A GENERALIZED MULTIPLE INSTANCE LEARNING ALGORITHM FOR LARGE SCALE MODELING OF MULTIMEDIA SEMANTICS

Milind R. Naphade and John R. Smith

IBM Thomas J. Watson Research Center
19 Skyline Drive, Hawthorne, NY 10532
naphade@us.ibm.com

ABSTRACT

Statistical learning techniques provide a robust framework for learning representations of semantic concepts from multimedia features [1]. The bottleneck is the number of training samples needed to construct robust models. This is particularly expensive when the annotation needs to happen at finer granularity. We present a novel approach where the annotations may be entered at coarser spatial granularity while the concept may still be learnt at finer granularity. This can speed up annotation significantly. Using the multiple instance learning paradigm, we show that it is possible to learn representations of concepts occurring at the regional level by using annotations for several images. We present a generalized multiple instance learning algorithm that can scale to a large number of training samples as well as a large number of instances per bag. The algorithm also provides the ability to plug in different density modeling or regression techniques. Using the TREC 2001 Corpus we demonstrate the superior performance of the proposed algorithm over the existing diverse density algorithm [2].

1. INTRODUCTION

Enabling semantic detection and indexing is an important task in multimedia content management. Learning and classification techniques are increasingly relevant to the state of the art content management systems. From relevance feedback to semantic detection, there is a shift in the amount of supervision that precedes retrieval from light weight classifiers to heavy weight classifiers. It is therefore natural that machine learning and classification techniques are making an increasing impression on the state of the art in media indexing and retrieval. Techniques such as relevance feedback can be thought of as non-persistent lightweight binary classifiers using incremental learning to improve retrieval performance. Techniques for detection of explosion, outdoors [3], and over thirty other visual concepts [4] etc. on the other hand require considerable supervision during the training phase. We need techniques that utilize the annotation procedure intelligently. One way to speed up annotation is to deploy active learning during annotation. An orthogonal approach for concepts that have regional support is to accept annotations at coarser granularity. While building a model for the regional concept *Sky*, the user is thus not required to select the region in the image which corresponds to this regional label. It is up to the system then, to learn from several possible positive and negatively annotated examples, how to represent the concept *Sky* using regional features. This learning paradigm which disambiguates across granularity is called multiple instance learning [5] and was originally applied to problems in drug discovery. Ratan et al [2] presented the application of this framework to content based retrieval. Unfortunately the application of the diverse density algorithm to large scale media data sets exposes several problems in the algorithm. In this paper we propose a new generalized multiple instance learning (GMIL) algorithm that is designed to avoid the shortcomings of the diverse density algorithm and also provide the opportunity to experiment with different existing classification and regression algorithms to be plugged into the generalized algorithm based on the domain and the problem at hand.

2. LEARNING REGIONAL CONCEPTS FROM GLOBAL ANNOTATION

The essence of applying multiple instance learning to disambiguate across granularity is shown in Figure 1. Here we use the same notation of *Bags* and *Instances* as in [5]. A Bag is a collection of instances. Annotation is provided at the bag level but actually reflects the label of one or more instances in that bag. If at least one instance (region) that is positive the corresponding bag is labeled positive. Conversely a bag is labeled negative when all instances (regions) are negative for the semantic concept. The problem is to then learn in some feature space a concept point or a set

This material is based upon work funded in part by the U.S. Government. Any opinions, finidings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the U.S. Government

of concept points that are closest to maximum possible positive bags (i.e. instances in these bags) and simultaneously away from as many negative bags (i.e. negative instances) as possible. Figure 1 uses a 2 dimensional feature space to illustrate this idea.

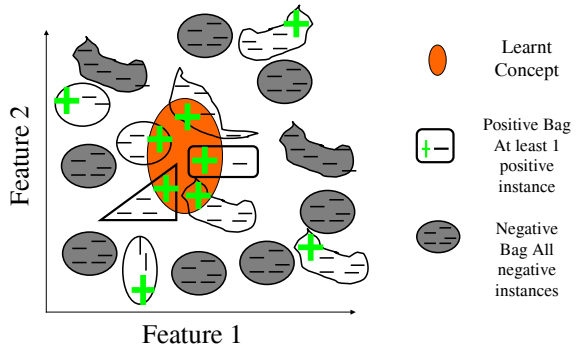

Fig. 1. This figure shows a distribution of *bags* in a two dimensional feature space. Only bags are labeled. Multiple instance learning can result in the region in orange as the target concept as it is closest to as many positive bags as possible while farthest from many negative bags.

