Pattern Recognition ∎ (∎∎∎∎) ∎∎∎–∎∎∎

Contents lists available at ScienceDirect

Pattern Recognition

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

Diversified dictionaries for multi-instance learning

Maoying Qiao [a], Liu Liu [a], Jun Yu [b,*], Chang Xu [c], Dacheng Tao [c]

[a] The Centre for Quantum Computation and Intelligent Systems, The Faculty of Engineering and Information Technology, University of Technology Sydney, Sydney, Australia
[b] The Key Laboratory of Complex Systems Modeling and Simulation, School of Computer Science and Technology, Hangzhou Dianzi University, Hangzhou 310018, China
[c] The Centre for Quantum Computation and Intelligent Systems, the Faculty of Engineering and Information Technology, University of Technology Sydney, 81 Broadway Street, Ultimo, Sydney, NSW 2007, Australia

ARTICLE INFO

Article history:
Received 29 May 2016
Received in revised form
28 July 2016
Accepted 22 August 2016

Keywords:
Multi-instance learning
Diversified learning
Dictionary learning

ABSTRACT

Multiple-instance learning (MIL) has been a popular topic in the study of pattern recognition for years due to its usefulness for such tasks as drug activity prediction and image/text classification. In a typical MIL setting, a bag contains a bag-level label and more than one instance/pattern. How to bridge instance-level representations to bag-level labels is a key step to achieve satisfactory classification accuracy results. In this paper, we present a supervised learning method, diversified dictionaries MIL, to address this problem. Our approach, on the one hand, exploits bag-level label information for training class-specific dictionaries. On the other hand, it introduces a diversity regularizer into the class-specific dictionaries to avoid ambiguity between them. To the best of our knowledge, this is the first time that the diversity prior is introduced to solve the MIL problems. Experiments conducted on several benchmark (drug activity and image/text annotation) datasets show that the proposed method compares favorably to state-of-the-art methods.

1. Introduction

Multi-instance learning (MIL) is a variant of a traditional learning setting. Rather than one instance with one label, MIL accepts one bag with one label as its input, where one bag usually contains more than one instance/pattern. A positive bag is dominated by one or more positive instances, while a negative bag has no positive instances within it. This MIL framework originally arises from drug activity prediction [1] and now is necessitated by more complex situations. For instance, weak bag-level labeling information can be easily obtained from the Internet, while the accurate instance-level label annotation is expensive and time-consuming [2–4]. Another complex situation is that a bag's label depends on its constituent instances and therefore may not be consistent because of different semantic meanings among the instances [5]. Due to the high demanding for it, MIL has been studied as it applies to stock market prediction [6]; text categorization [7,8]; image processing (natural scene classification [9–11], content-based image retrieval [12–14], and image categorization [15–19]); etc.

There are two main methods for adapting MIL to mature one-instance learning algorithms by bridging instance-level representations to bag-level labels. The first one focuses on resolving the obstacle caused by missing instance-level labels. Many effective MIL algorithms belong to this category. For example, miSVM [7] treats instances in positive bags as unknown integer variables and learns max-soft-margin hyperplanes by jointly exploiting possible label assignments. Rather than assuming at least one dominating instance in a positive bag, MIBoosting [20] treats instances independently and equally and employs a boosting algorithm to do MIL classification. In contrast, miGraph [21] supposes instances are non-i.i.d. and regards a bag as an undirected graph to do MIL learning by exploiting relations among instances. These algorithms achieve decent accuracy rates among various MIL applications.This result is partially attributable to the bag representations' preservation of the information contained in all instances. At the same time, however, this cumbersome bag representations increase the computation burden for classification.

The second method mainly focuses on constructing bag-level representations via codebooks constructed from instances, specifically in an unsupervised manner. For example, MILES [22] simply treats each instance from a training bag as an entry of a codebook and re-represents each bag via similarities between the codebook and bag instances. To reduce redundancy of the codebook, a ℓ_1-SVM is trained for classification. Instead of reducing codebook

* Corresponding author.
E-mail addresses: Maoying.Qiao@student.uts.edu.au (M. Qiao), Liu.Liu@student.uts.edu.au (L. Liu), yujun@hdu.edu.cn (J. Yu), changxu1989@gmail.com (C. Xu), Dacheng.Tao@uts.edu.au (D. Tao).

http://dx.doi.org/10.1016/j.patcog.2016.08.026
0031-3203/

redundancy in the classifier training stage, CCE [23] does it in the codebook constructing stage. It first groups instances into clusters and then re-represents each bag with a binary vector, and finally, a classifier ensemble is employed to carry out the classification task. Similarly, miVLAD [24] applies the same clustering step to form a codebook, but it employs a different mapping function to project a bag into a feature vector. Thanks to the efficacy of the transformed feature vector, even linear classifiers can achieve respectable accuracy. In summary, the algorithms from this category accomplish an efficient transformation from instance representations to bag-level vector representations and produce succinct inputs to traditional one-instance classification algorithms. They may, however, lose pattern information during the transformation, and useful label information is dropped when learning a codebook.

