Journal of Theoretical and Applied Computer Science
ISSN 2299-2634 (printed), 2300-5653 (online)
Vol. 7, No. 3, 2013, pp. 29–39
http://www.jtacs.org

Future directions in Multiple Instance Learning

David Chiu[1], Iker Gondra[2], Tao Xu[1]

[1] School of Computer Science, University of Guelph, Guelph, Canada
[2] St. Francis Xavier University, Antigonish, Canada

xut@uoguelph.ca

Abstract: *In Multiple Instance Learning, each training sample consists of a set of unlabelled instances. The set as a whole is labeled positive if at least one instance in the set is positive, or negative otherwise. Given such training samples, the goal is to learn either an explicit description of the common positive instance(s) or a bag classifier that can assign labels to bags. Previous research has focused on this standard definition of the problem where instances in a set are independent. This raises a question: if we remove the independence assumption, can we generalize the goal of finding a description of the common instance(s) to that of finding a description of the common pattern(s) among instances? Similarly, can we generate bag classifiers that discriminate based on common pattern(s) among instances instead of just common instance(s)? This question raises many other related questions that have not been yet fully explored in the context of this problem. In this paper we first present a survey of existing methods that work with the standard definition of the problem and then elaborate on the previous question in the hope that researchers will investigate this exciting research direction.*

Keywords: *machine learning, pattern recognition, multiple instance learning, total entropy, partial entropy*

1. Introduction

Multiple Instance Learning (MIL) refers to a special kind of supervised learning problem that aims to discover the concept, i.e., what it is that results in a particular class label such as "positive", from collectively labeled data. Generalized to the standard supervised classification where every input has a corresponding label, MIL deals with data of which a collection (or bag) of inputs (or instances) is assigned with a single label. The labeling of MIL data amounts to the relevance of data with respect to the concept being sought. A bag is labeled positive if at least one instance in that bag is relevant to the concept or negative otherwise. The relevance can be numerically assessed in a metric space with a criterion such as a distance-based measure. Given a set of such labeled bags, the goal of MIL is to identify the concept that accounts for the labeled data; predictions can be made on new samples accordingly.

The importance of MIL is reflected through a variety of applications such as, e.g., drug behavior analysis [1], drug discovery [2], content-based image retrieval [3, 4, 5, 6, 7], supervised image segmentation [8], object recognition [9, 10, 11], automatic object tracking [12], web mining [13], text categorization [14], biological data analysis [15, 16]. Previous MIL research has focused on this standard definition of the problem where instances in a set are assumed to be independent. However, it could be that what generates the positive labelling of a bag is the inclusion of a particular pattern among its instances, e.g., a subset of its instances following a particular spatial arrangement. In such situations, the mere discovery of common instances among positive bags may not solve the problem since what is common are not indi-

vidual instances but rather multi-instance patterns. This observation has prompted us to think about the following question: if we remove the independence assumption, can we generalize the goal of MIL from the one of finding a description of the common instance(s) to that of finding a description of the common pattern(s) among instances? Similarly, can we generate bag classifiers that discriminate based on common pattern(s) among instances instead of just common instance(s)? The remaining of this paper is organized as follows: we first present a survey of existing methods that work with the standard definition of MIL. Then, we elaborate on the previous question. Some concluding remarks are given at the end.

2. Survey

Depending on whether the concept is explicitly given as the result of learning, existing methods can be classified into two basic types: *MIL concept learners*, include methods that produce an explicit description of the common concept(s), i.e., instances, among the positive bags; *MIL classifiers*, include algorithms that generate a classifier that can be used to classify, i.e., label, new bags. The concept learners provide an abstract view of the domain knowledge, which could further aid the design of domain-specific solutions, whereas the classifiers avoid the intermediate step of exploiting the compositional and structural constitution of the concept(s) and instead generate a decision function for classification of new bags. Although concept learning spans a broad problem domain, in this survey, the focus is made on the MIL problem in a metric space where distance measures are commonly adopted to support the commonality analysis among identically labelled samples.

