1}{2}\mathbf{H}\mathbf{C}^0\mathbf{H}$, with $\mathbf{C}^0_{ij} = t_{ij} - C_{ij}$ and centering matrix \mathbf{H} defined as $\mathbf{H} = \mathbf{I} - \frac{1}{n}\mathbf{1}\mathbf{1}^T$, where $\mathbf{1}$ is a vector of n ones. Let the eigenvector composition of $\mathbf{B} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^T$. The optimal \mathbf{P} that minimizes Eq.5 can be obtained as $\mathbf{P} = \mathbf{V}\mathbf{\Lambda}^{\frac{1}{2}}$.

3.2 Uncertainty Measure

After embedding each class label \mathcal{C} into a point $z_{\mathcal{C}}$ in semantic space, and suppose we have a probability mass $p_{\mathcal{C}} = p(y = \mathcal{C}|x)$ at the point $z_{\mathcal{C}}$ given instance x , the variance can be defined as the trace of the covariance matrix of all the label points, which is similar to A-optimality for the information matrix [12]:

$$\text{Var}(y; x) = \text{Tr}\left(\sum_{\mathcal{C}=1}^m p_{\mathcal{C}} z_{\mathcal{C}} z_{\mathcal{C}}^T - d_y d_y^T\right) \quad (6)$$

where $d_y = \sum_{\mathcal{C}=1}^m p_{\mathcal{C}} z_{\mathcal{C}}$ is the mean of all label vectors. From an estimation point of view, the larger the variance of an instance is, the more difficult it is to estimate the true label of this instance. By encoding the structure of the labels into the location of the label points, the variance measure is able to exploit this information and achieves better sampling efficiency.

If all the labels are independent(not hierarchical structure), they can be embedded uniformly into a high dimensional Euclidean space by setting the projection matrix $\mathbf{P} = \mathbf{I}$. Suppose $y \in \{1, 2, \dots, m\}$, then we can embed them into a m -dimensional space. In particular, we can embed label $y = \mathcal{C}$ to $e_{\mathcal{C}}$, where $e_{\mathcal{C}}$ is a vector with its \mathcal{C} -th element 1 and the rest of element zero. Then the variance Eq.6 can be derived as:

$$\text{Var}(y; x) = \sum_{\mathcal{C}=1}^m p_{\mathcal{C}}(1 - p_{\mathcal{C}}) \quad (7)$$

This shows that the proposed methods is actually a generalization of the basic variance-based measure. The proposed approach for uncertainty measure is depicted in Alg.1.

Algorithm 1 Variance-based Uncertainty Measure Embedding Label Tree

- 1: Input: The label tree $T = (V; E)$, the labeled data $(x_1; y_1), \dots, (x_N; y_N)$ ($y_i \in \{1, 2, \dots, m\}$), an unlabeled instance $x \in U$
 - 2: Output: uncertainty measure $\text{Var}(y; x)$ given x
 - 3: Take label tree $T = (G; E)$, all the labels $y \in \{1, 2, \dots, m\}$ with their accompanied training datasets, make the embedding functions using Eq.5.
 - 4: Obtain the projection matrix \mathbf{P} by minimizing Eq. 5 using eigenvector composition.
 - 5: Project each $e_{\mathcal{C}}$ into $z_{\mathcal{C}}$ using Eq. 2.
 - 6: Compute each $p_{\mathcal{C}} = p(y = \mathcal{C}|x)$ at the point $z_{\mathcal{C}}$ given instance x .
 - 7: Compute the uncertainty using Eq.6.
 - 8: Return $\text{Var}(y; x)$.
-

In each step, the uncertainty measure is computed for all the unlabeled training data, and the unlabeled training data

Figure 2: Description of the synthetic dataset

with the largest uncertainty measure is given to the human labeler for labeling.

4. EXPERIMENTAL RESULTS

In order to demonstrate the effectiveness of our proposed approach, we evaluate it on both synthetic and real-world datasets. To make a prediction of the label given an input x , the hierarchical SVM [3, 4] is employed. The hierarchical SVM algorithm traverses the tree from the root until it reaches a leaf node, and at each node, follows the child that has the largest classifier score. The label at the final leaf node is outputted as the classification label. In all the experiments, the proposed hierarchical active learning is compared with two baseline methods: the entropy-based uncertainty sampling, and random sampling. In the paper, we report prediction accuracy and tree-loss [3] with the approaches on different datasets.

4.1 Synthetic Data

First we use synthetic data to clearly illustrate our active learning approach. We use a simple 2-level label tree shown in Fig.2a. The four leaf classes are assumed to be drawn from four independent Gaussian distributions, respectively, and 1000 sample points are drawn from each of four Gaussian distributions. The mean of the Gaussian of the four labels: $\backslash a1$, $\backslash a2$, $\backslash b1$ and $\backslash b2$ is set to $[0,0,0,0]$, $[0,0,1,1]$, $[0.5,0.5,-1,-1]$, $[0.5,0.5,-2,-2]$ respectively as shown in Fig.2b, and the standard derivation of each Gaussian is set to 0.5. We randomly divide the data in each leaf node into three parts: 200 for initial training, 400 samples for active learning and 400 for testing. In each step one sample is selected to be labeled, and we then tracked the classification accuracies after each step. We plot the test accuracies as well as the tree-loss as the various methods learn each additional sample selected in every active learning step, and the comparisons between different selection approaches are shown in Fig.3. Hierarchical active learning maintains the best performance under both measures compared to the entropy-based method and random sampling.

