Pattern Recognition 41 (2008) 3214 – 3223

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/pr

Exploiting generalized discriminative multiple instance learning for multimedia semantic concept detection

Sheng Gao*, Qibin Sun

Institute for Infocomm Research, 21 Heng Mui Terrace, Singapore 119613, Singapore

ARTICLE INFO

Article history:
Received 8 October 2007
Received in revised form 4 February 2008
Accepted 31 March 2008

Keywords:
Multiple instance learning
Discriminative training
Semantic concept detection
Area under the ROC curve
Classification accuracy

ABSTRACT

A generalized discriminative multiple instance learning (GDMIL) algorithm is presented to train the classifier in the condition of vague annotation of training samples GDMIL not only inherits the original MIL's capability of automatically weighting the instances in the bag according to their relevance to the concept but also integrates generative models using discriminative training. It is evaluated on the task of multimedia semantic concept detection using the development data set of TRECVID 2005. The experimental results show GDMIL outperforms the baseline systems trained on MIL with diverse density and expectation–maximization diverse density and the system without MIL.

1. Introduction

Enormous digital multimedia is archived and it is still growing exponentially with the popularity of Internet and personal digital multimedia devices. However, efficiently managing (e.g., indexing, search and browsing) such giant multimedia database at the semantic level is still a challenge. In the past, extensive studies on content-based image retrieval have been done on retrieving the image based on low-level feature similarity. Retrieving from the video stream at the semantic level attracts more attention in recent years. The techniques are being advanced by the annual TREC video retrieval evaluation (TRECVID) organized by NIST.[1] They exploited various machine learning algorithms for detecting multimedia semantic concept. For each interested semantic concept, the training samples (i.e., keyframes) are manually annotated based on the visual content. A keyframe is labeled as the positive class if it is relevant to the concept. Otherwise, it is negative [1]. Supervised learning algorithms are then applied to train a classifier based on the annotated training set. Finally, the classifier is used to score and sort the video shots.

In semantic concept detection, the keyframe label is a weak annotation. It means that the concept label is given to the whole image and it is not known which regions or patches are exactly relevant to the concept. For example, an image annotated as a *Car* only tells us that there is a car in the image, but it is not known where the car is. Furthermore, the real-world image may have complex scenes. For example, the annotated *Car* image may contain other objects such as *building*, *tree*, *people*, etc. Thus the positive training image will have many irrelevant regions associated with the concept. Sometimes it may become more severe; for instance, the interested concept may only occupy a small part in the whole image. Thus it is difficult to learn a good concept model from this training set. To eliminate the effects of such noises, the ideal way is to build a training set by exactly annotating the concept in the regions. However, it is infeasible for large-scale images because labeling is very time consuming. Large-scale images annotated at the whole image level could be easily collected through the website. For example, we can use the image search engine such as Google image search [2] to collect the image data for a specific concept. Therefore, it is interesting to develop a learning algorithm that can learn a "*good*" (in terms of selected metric) classifier for the semantic concept.

Multiple instance learning (MIL) is such a framework [2,3,21,22,24–26]. MIL learns the classifier from the labeled bag samples. Each bag is a container, which has multiple instances. For example, in multimedia semantic concept detection, the bag is the whole image and the instance is a patch or region represented in a *D*-dimensional feature space. MIL is originally introduced to predict the drug activity in Ref. [3]. But it now has succeeded in a few other applications

* Corresponding author. Tel.: +65 68748531.
E-mail addresses: gaosheng@i2r.a-star.edu.sg (S. Gao), qibin@i2r.a-star.edu.sg (Q. Sun).

[1] http://www-nlpir.nist.gov/projects/tv2005/tv2005.html.

[2] http://images.google.com/.

0031-3203/$30.00 © 2008 Elsevier Ltd. All rights reserved.
doi:10.1016/j.patcog.2008.03.029

such as content-based image retrieval [4–6], image classification [7,8], semantic concept detection [9], object detection [10], face labeling in news video [11], text categorization [12], etc.

In this paper, we will systematically study the generalized discriminative multiple instance learning algorithm (GDMIL) for semantic concept detection. This is an abstract extension of our previous work in Ref. [13]. We will empirically study the usefulness of GDMIL in semantic concept detection and compare with other two popular MIL algorithm such as diverse density (DD) [3] and expectation–maximization diverse density (EM-DD) [14]. GDMIL can train any type of generative models such as Gaussion mixture model (GMM), hidden Markov model (HMM), etc. Thus it owns the original MIL's capability of automatically weighting the instance in the bag according to their relevance to the concept. Moreover, it integrates expressive power of generative models using discriminative training. Its efficiency on semantic concept detection is evaluated on the large-scale development set in TRECVID 2005. We compare the algorithms in terms of classification accuracy, ranking performance metric such as AUC and precision–recall (PR) curve. Up to now, we have not seen much work for evaluating MIL on the large-scale real-world TREC video image using the ranking measure. Much of the past work is on the photographic image and is only reported on classification accuracy [4,5,7,8]. We shall also demonstrate in our experiments that classification accuracy is not a good measure for evaluating MIL algorithms on semantic concept detection task.

