Pattern Recognition Letters 40 (2014) 19–26

Contents lists available at ScienceDirect

Pattern Recognition Letters

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

Multiple instance learning based on positive instance selection and bag structure construction ☆

Zhan Li [a,*], Guo-Hua Geng [a], Jun Feng [a], Jin-ye Peng [a], Chao Wen [a], Jun-li Liang [b]

[a] School of Information Science and Technology, Northwest University, Xi'an 710069, China
[b] School of Computer Science and Engineering and School of Automation and Information Engineering, Xi'an University of Technology, Xi'an 710048, China

ARTICLE INFO

Article history:
Received 11 April 2013
Available online 3 December 2013

Keywords:
Multiple instance learning (MIL)
Support vector machine (SVM)
K-means clustering
Multiple kernel

ABSTRACT

Previous studies on multiple instance learning (MIL) have shown that the MIL problem holds three characteristics: positive instance clustering, bag structure and instance probabilistic influence to bag label. In this paper, combined with the advantages of these three characteristics, we propose two simple yet effective MIL algorithms, CK_MIL and ck_MIL. We take three steps to convert MIL to a standard supervised learning problem. In the first step, we perform K-means clustering algorithm on the positive and negative sets separately to obtain the cluster centers, further use them to select the most positive instances in bags. Next, we combine three distances, including the maximum, minimum and the average distances from bag to cluster centers, as bag structure. For CK_MIL, we simply compose the positive instance and bag structure to form a new vector as bag representation, then apply RBF kernel to measure bag similarity, while for ck_MIL algorithm we construct a new kernel by introducing a probabilistic coefficient to balance the influences between the positive instance similarity and bag structure similarity. As a result, the MIL problem is converted to a standard supervised learning problem that can be solved directly by SVM method. Experiments on MUSK and COREL image set have shown that our two algorithms perform better than other key existing MIL algorithms on the drug prediction and image classification tasks.

1. Introduction

The multiple instance learning (MIL) is firstly introduced by Dietterich et al. (1997) on classifying molecular pharmaceutical ability according to its "musky" ability, where each molecular (called a bag) suitable for drug design has multiple low-energy shapes or conformations (instances). In the MIL problem, each training example is a bag of instances. A bag is positive if it contains at least one positive instance, and negative otherwise. In another word, negative bags only contain negative instances, while the positive bags contain both positive and negative instances. Thus, the difficulty of MIL lies in its weak label problem arising from positive bag, implying that the labels of instances in positive bag are ambiguous.

Since MIL is suitable to represent many applications, it has emerged in a variety of challenge learning tasks, such as drug activity prediction (Dietterich et al., 1997), image categorization (Chen et al., 2006; Chen and Wang, 2004), image retrieval (Zhang et al., 2002; Li et al., 2011), text categorization (Andrews et al., 2003; Settles et al., 2008), computer security (Ruffo, 2000), face detection (Viola et al., 2006; Zhang and Viola, 2008), pedestrian and car detection (Fung et al., 2007), computer-aided medical diagnosis (Fung et al., 2007), web recommendation (Zhou et al., 2005) etc. In this paper, we focus on efficient MIL algorithm design with special regard to drug activity prediction and image classification.

During the past decade, numerous of MIL algorithms have been proposed, therefore it is difficult to list all existing MIL algorithms here. We mainly focus on algorithms that are most related to our work. The first MIL algorithm, Learning axis-parallel hyper-rectangle concepts (APR), is proposed by Dietterich et al. (1997) in drug design, tempting to find a hyper-rectangle in instance space containing at least one instance within positive bags. Diverse Density (DD) is presented in Maron and Lozano-Pérez (1998) as a general framework for solving multiple instance learning problems. The main idea of DD approach is to find a concept point in the feature space that are close to at least one instance from every positive bag and meanwhile far away from instances in negative bags. To make DD optimization more efficient, Zhang and Goldman (2002) combines DD method with Expectation–Maximization (EM) to speed up the optimization process. In Wang and Zuker (2000), modifys the k-nearest neighbor (KNN) algorithm to solve the MIL by adopting the modified Hausdorff distances.