2.1. The Diverse Density Algorithm

To learn the target concept from annotated bags as against annotated instances Maron et al [5] proposed applying the diverse density algorithm. This algorithm attempts to find areas in feature spaces that are close to at least one instance from every positive bag and far from every negative instance. Notation used in this paper resembles [5] and details can be found in [5]. Let B_i denote a bag, B_i^+ be the positively labeled bags and B_i^-, the negatively labeled bags. For each bag let B_{ij} denote the instance (as in a region in our case) in the bag. Under a simplistic assumption that the true concept is a point mass in some feature space, this true concept t can be found by maximizing the probability $P(t|B_1^+, \ldots, B_p^+, B_1^-, \ldots, B_n^-)$ over the training set (where p is the total number of positively labeled bags and n the negatively labeled bags. If a uniform prior $P(t)$ is assumed we can use Bayes' rule to select t using Equation 1

$$arg\max_t P(B_1^+, \ldots, B_p^+, B_1^-, \ldots, B_n^-|t) \qquad (1)$$

Under the assumption that bags are conditionally independent given t the target concept, and the assumption of uniform prior over t we can then separate the contributions from the positive and negative bags as in Equation 2 and this is what Maron et al [5] define as diverse density.

$$arg\max_t \prod_{i=1}^{i=p} P(t|B_i^+) \prod_{i=1}^{i=n} P(t|B_i^-) \qquad (2)$$

For the experiments reported in this paper, a minor variant of the noisy OR model is tested as in [5] i.e. $P(t|B_i^+) = 1 - P(\bar{t}|B_i^+)$ where $P(\bar{t}|B_i^+)$ is the probability of not being the target instance given this particular positive labeled bag and its instances.

$$P(\bar{t}|B_i^+) = \prod_{j=1}^{j_k} P(\bar{t}|B_{ij}^+) = \prod_{j=1}^{j_k} (1 - P(t|B_{ij}^+)) \qquad (3)$$

where j_k is the number of instances in the j^{th} bag. Similarly $P(t|B_i^-) = \prod_{j=1}^{j_k}(1 - P(t|B_{ij}^-)$ The probability of the target concept given the features of an instance can be defined in several different ways. Perhaps the simplest is a Gaussian like distribution as in Maron et al. [5]. The problem with this optimization is that it is highly nonlinear and the function is not convex. The algorithm sets this up as a nonlinear root finder availing standard mathematical packages. The diverse density algorithm faces a challenge from the number of samples as well as from the large number of instances per bag. This was not evident in the original application to retrieval because of the small number of user provided training samples [2] but in our concept modeling task, this is a formidable challenge. We alleviated the number of instances per bag by segmenting the keyframes and selecting the five largest regions in the keyframe as our instances. Despite this the algorithm was extremely slow to converge.

3. THE GENERALIZED MULTIPLE INSTANCE LEARNING ALGORITHM

The scale and complexity of multimedia concept modeling exposes several problems with the formulation of the diverse density. The scaling problem already mentioned previously (both in terms of the number of bags and the number of instances per bag), the asymmetry in the number of positive and negative training samples and the use of generic hill climbing nonlinear optimization are the three main shortcomings of the diverse density algorithm. The diverse density algorithm formulation penalizes proximity from negative samples so heavily that in most multimedia learning problems with highly nonlinear separating decision boundaries between positive and negative hypotheses, the diverse density solution is usually not robust. Based on these observations we propose the generalized multiple instance learning algorithm. In Algorithm 1 the first step is to recognize the fact that the information in negative bags provides certainty as against the information in positive bags which hold

Training
Data : Labeled Bags B_i^-,B_i^+
for instance j in bag i feature vector X_{ij}

