MULTI-CLASS ACTIVE LEARNING FOR VIDEO SEMANTIC FEATURE EXTRACTION

Rong Yan, Alexander Hauptmann

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213
{yanrong, alex+}@cs.cmu.edu

Abstract

Active learning has been demonstrated to be a useful tool to reduce human labeling effort for many multimedia applications especially for those handling large video collections. However, most of the previous work on active learning only focuses on binary classification, which greatly limits the applicability of active learning. In this paper, we present a multi-class active learning approach which extends active learning from binary classification to multi-class classification using a unified representation with margin-based loss functions. The experimental results on the TREC03 semantic feature extraction task shows that the proposed active learning approach works effectively even with a significantly reduced amount of labeled data.

1. INTRODUCTION

As the increasing amount of multimedia information is driving the demand for content-based access to video data, new challenges have been created. Machine learning approaches are becoming more and more useful for improving the performance of multimedia information retrieval, but a large amount of human effort is still required to the annotate the training data. Unfortunately, manually annotating training data is not only labor intensive and time-consuming for large video archives, but also subject to human errors. The easiest way to reduce the labeling effort is to request a human to label some randomly selected data, and automatically propagate the labels to the entire collection using a supervised learning algorithm. However, random sampling cannot always provide the most useful data to label and thus can waste a lot of human labeling effort. A better approach would be to select non-random examples which, if labeled, will provide the most information for the learning algorithm. This motivated us to develop an incremental learning framework with interaction/supervision from a human. This framework is known as active learning.

The effectiveness of active learning to reduce labeling cost has been demonstrated by previous work [1, 2, 3]. An

Fig. 1. Illustration of multi-class active learning

active learner may begin with a pool of unlabeled data, select a set of unlabeled examples to be manually labeled as positive or negative and learn from the newly obtained knowledge repetitively. This type of problem can also be called "query learning" or "selective sampling" [1]. Typically, the unlabeled examples can be selected by means of either minimization of the learner's expected error [1] or maximization of information gain / version space reduction [2, 3].

On the other hand, multi-class classification is widely applied in many areas of multimedia processing such as face recognition, people identification, image categorization and semantic feature extraction from video. However, most previous studies on active learning simply focused on binary classification problems, and few of them mention selection strategies in the context of multi-class classification, which greatly limits the applicability of active learning. In this paper, we present a multi-class active learning approach using a unified representation with margin-based loss functions to minimize human effort in the task of semantic feature extraction. Figure 1 illustrates the process of multi-class active learning. It enables a learning algorithm to select the most informative unlabeled data for all classes simultaneously instead of just for the binary classes. The experiments on the TREC03 feature extraction task indicate that our multi-class

active learning approach works effectively even when the amount of training data is significantly reduced.

2. MULTI-CLASS ACTIVE LEARNING

We present a unified multi-class active learning framework in this section¹. In the following discussion, let $\mathcal{X}$ denote the domain of possible examples, $\mathcal{Y}$ be a finite set of classes. Formally, the learning algorithm takes a set of training examples $(x_1, y_1), \dots, (x_m, y_m)$ as input, where $y_i \in \mathcal{Y}$ is the label assigned to example $x_i \in \mathcal{X}$. We pay particular attention to learning algorithms that attempt to minimize a margin-based loss function, called margin-based learning algorithms [5]. This includes a large family of well-studied algorithms with different loss functions and minimization algorithms, such as decision trees, logistic regression, support vector machines (SVMs) and AdaBoost. The margin-based learning algorithms always minimize the loss function with respect to the margin, i.e. $\frac{1}{m} \sum_{i=1}^m L(y_i f(x_i))$, where L is some loss function $L: \mathbb{R} \rightarrow [0, \infty)$.

First we want to extend margin-based learning algorithms to multi-class classification since most of the algorithms were originally devised for binary classification. Allwein et al. [5] have proposed a unified approach for decomposing multi-class problems into a set of binary-class problems. Each decomposition can be represented by a coding matrix $M \in \{-1, 0, +1\}^{k \times l}$, where k is the number of classes and l is the number of binary classification problems. $M_{rs} = 1$ indicates that the examples in the class r are considered as positive examples for classification problem s . Similarly, for classification problem s , if $M_{rs} = -1$ the examples in the class r are considered negative examples, and if $M_{rs} = 0$ we don't care how the learner categorizes the example in the class r .

Orthogonal to the problem of coding matrix selection, the learning algorithm has to assign examples to a predicted class $\hat{y} = 1, \dots, k$ given the labels provided by the binary classifiers. We adopted a loss-based decoding scheme

$$d_L(M_r, f(x)) = \sum_{s=1}^l L(m_{rs} f_s(x)), \quad (1)$$

where L is a loss function for both decoding schemes. The predicted label $\hat{y} = \arg \min_r d(M_r, f(x))$.