Motived by the success of support vector machine (SVM), many researchers have presented different methods to make SVM suitable for MIL problem. Among these methods, Andrews (Stuart et al., 2003) reformulates the support vector machine (SVM) with label and bag style constraints, and further propose two MIL algorithms, mi-SVM and MI-SVM. Observing that the MIL can be

☆ This paper has been recommended for acceptance by M.A. Girolami.

* Corresponding author. Tel.: +86 0139 91898078; fax: +86 29 88788315.
E-mail address: lizhan@nwu.edu.cn (Z. Li).

0167-8655/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.patrec.2013.11.013

treated as a special form of semi-supervised learning after training bags rearrangement, Zhou and Xu (2007) packs all the instances together according to their bag labels, and puts forward the MissSVM algorithm. Unlike (Zhou and Xu, 2007) that solve MIL through modifying the SVM optimization constraints, some researchers attempts to construct proper MIL kernels. For example, Gartner (Thomas et al., 2002) computes the statistics of bags, and uses them to construct MIL kernel. Inspired by the fact that the bag label probabilistically depends on the instance label, Kwok and Cheung (2007) computes the distances between multiple DD concepts and bags as probability variable and proposes a marginal kernel. Taking the similar idea, Wang et al. (2008) divides MIL applications into drug and image models and further presents a PPMM kernel. Considering the instances in same bag should not be independent and identical distribution, Zhou et al. (2009) maps bags to an undirected graphs and constructs two kernels, MIGraph and miGraph. To generate a new space suitable for representing MIL bags, Chen proposes two algorithms DD-SVM (Chen and Wang, 2004) and MILES (Chen et al., 2006) to select the important instances either by multiple DD concepts or 1-norm SVM.

Inspired by three existing MIL characteristics that we will disscuss in the next section, in this paper, we present two novel MIL algorithms, CK_MIL and ck_MIL, to fully utilize the advantages of these three characteristics. We firstly employ k-means clustering algorithm on positive and negative instance space to obtain the clustering centers, which are further used to select the most positive instances in bags. Next, we compute the minimum, maximum and average distances from bags to positive cluster centers as bag structures. Finally, for the CK_MIL algorithm we combine the most positive instance and bag structure together to represent the bag, whose similarity is computed by using radial basis function (RBF) kernel. For ck_MIL, we construct a multiple kernel by using a probabilistic coefficient to balance positive instances and bag structures' similarity. We summarize our contribution as follows.

- We adopt the K-means clustering algorithm to select the target concept from the positive instance training set. In this way, ambiguous problem of MIL is efficiently solved. Unlike the APR (Dietterich et al., 1997) and DD (Maron and Lozano-Pérez, 1998) methods which are sensitive to noise, the CK_MIL and ck_MIL algorithms easily select the most positive instances in bags and have strong robustness for our method is modeled on positive and negative instance space seperately.
- Note that the bag label mainly depends on the positive instance's appearance and the other instances in bag can improve the prediction accuracy, in ck_MIL algorithm we introduce a probabilistic coefficient to balance the influence of the most positive instance and bag structure, which accurately represents the essence of MIL learning problem.
- Through analysis the relationship between ck_MIL and multiple kernel learning, we propose a novel multiple kernel MIL algorithm, which can be further used to different kernel classification methods, such as SVM or Gaussian Processes.

The organization of this paper is as follows: Section 2 gives the analysis of existing MIL algorithms we have mentioned in Section 1. Section 3 introduces the CK_MIL and ck_MIL algorithms. Experimental results and analysis are presented in Section 4. Section 5 contains our conclusion and future work.