2.1. Concept learners

2.1.1. Axis-parallel rectangles

In the axis-parallel rectangle (APR) approach [1], the concept is confined to the smallest APR that includes at least one instance from each positive bag while excludes all instances from the negative bags. To find such a rectangle, the execution order could be either "outside-in" or "inside-out". In the former case, the algorithm starts with a bound that includes all the instances in the positive bags and shrinks it until all the false positive instances are excluded. The latter case starts with an instance in a positive bag and grows it until the smallest APR is obtained. Obviously, the smallest APR gives a representation of the target concept. Any feature vector that falls inside the APR can be thought of as a realization of it.

Theoretical analysis of APR has been carried out under the *probably approximately correct* (PAC) framework [17, 18, 19]. The PAC learnability of APR mainly concerns the statistical conditions such as the sample size, distribution and independence relation between the instances for an algorithm to be able to find the smallest APR. It is shown in their study that finding such an APR is NP-hard.

2.1.2. Diverse density

Diverse density (DD), a probabilistic model of MIL describes the concept as a region that is highly dense of diverse positive bags while sparse of negative ones. Let $B_i^+ = \{B_{i,1}^+, \ldots, B_{i,l}^+\}$ be the i^{th} positive bag, and $B_{i,j}^+ \in \Re^n$ be the j^{th} instance in that bag, and the k^{th} feature for that instance is denoted as $B_{i,j,k}^+$. A negative instance $B_{i,j}^-$ is similarly defined. Assuming that in the entire instance space there is a unique representation of the target

concept centered at $\mathbf{t} \in \Re^n$ that determines all the bag labels, $\mathbf{t}$ can be located by examining a point $\mathbf{x}$ that maximizes the likelihood [3]

$$h_{ML} = \arg\max \prod Pr(B_i^+|\mathbf{x} = \mathbf{t}) \prod Pr(B_i^-|\mathbf{x} = \mathbf{t}).$$

conditioned on the independence between bags given the target concept $\mathbf{t}$. Assuming a uniformly distributed probability over the target concept, by applying Bayes's rule, this becomes

$$h_{ML} = \arg\max \prod Pr(\mathbf{x} = \mathbf{t}|B_i^+) \prod Pr(\mathbf{x} = \mathbf{t}|B_i^-),$$

where

$$Pr(\mathbf{x} = \mathbf{t}|B_i^+) = Pr(\mathbf{x} = \mathbf{t}|B_{i,1}^+, \ldots, B_{i,l}^+) = 1 - \prod_j (1 - Pr(\mathbf{x} = \mathbf{t}|B_{i,j}^+)).$$

and

$$Pr(\mathbf{x} = \mathbf{t}|B_i^-) = Pr(\mathbf{x} = \mathbf{t}|B_{i,1}^-, \ldots, B_{i,l}^+) = \prod_j (1 - Pr(\mathbf{x} = \mathbf{t}|B_{i,j}^-)),$$

based on the assumption that the failing of one cause is independent of others. Given

$$Pr(\mathbf{x} = \mathbf{t}|B_{i,j}) = \exp(-\left\|B_{i,j} - \mathbf{x}\right\|^2),$$

the estimate of $\mathbf{t}$ is nothing but the mean of an un-normalized Gaussian distribution. With an exponential function as the probability measure, the diverse density at the intersection of n bags is exponentially higher than at the intersection of $n-1$ bags [3].

Based on DD, Zhang and Goldman [4] used hidden variables to model the unknown instance labels. They proposed an iterative expectation-maximization(EM) process to refine the maximum likelihood estimate. In each E-step, the most-likely-cause positive instance in each bag is selected and used in estimating the new target concept in M-step. EM-DD largely reduces the computational burden of MDD and achieves a remarkable performance improvement on the "MUSK" datasets.