2.1. Optimal Multi-Class Active Learning

One of the most important components in active learning is the sample selection function which aims at selecting a set of informative examples to label. Specifically, it will be reasonable for a margin-based learning algorithm to use the

minimization of the training loss as the criterion [5]. Therefore the goal for an optimal active learner is to search for those unlabeled examples which can minimize the expected loss on the data set.

Let $P(y|x)$ be the conditional distribution over an example x , and $P(x)$ be the marginal distribution of x . The learner has been given a labeled training set D , and output a set of estimated loss $d_L(M_y, f^D(x))$ for every x in the pool $\mathcal{P}$. We denote $d_L(M_y, f^D(x))$ as $d_L(f^D)$ for the sake of simplicity in the following discussion. We can then write the expected risk function of the learner as follows,

$$R(f^D) = E_x E_{y|x} (d_L(f^D)) = \int_{\mathcal{X}} \sum_{y \in \mathcal{Y}} d_L(f^D) P(y|x) P(x) dx.$$

A multi-class active learner has to select a set of unlabeled examples, or query set D^+ from the pool and ask a human for their labels. After every example x_i^* in D^+ is given labels $y_i^* \in \mathcal{Y}$ and added in the training set, an updated learner will be trained on the training set $D^* = D \cup D^+$. The optimal learner can choose the optimal query set D_{opt}^+ so that the updated learner should have the largest risk reduction,

$$D_{opt}^+ = \arg \max_{D^+} (R(f^D) - R(f^{D^*})). \quad (2)$$

Because it is rather difficult to estimate the expected risk over the full distribution, $P(x)$, it is more feasible to measure the risk over all the examples in the pool. Therefore, the minimization function becomes

$$R(f^D) - R(f^{D^*}) = \sum_{x \in \mathcal{P}} E_{y|x} (d_L(f^D) - d_L(f^{D^*})). \quad (3)$$

In theory, the maximization of (3) straightforwardly leads to the optimal query set D^+ . Unfortunately, in practice, it is intractable to compute all the $2^{|\mathcal{P}| - |\mathcal{D}|}$ possible combinations even in one round, not to mention selecting unlabeled examples iteratively. One of the feasible solutions is to select only one unlabeled example each time, such that the choice of examples greatly reduces to $|\mathcal{P}| - |\mathcal{D}|$. Plus, many learning algorithms such as SVMs and Naive Bayes have efficient algorithms for incremental learning. However, even with these optimizations usually the computation is not affordable for many real-world applications, which leads us to develop some approximated strategies with similar performance but significantly less computational effort.

2.2. Approximated Sample Selection Strategies

In this section, we describe several practical sample selection strategies as alternatives for optimal multi-class active learning. As noted before, it is computationally intensive to maximize (3) due to the re-learning of the classifiers to estimate the new expected risk. Moreover, because typically

¹More details of this algorithm can be found in [4]

only a small number of labeled data are available for training, the estimation for $P(y|x)$ might be unreliable. To make multi-class classification more practical, we will use some simple heuristics to simplify our selection strategies.

The first step of the approximation is to eliminate the computation-intensive components that have to be reestimated after adding additional data, which is the prediction function $f(D^*)$. Based on two assumptions presented in [4], we can rewrite (3) into the following equation which maximizes the expected risk for only one example,

$$D^+ = \arg \max_x E_{y|x}(d_L(f^D)). \quad (4)$$

Substituting (1) into $E_{y|x}(d_L(f^D))$ in (4), we get

$$D^+ = \arg \max_x \sum_{y \in \mathcal{Y}} \sum_{s=1}^l P(y|x) L(m_{ys} f_s^D(x)) \quad (5)$$

A probability estimation scheme needs to be provided for the conditional probability $P(y|x)$. Note that the classification confidence presented in the loss function is not the posterior probability, so we cannot straightforwardly treat the confidence as the estimation of $P(y|x)$. To address this, we adopt the best worst case model which has been suggested in several papers [2, 1]. This model expects that the predicted label is exactly the true label of the unlabeled data, e.g. $P(y|x) = 1$ if y is predicted otherwise $P(y|x) = 0$. Therefore it approximates the expected loss function with the smallest loss function among all the possible labels. Thus (5) can be rewritten as

$$D^+ = \arg \max_x \sum_{s=1}^l L(m_{ys} f_s(x)), \quad (6)$$

where y_x is the predicted label for example x . The rationale for this model is to choose the most ambiguous examples with the maximum expected loss for the predicted label. In the case of $l = 1$ where only binary classes are predicted, this strategy can be reduced to the problem of choosing the example x with $\max_x L(y, f(x))$. This can be interpreted as selecting the examples closest to the decision boundary, which is a common sample selection criterion in binary active learning tasks [3].