2. Analyzing MIL algorithms

To understand the characteristics of MIL and help us to design a more efficient MIL method, we will review and analyze the MIL methods listed in Section 1. Let us firstly turn back to the two most important MIL algorithms, APR (Dietterich et al., 1997) and DD (Maron and Lozano-Pérez, 1998; Zhang and Goldman, 2002), and find what they have in common. The key idea of APR is to begin with a shrinking or expanding hyper-rectangle, and to represent a region that contains at least one positive instance of each positive bag. Unlike the APR method that locates a dense positive region by a hyper-rectangle, DD method attempts to find the maximum DD point whose ellipse region contains instances of positive bags, while is far away from instances of negative bags. Noting the DD method's success, many researchers (Chen and Wang, 2004; Kwok and Cheung, 2007) follow the similar idea and adopt the extension of DD to solve MIL. It is not difficult to conclude that both APR and DD imply positive instance cluster characteristic, because both of them find a dense positive cluster region in instance space. The MIL graph kernels, most recently proposed by Zhou et al. (2009), attempt to acquire the special structure of bag, and utilize the similarity of bag structures to solve MIL problem. The results of this method on different data sets have proved that indeed there is some potential identified structure in MIL bags. Both marginal kernel (Kwok and Cheung, 2007) and PPMM kernel (Wang et al., 2008) consider that the bag label is probabilistically influenced by instance label. The difference between them is that the marginal kernel introduces a coefficient for every instance of bag, while PPMM directly classifies the MIL problem into two probability models, drug mode and image mode. Now, through above analysis, it is obvious that the MIL holds three characteristics: positive instance clustering, bag structure and instance probabilistic influence to bag label.

Motived by these three characteristics' success in MIL problem, the nature idea is to combine them together to find a more efficient MIL method. Thus, we advocate here to integrate these three characteristics into a MIL framework, and propose CK_MIL and ck_MIL algorithms to solve MIL problem.

3. CK_MIL and ck_MIL algorithms

3.1. Overview of our method

To describe our method, we first introduce some notations. Let $T = \{(B_1, y_1), (B_2, y_2), \ldots, (B_S, y_S)\}$ denote the training set consisting of S bags including P positive and Q negative, where $y_i \in \{-1, +1\}$ is the labels of positive and negative bag, and $B_i = \{x_{i1}, x_{i2}, \ldots, x_{in_i}\}$ is a collection of n_i instances, each instance $x_{ij} \in \mathrm{R}^d$ is a d-dimensional feature vector. The objective of MIL is to learn a classification function that can accurately predict the label of any unseen bag. For the sake of convenience, we line up all the instances in every positive training bag together, and re-index them as $T^+ = \{x_i | i = 1, 2 \ldots, r^+\}$. Here $r^+ = \sum_{i=1}^{P} |B_i|$ is the total number of instances within positive training bags. Taking the similar way, another set $T^- = \{x_i | i = 1, 2 \ldots, r^-\}$ is obtained, where $r^- = \sum_{i=P+1}^{P+Q} |B_i|$ is the total number of instances within negative training bags.

It is well known that the traditional kernel function, such as RBF, can only handle singleton sample similarity problem, which makes it can't be applied directly to MIL problem for the MIL bag is a set not a singleton example. However, if we convert MIL bag to a fix-length vector, the bag similarity can be transformed to one-to-one similarity problem, and the traditional kernel function can be used. Along this idea, our MIL solution is to find a bag mapping function $K(B_i, B_j; \pi) = \{(x_{i1}, x_{\pi_j 1}), \ldots (x_{ik}, x_{\pi_j k})\}$, where π maps the bag B_i, B_j to a same dimension space and make their similarity be a one-to-one similarity problem. Here the B_i and B_j represent two different bags and x_{i1},$x_{\pi_j 1}$ are one-to-one similarity matching instances in these two bags.

We use Fig. 1 to illustrate our one-to-one MIL matching idea, where the task is to classify the horse and non-horse images. If