Information Sciences 467 (2018) 559–578

Contents lists available at ScienceDirect

Information Sciences

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

Bag encoding strategies in multiple instance learning problems

Emel Şeyma Küçükaşcı[a,*], Mustafa Gökçe Baydoğan[b]

[a] Department of Industrial Engineering, Istanbul Commerce University,İstanbul 34840, Turkey
[b] Department of Industrial Engineering, Boğaziçi University, İstanbul 34342, Turkey

article info

Article history:
Received 30 October 2017
Revised 4 August 2018
Accepted 8 August 2018
Available online 10 August 2018

Keywords:
Multiple instance learning
Classification
Bag encoding
Decision trees

abstract

Multiple instance learning (MIL) deals with supervised learning tasks, where the aim is to learn from a set of labeled bags containing certain number of instances. In MIL setting, instance label information is unavailable, which makes it difficult to apply regular supervised learning. To resolve this problem, researchers devise methods focusing on certain assumptions regarding the instance labels. However, it is not a trivial task to determine which assumption holds for a new type of MIL problem. A bag-level representation based on instance characteristics does not require assumptions about the instance labels and is shown to be successful in MIL tasks. These approaches mainly encode bag vectors using bag-of-features representations. In this paper, we propose tree-based encoding strategies that partition the instance feature space and represent the bags using the frequency of instances residing at each partition. Our encoding implicitly learns generalized Gaussian Mixture Model (GMM) on the instance feature space and transforms this information into a bag-level summary. We show that bag representation using tree ensembles provides fast, accurate and robust representations. Our experiments on a large database of MIL problems show that tree-based encoding is highly scalable, and its performance is competitive with the state-of-the-art algorithms.

1. Introduction

Classification, one of the important class of supervised learning problems, vastly takes place in data mining tasks. In traditional classification tasks, each object is represented with a feature vector, and the aim is to predict the label of the object given some training data. However, this representation is not flexible when the data has a certain structure. For example, in image classification, images are segmented into patches and instead of a single feature vector, each image is represented by a set of feature vectors derived from the patches. This way, important information regarding the certain invariances such as location and scale can be taken into account [4]. Change of object representation provides benefits for a wide range of applications such as bioinformatics [15], document retrieval [3], computer vision [18] and etc. This type of applications fits well to Multiple Instance Learning (MIL) setting where each object is referred to as bag and each bag contains certain number of instances.

Most of the MIL approaches generally solve the binary classification problem, where bags are labeled as either positive, or negative [15,27,45,47]. The firstly described formal MIL problem is a drug activity prediction problem, which considers molecules as bags and distinct shapes of the same molecule as instances [15]. A molecule is positively labeled if it includes at least one effective shape, otherwise it is negatively labeled. In text categorization problems [50], each document can be considered as a bag and its instances are the collection of relevant passages inside it. In all these applications, training bags are labeled and instances belonging to each bag do not necessarily have labels. The aim of MIL is to learn a classifier on the training bags to predict the label of a test bag.

Ambiguity about the instance labels has made researchers focus on certain assumptions regarding the instance labels. The so called standard MIL assumption is given as: if a bag is labeled positive, then at least one instance in that bag is labeled as positive; otherwise, labels of all instances in

* Corresponding author.
E-mail addresses: eskucukasci@ticaret.edu.tr (E. Şeyma Küçükaşcı), mustafa.baydogan@boun.edu.tr (M. Gökçe Baydoğan).

https://doi.org/10.1016/j.ins.2018.08.020
0020-0255/

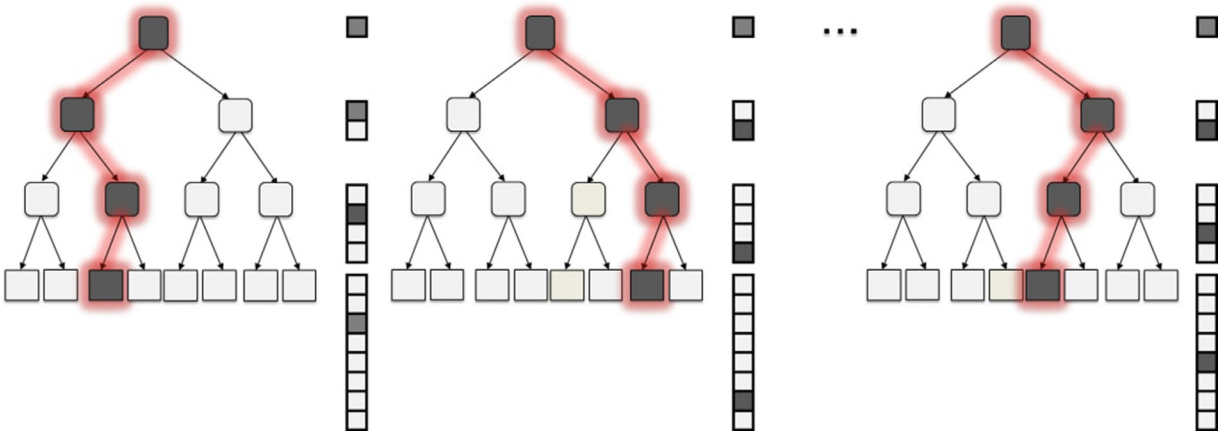

Fig. 1. An example of path-encoding.