Result : Negative and Positive hypotheses models H_0 and H_1
Initialization: Using all instances in all negative bags B_i^- build a negative hypothesis model H_0
for *Each positive bag* B_i^+ **do**
  $sumbag_i = 0$
  for *Each instance* j *in positive bag* B_i^+ **do**
    Evaluate $P(X_{ij}|H_0)$
    $sumbag_i = sumbag_i + P(X_{ij}|H_0)$
  end
  for *Each instance* j *in positive bag* B_i^+ **do**
    $w_{ij} = \frac{1-P(X_{ij}|H_0))}{sumbag_i}$
  end
end
Now build model H_1 for the positive hypothesis using X_{ij}, w_{ij} for all instances in the positive bags, where the contribution of each instance to the model is weighed using w_{ij}

Classification
Data : Negative and Positive hypotheses models H_0 and H_1
Unlabeled bag U
For instance j in bag U feature vector U_j

Result : Classification C for bag U
C=0
for *Each instance* j *in unlabeled bag* U **do**
  Evaluate $P(X_{ij}|H_0)$
  Evaluate $P(X_{ij}|H_1)$
  Evaluate likelihood ratio test $L_j = \frac{P(X_{ij}|H_1)}{P(X_{ij}|H_1)}$
  $C = \max(L_j, C)$
end
C is strength with which the bag is classified as positive

Algorithm 1: Generalized Multiple Instance Learning Algorithm

ambiguity. So a garbage model for the negative hypothesis H_0 can be constructed independently. Further this is used to rank the instances in the positive bag in the order in which each instance was least likely to be generated by the negative hypothesis. Subsequently the model for the positive hypothesis can be trained using some or all instances from each positive bag. During the classification phase, all instances of an unlabeled bag are scored using the likelihood ratio test in Algorithm 1 and the highest ratio is selected as the classification strength of the bag being positive.

We implement this generalized multiple instance using gaussian mixture models for representing the positive and negative hypotheses. The model for the negative hypothesis H_0 is referred to as $N(X^-, \mu, \Sigma)$ where $\mu_\mathbf{i}$ is an n-dimensional vector, and Σ_i is an $n \times n$ matrix. It is estimated using the expectation maximization (EM) algorithm [6] using all instances from the negative bags. The model for the positive hypothesis H_1 is referred to as $N(\{X^+, w\}, \mu, \Sigma)$, and is estimated using the instances from the positively labeled bags from the training set and the weight w assigned to each instance in a positive bag by H_0. The EM algorithm is used for the estimation. In a simplification of the estimation for the positive hypothesis model, we can also use one instance from each positive bag with the highest assigned weight. While the experiments reported in this paper use gaussian mixtures to fit the hypotheses, any other modeling or regression technique can be used. The generalized algorithm does not force the learnt concept to be far from **each and every negative instance**. At the same time since it reverse ranks the instances in the positive bags, this leads to a statistical concentration of instances in positive bags that are in general far from most negative instances. Also since the generalized algorithm is a 2 step procedure and the modeling of each hypothesis is separately done, well known efficient estimation procedures can be easily incorporated as against the diverse density algorithm. For example the convergence of the estimation process for H_0 and H_1 with gaussian mixtures and the EM algorithm is orders of magnitude faster than the nonlinear hill climbing of the diverse density algorithm. A further advantage of the generalized algorithm proposed in this paper is that it provides us with a natural extension to the multiple instance learning of temporal concepts. By merely replacing gaussian mixture models for H_0 and H_1 with hidden Markov models with gaussian mixture observation densities, we can convert the static multiple instance learning algorithm into a dynamic multiple instance learning algorithm that can learn spatial and temporal instances and disambiguate in spatial and temporal features.

4. EXPERIMENTAL SETUP

4.1. The TREC Video 2001 Corpus

For the Video TREC 2001 Benchmark the National Institute for Standards and Technology NIST provided a data set of 11 hours of NIST documentary video (30 clips) comprising more than 7000 shots.

4.2. Feature Extraction

After performing shot boundary detection and key-frame extraction each keyframe was analyzed to detect the 5 largest regions homogeneous in color and texture described by their

bounding boxes. The system then extracts low level visual features at the region level including a color histogram (24-bin linearized HSV histogram) an edge orientation histogram (Sobel filtered and quantized to 24 angles) and Dudani's moment invariants (6) modified for gray-level intensity as shape desriptors.