Some of the most relevant work to our approach was done by Tong et al [6]. Viewing the multi-class classification problem as an extension of the binary case, they propose a simple heuristic to select the next unlabeled example that minimizes the maximum model loss

$$x = \arg \min_x \max_y \prod_i Area(\mathcal{V}_{x,y}^{(i)}). \quad (7)$$

By approximating the size of the version space $Area(\mathcal{V}_{x,y}^{(i)})$ with $(1 + y_i f_i(x)) Area(\mathcal{V}^{(i)})/2$, we can show that this heuristic is a special case of our best worst case model with loss function $\log 1/(1 + x)$.

3. EXPERIEMENTS

In this section, we describe experiments on semantic feature extraction using the development set of the TREC03 Video Track Feature Extraction Task [7] to demonstrate the effectiveness of our multi-class active learning approach. To construct a multi-class data set, we randomly sampled 893 video shots from 5 mutually exclusive categories using the labels provided by the common annotations, including 277 shots of studio settings, 215 shots of weather news, 123 shots of hockey events, 175 shots of basketball events and 103 shots of baseball events.

Low-level features including color, edge and face are generated to learn the semantic features. After dividing an image into 5*5 equal regions, the color features in each region are computed as the color histograms for each separate color channel, where the three channels are derived from the HSV color space. We use 16 bins for hue and 6 bins for both saturation and value. A canny edge detector was applied to extract edges from the images. The edge histogram includes a total of 73 bins. The first 72 bins represent the edge directions quantized at 5 degree intervals and the last bin represents a count of the number of pixels that have not contributed to any edges. Schneiderman's face detector algorithm [8] was used to extract frontal and profile faces. RBF-kernel support vector machine(SVM) was trained as the base classifier with parameter 0.05.

Fig. 2. Classification error rate against the number of training examples for multi-class active learning and supervised learning with random sampling

In figure 2, we compared the performance between supervised learning with random sampling, binary class active learning which independently selected examples for each decomposed binary classifiers using **Simple** strategy[3], and the multi-class active learning approach with best worst case model. The one-against-all coding scheme and hinge loss function $(1 - y)^+$ [5] are used. We started with 50 training examples, and added 5 more training examples for each

Fig. 3. Key frames of video shots selected by active learning in first 4 iteration. Texts above are the category for each shot.

iteration until it reached 200 examples.² For the sake of simplicity, we simulated the human labeling process using complete, true data labels without actually asking a human for the labels at each step. The results show that the classification error rate of multi-class active learning decreased much faster and more consistently than both the random sampling strategy and binary active learning. When comparing both methods with 200 training data, active learning resulted in the lowest error rate, which was only half of the error of random sampling. For another view of point, the error rate of active learning was reduced to 24% with only 30 additional training examples while the random sampling required more than 150 additional training examples to achieve a similar performance. Figure 3 is depicted to show the key frames of video shots selected by the proposed selection strategies in first 4 sampling iteration.

4. CONCLUSION

In this paper, we presented a multi-class active learning framework to reduce the human labeling effort for multi-class classification. Our experiments on semantic feature extraction demonstrate that an active learner with approximated sample selection strategies can achieve remarkably good performance with much less human labeling effort compared to supervised learning with random sampling.

However, it may be too early to claim that the active learning definitely outperforms supervised learning with random sampling, because active learning requires users to wait for the learning algorithm to output the classification results, which might take a long time as a large video collection is processed. A better cost function needs to be defined to take all these issues into account. Another promising avenue for

future work is to use user feedback to adjust the weights of multi-modal information such as the output of text classification and image classification.

5. REFERENCES

- [1] C. Campbell, N. Cristianini, and A. Smola, "Query learning with large margin classifiers," in *Proc. 17th International Conference on Machine Learning (ICML00)*, 2000, pp. 111–118.
- [2] G. Schohn and D. Cohn, "Less is more: Active learning with support vector machines," in *Proc. 17th International Conf. on Machine Learning*, 2000, pp. 839–846, Morgan Kaufmann, San Francisco, CA.
- [3] S. Tong and E. Chang, "Support vector machine active learning for image retrieval," in *ACM Intl. Conf. on Multimedia*, 2001, pp. 107–118.
- [4] R. Yan, J. Yang, and A. G. Hauptmann, "Automatically labeling data using multi-class active learning," in *Proceedings of the Intl. Conf. on Computer Vision*, 2003, pp. 516–523.
- [5] E. L. Allwein, R. E. Schapire, and Y. Singer, "Reducing multiclass to binary: A unifying approach for margin classifiers," in *Proc. 17th International Conference on Machine Learning (ICML00)*, 2000, pp. 9–16.
- [6] S. Tong, *Active Learning: Theory and Applications*, Ph.D. thesis, Stanford University, CA, 2001.
- [7] TREC2003, "TREC2003 video track, <http://www-nlpir.nist.gov/projects/tv2003/tv2003.html>," .
- [8] H. Schneiderman and T. Kanade, "Object detection using the statistics of parts," *International Journal of Computer Vision*, 2002.

²For the binary active learning, we pick the most ambiguous example for each class which exactly amounts to 5 examples per iteration.