negative bags are negative [15]. It is obvious that when a bag is known to be positively labeled, the labels of its instances are not completely known. Standard MIL assumption is too restrictive to handle real-life problems. For example, optimal combination therapy is used in cancer treatment to overcome drug resistance. An optimal combination of drugs is considered to be capable of circumventing drug resistance among individual patients. Since there exists enormous number of possible drug combinations, the prediction problem of optimal combinatorial therapy can be modeled as a MIL problem where drugs are the instances, and collections of drugs are the bags. A bag is positive if a subset of its instances forms an effective drug combination, otherwise the bag is negative. Optimal combination therapy discovers an effective combination of drugs, rather than identifying a single type of drug that supports the treatment. Instead of a single positive instance, this MIL problem searches for a combination of multiple instances referring to various drugs.

Criticizing the potential problems with the standard MIL assumption, MIL problems are categorized as presence-based, threshold-based and count-based MIL problems [43]. A specific region of feature space where the positive instances are located are referred to as a concept by Weidmann et al. [43]. Presence-based MIL has the standard MIL assumption for multiple concepts, whereas threshold-based MIL forces a lower bound on the number of necessary instances of each concept. Finally, in addition to the previous lower bound, count-based MIL also requires an extra upper bound on the number of necessary instances from each concept. Extensions and variations of the described categorization of generalized MIL problems are also presented in [1,2,19]. Based on the experiments on synthetic and real datasets following different assumptions, the bag-level classification is indicated to be successful on datasets from different categories [1]. These approaches require each bag to be represented with a feature vector that will summarize the instance level information. Since bag-level methods are competitive, we focus on bag classification by representing each bag with a single feature vector in this study.

Earlier, many approaches from the computer vision literature utilized the well-known Bag-of-Features (BoF) or Bag-of-Words (BoW) representations to perform similar tasks. After clustering the patches (i.e. instances), the image (i.e. bag) is represented by the frequency of cluster assignments of the corresponding instances in the simple BoW setting [12]. These approaches implicitly transform the instance-level probabilistic distribution information to a bag-level summary [27,42]. Recently, [10] has approached the problem by considering the geometric view of the instance space and obtain a bag-level summary using the similarities between the instances. Motivated by the success of the bag-level representations and their robustness to the MIL assumptions, this study proposes bag encoding strategies for MIL problems. Fig. 2 presents a summary of the bag representation algorithms, each of which will be discussed in detail in Section 4.

Most of the existing proposals to obtain bag-level summary require numerical features as an input since they involve transformations such as principal component analysis [42], density estimation [27] or distance calculations [10,42]. However, a MIL dataset can have features other than numeric. When there are categorical features, dummy variables are required to be introduced. Moreover, standardization/normalization is required but standardization of the dummy variables introduced to represent categorical variables is not well-defined. Hence, an approach that can treat each variable without any modification may be required for certain applications. Considering this fact, our approach utilizes tree-based ensembles to partition the instance feature space. A tree learner trained on the raw data assigns each instance to a terminal node of the tree.

Use of trees for feature induction is a relatively new research direction, which is also named as hashing [40]. This method transforms each node in the tree to a feature. Moreover, the new representation is easy to be modified by changing the tree parameters. Each level of the tree provides a different partition of the instance feature space as they imply simple splitting rules on the features. An instance traverses the tree based on the splitting rules (i.e. follows a path in the constructed tree). The path followed by an instance implies regions of the feature space an instance belongs to and it provides an hierarchical information regarding the feature space an instance resides. Tree-based encoding of the feature space does not require scaling of the data as opposed to the approaches requiring distance calculations or density estimation. Fig. 1 illustrates the path-encoding of an instance. Next to each tree in Fig. 1, the traversed path by an instance is detected, and a binary vector is encoded conceiving whether a node is on that path, or not. Thus, these paths can be used to learn a BoW type representation. Our approach inherits the properties of tree-based learners. That is, it can handle numerical or categorical data. Besides, tree-based encoding is scale invariant and robust to missing values. The same tree can be used to encode the instances based only on terminal nodes. Earlier, Moosmann et al. [29] used a similar strategy for image classification problems