To embed label information into a codebook learning procedure, in this paper, we propose a diversified dictionaries scheme for MIL. Our main contributions are as follows: (i) We propose a diversity regularizer among a set of dictionaries, based on the knowledge of determinantal point process (DPP). By defining a similarity measure between dictionaries, we show that the determinant of this similarity matrix geometrically encourages diversity among the dictionaries. (ii) We propose a diversified dictionaries learning scheme for MIL. Similar to dictionary-based MIL [2], bag-level labels are utilized to conduct class-specific dictionary learning. To reduce overlap among dictionaries, and at the same time, to increase the discrimination of those dictionaries, we combine the proposed diversity regularizer to encourage dissimilarity between pairwise dictionaries. As a consequence, each class is accompanied by a unique dictionary. Because a more diverse dictionary set tends to cover a larger representation space, our proposed method is a more generalized model. (iii) The proposed model is solved by a coordinate ascent method combined with an active set algorithm and a proximal gradient method. (iv) Our method achieves satisfactory results in terms of classification accuracy over several benchmark datasets.

This paper is organized as follows. Related works on MIL are described in Section 2. The proposed model and its solution are presented in Sections 3 and 4, respectively. Experimental results are shown in Section 5, and finally, Section 6 concludes this paper with a brief discussion.

2. Related work

In recent years, various works related to multi-instance learning have appeared in the literature. We refer interested readers to [25–28] for refined MIL-related reviews. Here, in terms of how to cope with the gap between instance-level representations (patterns) and bag-level labels, we roughly divide these works into two complementary categories, namely algorithms transforming instance-level representations to bag-level representations and algorithms transferring bag-level labels to instance-level labels.

Efficient and effective algorithms pertaining to the first category include MILES [22], CCE [23], and miVLAD/miFV [24,29]. These algorithms have similar instance-level pre-processing to extract bag-level representations later by mapping bags into instance-based codebooks. MILES and CCE differ in terms of the codebook's formation, as the former one collects all instances as atoms of a codebook, while the latter groups instances into clusters and treats each cluster as one single atom of a codebook. Although both CCE and miVLAD do the same clustering pre-processing over instances, they employ different mapping functions: CCE maps bags into binary vectors, while miVLAD maps bags with a similarity measure. Finally miVLAD and miFV can encode more information than the above-mentioned algorithms, and thus, both produce excellent results. Coupling the obtained bag-level representations with bag-level labels, any classifiers can be employed to do bag-level classification tasks. However, it is easily seen that no available label information is exploited during the codebook learning, which may reduce classification performance.

Algorithms falling into the second category aim to transfer bag-level labels to instance-level labels and then conduct classification over instance pairs of representations and labels. Most of the algorithms in this category are built on the concept of diverse density (DD) [9,30]. DD at a point defines a measure of how near instances within positive bags are to it and how far instances within negative bags are from it. In other words, DD is used to find instances that are shared among positive bags but do not exist in negative bags. Thus, it is natural that the common instances inherit labels of positive bags. Based on this idea, different algorithms attempt to integrate DD to develop new schemes for MIL. The EM-DD algorithm [31], as its name indicates, finds responsible instances for different labels. MILES [22] employs DD to choose 'concepts' and use them as an embedded feature space to obtain bag-level representations. MIBoosting [20] alters the original Gaussian-like instance density distribution into an independent and equal contributed instance assumption and combines it into training a boosting classifier for MIL. GD-MIL [32,2] combines DD with dictionary learning and learns class-specific dictionaries in a supervised way. They implicitly weight each instance's contribution to the bag label via DD distribution. As has been mentioned and emphasized above, these learned dictionaries for different classes may be quite similar to one another. Other supervised MIL algorithms include [33–36].

3. Model

Given a data set $\{(X_i, l_i)\}_{i=1}^N$, where $X_i = \{x_{i1}, \ldots, x_{ij}, \ldots, x_{in_i}\}$ represents one bag containing n_i instances, and each instance x_{ij} is a D-dimensional feature representation, namely, $x_{ij} \in \mathcal{R}^D$. The symbol $l_i \in \{1, \ldots, C\}$ denotes the bag-level label for bag X_i, and N is the number of bags. MIL aims to learn a classifier to predict labels for unseen bags. Before training the classifier, bag-level feature representations $\{Y_i\}_{i=1}^N$ for all bags are needed to be learned.

In this section, we first review dictionary learning for traditional one-instance learning. Then, we show how the dictionary learning method has been extended to MIL. Finally, we present our proposed diversified dictionaries for MIL.

3.1. Review of instance-level dictionary learning

Instance-level dictionary learning is where the instance number of every bag is exactly equal to 1, i.e., $j=1$, for all $\{x_{ij}\}_{i=1}^N$. We learn a dictionary $\mathcal{D}^c \in \mathcal{R}^{D\times N_c}$ for each class c, where N_c is the atom number of dictionary for class c and D, as described above, is the dimension for instance-level representations. Correspondingly, symbols $\{\mathcal{D}^c_{\cdot n_c}, n_c = \{1, \ldots, N_c\}\}$ denote the n_c-th atom of dictionary for class c. We collect all instances belonging to class c together and symbolize it as $X^c = \{x_{ij} | x_{ij} \in X_i, l_i = c, j = 1\}$. Their corresponding sparse coding representations are denoted as $Y^c = \{y_{ij} | y_{ij} \in \mathcal{R}^{N_c}, l_i = c, j = 1\}$, each instance of which is denoted as $y_{ij}{}^c$.

In dictionary learning, each instance should be reconstructed by sparsely combining atoms of a given dictionary. In other words, The objective of dictionary learning is to minimize reconstruction errors, or, equivalently, to maximize their corresponding energy functions. Formally, the energy function for each instance is defined as

$$E(\mathcal{D}^c, y_{ij}^c) = \exp(-\sigma \| x_{ij} - \mathcal{D}^c y_{ij}^c \|_2^2), \tag{1}$$