The paper is organized as follows. Next, we review the prior work related to MIL. In Section 3, we discuss the classical MIL formulation in detail. GDMIL will be presented in Section 4. We report our experimental results and analyses in Section 4.1. Finally, we conclude our findings in Section 4.2.

2. Related work

The classical MIL is originally proposed in Ref. [2] to predict the drug activity. The objective is to predict whether a drug molecule can bind well to a target protein or not. In the task, the bag is the molecule and instances are the shapes in the molecule. A bag is positive if at least one shape binds well and it is negative otherwise. The axis-parallel rectangles learning algorithm is proposed to solve the MIL problem. Following the work, the DD algorithm is presented to find the target point in the instance space [3]. It assumes the target is one or more feature point in the instance space. A noisy-or model is employed to calculate the prediction probability of a bag labeled as the target (the positive) or non-target (the negative). With the assumption of target concept having a Gaussian distribution, then the target model is estimated through maximizing the likelihood. Since DD tries several starting points selected from all instance in positive bags, it converges slowly and computation cost is high. Thus EM-DD is introduced [14]. EM-DD uses two steps for estimating the target. In the first expectation step, one instance most closely to the target hypothesis is picked from each bag and other instances are ignored. In the second maximization step, the target hypothesis is updated using the DD algorithm. Since only one instance in each bag is kept in the expectation step, computation cost for the following DD is significantly reduced. Other researchers also modify the kNN (Citation kNN) [15] or SVM to handle MIL problem [12]. Unlike DD, EM-DD and Citation kNN, SVM-based MIL is a discriminative training version for MIL.

Another way for address MIL problem is to use the traditional supervised learning approach. By assuming all instances in the positive bag belong to the positive, then the MIL is a standard supervised learning problem with the complete label information on all instances. In Ref. [16], the empirical comparison between the supervised learning and MIL is studied on various MIL data sets. Some interesting conclusions were obtained such as (1) there is no MIL algorithm superior to others across all domains and (2) some MIL algorithms are consistently superior to their supervised counterparts.

It is common to index the image content using a set of feature vectors among which only a few are relevant to the concept. Thus the task of classifying the image will naturally fit well with MIL framework. Therefore, it is not surprising that there are some previous works on applying MIL to image classification, object detection and content-based image retrieval [4–11].

Since the image feature is often continuous, one popular model to characterize the feature distribution is GMM. For instance in Ref. [17], GMM is used to model the feature distribution of image database in the scenario of content-based image retrieval with relevance feedback. They assumed that one component in GMMs the distribution of the semantic concept while others model the negative class. Since the correspondence between components and models are unknown, active concept learning is proposed. Due to the image features being extracted from the whole image rather than the regions, thus active concept learning in Ref. [17] loses the capability of MIL approaches. In GDMIL, image features are extracted from regions and GMM is applied to model the distributions of the positive class as well as that of the negative class separately. Combining with MIL learning, we can automatically weigh the features in an image in terms of their relevance to the positive class and the negative class.

3. Multiple instance learning

The training samples in MIL are provided at the *bag* level, each bag containing multiple *instances* [3]. Annotation is given to the *bag* while the labels associated to instances are unknown. For the semantic concept detection task, the bag is the whole image while the instances are the particular regions or patches. A D-dimensional feature vector is extracted from each region or patch to represent image content. The union of instances describes the bag. A bag is annotated as the positive if at least one of its instances (i.e., region or patch) is associated to the concept. Otherwise, it is negative. We use the similar notations as in Ref. [3]. B_i^+ is the i-th positive bag with $|B_i^+|$ being the number of instances, and its j-th instance is B_{ij}^+. B_i^- is the i-th negative bag with $|B_i^-|$ being its instance size, whose j-th instance isB_{ij}^-. We assume there are M positive bags and N negative bags in the training set. The target concept t to be estimated is a single or multiple points in the D-dimensional feature space.

The target, t^*, is estimated by maximizing the joint probability of the training samples defined in Eq. (1),

$$t^* = \max_t P(B_1^+, \ldots, B_M^+, B_1^-, \ldots, B_N^- | t) \tag{1}$$

When assuming the bags are conditionally independent given the target and the target has a uniform prior over the concept location, then Eq. (1) will be

$$t^* = \max_t \prod_i^M P(t|B_i^+) \prod_i^N P(t|B_i^-) \tag{2}$$

This is a general definition in Ref. [3]. To define the conditional probability in Eq. (2), a *noisy-or* model is applied. Thus the probability of predicting a positive bag as the positive is calculated as

$$P(t|B_i^+) = 1 - \prod_j (1 - P(t|B_{ij}^+)) \tag{3}$$

Similarly, the probability of negative bag predicted as the negative is calculated as

$$P(t|B_i^-) = \prod_j (1 - P(t|B_{ij}^-)) \tag{4}$$