5. PRELIMINARY RESULTS

From the available TREC Video corpus keyframes, 4688 keyframes were annotated and used for this experiment. The annotated keyframes were split into a training set of 3292 keyframes and a test set of 1396 keyframes. We now compare the performance of our proposed generalized multiple instance learning algorithm with the diverse density [5] using the TREC 2001 corpus and the concept *sky*. In addition we also build a single instance learning model of the concept *sky*. The multiple instance learning algorithms are experimented using annotation at the keyframe level. The single instance learning algorithm uses annotations for the same keyframes but provided at the regional level. The single instance learning model is trained using a binary support vector machine classifier [4]. Figure 2 shows a precision recall plot for the three algorithms. Apart from being much faster than the diverse density algorithm in converging, the proposed generalized multiple instance learning algorithm also outperforms the diverse density and single instance SVM learning.

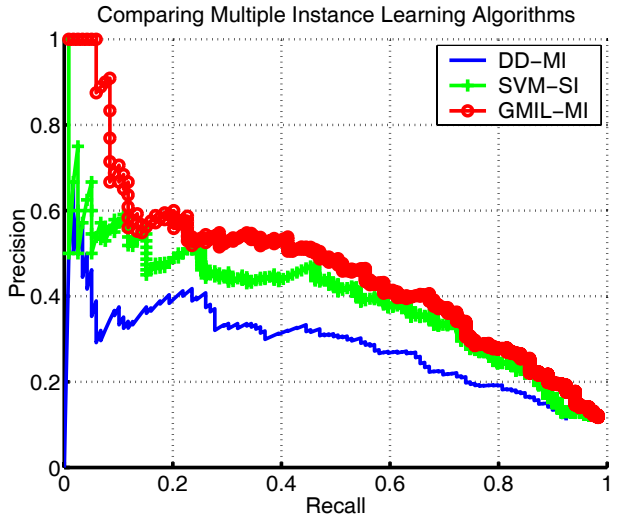

Fig. 2. Precision Recall Curves for *Sky* detection using the TREC Video 2001 corpus. The proposed generalized multiple instance learning (GMIL-MI). The proposed GMIL-MI algorithm outperforms both the diverse density multiple instance learning (DD-MI), and the single instance SVM learning (SVM-SI).

6. FUTURE DIRECTIONS

We present a novel generalized multiple instance learning algorithm to learn representations of regional concepts by using incomplete annotation in the form of global/frame level annotation. This can significantly reduce annotation time without degrading detection accuracy severely. Using the TREC Video Corpus and the regional semantic concept *Sky* we show how the multiple instance learner is able to learn the regional concept despite imperfect segmentation. Using the TREC corpus we demonstrate the superior performance of the proposed generalized multiple instance learning algorithm as compared to the diverse density algorithm. The proposed algorithm also scales to large training corpora. Our proposed algorithm also provides the ability to plug in different density modeling or regression techniques in each phase of multiple instance learning. Future work includes the application of the proposed generalized multiple instance learning algorithm to model spatial-temporal concepts by using temporal models such as hidden Markov models in the estimation procedure of the proposed generalized algorithm.

7. ACKNOWLEDGEMENTS

IBM TREC team (annotation), C. Lin (bounding boxes), A. Amir (shot detection).

8. REFERENCES

[1] M. R. Naphade, I. Kozintsev, and T. S. Huang, "On probabilistic semantic video indexing," in *Proceedings of Neural Information Processing Systems*, Denver, CO, Nov. 2000, vol. 13, pp. 967–973.

[2] A. Ratan, O.Maron, W. Grimson, and T. Lozano-Perez, "A framework for learning query concepts in image classification," in *Proceedings of Computer Vision and Pattern Recognition*, Fort Collins, CO, June 1999, vol. 1, pp. 423–429.

[3] M. Naphade, T. Kristjansson, B. Frey, and T. S. Huang, "Probabilistic multimedia objects (multijects): A novel approach to indexing and retrieval in multimedia systems," in *Proceedings of IEEE International Conference on Image Processing*, Chicago, IL, Oct. 1998, vol. 3, pp. 536–540.

[4] M. Naphade and J. Smith, "Learning visual models of semantic concepts," in *Proc. IEEE International Conference on Image Processing*, Sep 2003.

[5] O. Maron and T. Lozano-Perez, "A framework for multiple instance learning," in *Neural Information Processing Systems*. 1998, MIT Press.

[6] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Proceedings of the Royal Statistical Society*, vol. B, no. 39, pp. 1–38, 1977.