4.2 Real-world Data

We also performed experiments on the widely used benchmark set called RCV1-v2/LYRL2004 [8]. Here, the documents have been tokenized, stopped and stemmed to 47,236 unique tokens(features) and represented as L2-normalized $\log tf-idf$ vectors. The associated taxonomy of labels, which are the topics of the documents, has 101 nodes organized in a forest of 4 trees. We divided this dataset with 3000 data points into 3 subsets: 200 for initial training, 500 for training and 2300 for testing. We repeat the sampling step and measure the average of the accuracy and tree-loss at each active learning step. In Fig.4, we plot the average accuracy

(a) Accuracy comparison with 3 algorithms (b) Tree-loss comparison with 3 algorithms

Figure 3: Experimental results on the synthetic dataset

(a) Accuracy comparison with 3 algorithms (b) Tree-loss comparison with 3 algorithms

Figure 4: Experimental results on RCV1-v2 dataset

cies and tree-loss on the three competing methods at each learning step. We found that for this dataset, the proposed approach achieves better performances when the number of the labeled data is small. When the amount of labeled data is large, all the three methods achieve similar performance. These shows that the hierarchical active learning does improve the learning rate and classification performance compared to the baseline methods.

5. CONCLUSION

We have proposed an embedding-based uncertainty measure for hierarchical active learning. Both the label tree and the accompanied training data are embedded into a semantic space in which the uncertainty is computed. The proposed uncertainty measure is able to exploit the topology structure of the category labels to efficiently predict the most informative sample to query. Experimental results on both synthetic and real-world datasets demonstrated that the proposed active learning method for hierarchical classification task can improve upon the learning rate and performance compared to the baseline methods. The results indicate that utilizing the hierarchical structure of the category labels is very helpful to active learning. In future, we will consider some problems where the category labels are organized with more complicated structures.

6. ACKNOWLEDGMENT

This work is supported in part by NSF award numbers CCF-0621443, OCI-0724599, CCF-0833131, CNS-0830927, IIS-0905205, OCI-0956311, CCF-0938000, CCF-1043085, CC

F-1029166, and OCI-1144061, and in part by DOE grants DE-FG02-08ER25848, DE-SC0001283, DE-SC0005309, DE-SC0005340, and DE-SC0007456. The authors thank Dr. Weikeng Liao for helpful comments.

7. REFERENCES

- [1] S. Bengio, J. Weston, and D. Grangier. Label embedding trees for large multi-class tasks. In J. D. Laerty, C. K. I. Williams, J. Shawe-Taylor, R. S. Zemel, and A. Culotta, editors, *NIPS*, pages 163{171. Curran Associates, Inc., 2010.
- [2] K. Brinker. On active learning in multi-label classification. *From Data and Information Analysis to Knowledge*, 2006.
- [3] L. Cai and T. Hofmann. Hierarchical document categorization with support vector machines. In *Proceedings of the thirteenth ACM international conference on Information and knowledge management, CIKM '04*, pages 78{87, New York, NY, USA, 2004. ACM.
- [4] S. Dumais and H. Chen. Hierarchical classification of web content. In *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, SIGIR '00*, pages 256{263, New York, NY, USA, 2000. ACM.
- [5] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Wiley-Interscience, 2001.
- [6] J. Huang, S. Ertekin, Y. Song, H. Zha, and C. L. Giles. Efficient multiclass boosting classification with active learning. In *SIAM INTERNATIONAL CONFERENCE ON DATA MINING*. Society for Industrial and Applied Mathematics, 2007.
- [7] P. Jain and A. Kapoor. Active learning for large multi-class problems. *Computer Vision and Pattern Recognition, IEEE Computer Society Conference on*, 0:762{769, 2009.
- [8] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li. Rcv1: A new benchmark collection for text categorization research. *J. Mach. Learn. Res.*, 5:361{397, Dec. 2004.
- [9] D. J. MacKay. Information-based objective functions for active data selection. *Neural Computation*, 4:590{604.
- [10] E. M. N. Ebrahimi. Measuring informativeness of data by entropy and variance. *Slottje(ed.)*, 1999.
- [11] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional multilabel active learning with an efficient online adaptation model for image classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31:1880{1897, 2009.
- [12] B. Settles. Active learning literature survey. Technical report, 2010.
- [13] R. Yan, J. Yang, and A. Hauptmann. Automatically labeling video data using multi-class active learning. In *Proceedings of the Ninth IEEE International Conference on Computer Vision - Volume 2, ICCV '03*, pages 516{, Washington, DC, USA, 2003. IEEE Computer Society.