2.1.3. Adaptive kernel diverse density estimate

Adaptive kernel diverse density estimate (AKDDE) implements the notion of diverse density for each type of bags using density estimate techniques. It contrasts the likelihood estimate of DD [3]. Based on the intuition that the target concept of MIL falls into the region that is dense of diverse positive bags, one can define diverse density as the *probability density function of diverse positive bags*, then the AKDDE of N diverse positive bags is written as

$$\hat{f}_{\mathcal{B}^+}(\mathbf{x}) = \frac{c^+}{N} \sum_{i=1}^{N} \frac{1}{h_i^d} K\left(\frac{\mathbf{x} - B_{i,nrst}^+}{h_i}\right), \tag{1}$$

where N is the number of the positive bags (i.e.,. cardinality of $\mathcal{B}^+$), $B_{i,nrst}$ denotes the nearest instance in B_i to $\mathbf{x}$, c^+ is the normalization constant, and h_i is the maximum instance-bag distance from $B_{i,nrst}$ to all $N-1$ positive bags excluding B_i itself, i.e.,

$$h_i = \max_{B_j \in \mathcal{B}^+; i \neq j} \{dist(B_{i,nrst}^+, B_j)\}, \tag{2}$$

given the (Euclidean) instance-bag distance defined as

$$dist(\mathbf{x}, B_i) = \min_{B_{i,j} \in B_i} \{\|\mathbf{x} - B_{i,j}\|_2\}. \tag{3}$$

In short, Eq.1 together with Eq.2 and Eq.3 states that the AKDDE of positive bags at $\mathbf{x}$ is locally decided by the bound that contains the nearest instance from each positive bag. The AKDDE for negative bags is similarly defined. As the concept is highly concentrated of positive bags with minimum interference of negative ones, the objective is constructed as the AKDDE of positive bags regularized by the AKDDE of negatives ones. In one of our recent experiments (not yet published), AKDDE demonstrates superiority in learning concepts over other MIL concept learners.

2.1.4. Maximum partial entropy

Using an instance-bag distance as a measure of the degree of an instance belonging to a bag, the partial entropy taking into account all training samples is defined as the amount contributed by the positive bags to the total entropy [20]. Formally, let $Pr(\mathbf{x} \in B_i)$ be the probability of an unknown instance belonging to bag B_i (regardless of the bag's label), the entropy over the entire training data is

$$\begin{aligned} H(\mathbf{x} \in \mathcal{B}) &= H'(\mathbf{x} \in \mathcal{B}^+) + H'(\mathbf{x} \in \mathcal{B}^-) \\ &= -\sum_{B_i^+ \in \mathcal{B}^+} Pr(\mathbf{x} \in B_i^+) \log Pr(\mathbf{x} \in B_i^+) \\ &- \sum_{B_i^- \in \mathcal{B}^-} Pr(\mathbf{x} \in B_i^-) \log Pr(\mathbf{x} \in B_i^-). \end{aligned}$$

Taking away the amount contributed by the negative bags, we are left with the partial entropy contributed by the positive ones:

$$H'(\mathbf{x} \in \mathcal{B}^+) = -\sum_{B_i \in \mathcal{B}^+} Pr(\mathbf{x} \in B_i) \log Pr(\mathbf{x} \in B_i).$$

In a metric space, $Pr(\mathbf{x} \in B_i)$ can be defined inversely proportional to the distance from $\mathbf{x}$ to B_i to reflect our degree of belief that the closer $\mathbf{x}$ to B_i, the more likely it comes from bag B_i. Maximizing the partial entropy results in the reduction of the inter-class uncertainty and the increase of the intra-class certainty. When related to MIL, it is equivalent to locate in the instance space a point that is far from the negative bags while close to all the positive bags simultaneously. Thus, the point that maximize the partial entropy gives rise to the concept being sought.

2.2. Classifiers

2.2.1. Logistic linear models

Logistic regression functions, a widely used method to predict the outcome of a categorical variable $\mathbf{x} \in \Re^d$, is defined by

$$\sigma(\mathbf{x}) = \frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x})},$$

where $\mathbf{w} \in \Re^d$ denotes the linear coefficients of the exponential. It builds a straight connection between a single trial and the probability of the corresponding outcome. Under the MIL framework, $\sigma(\mathbf{x})$ is interpreted as the instance-level probability of an instance being positive governed by the linear coefficient vector $\mathbf{w}$. Since a positive bag contains at least one positive instance, adopting the noisy-OR model [3], the conditional probability for a bag to be positive is thus written as

$$p(y = 1|B_i) = 1 - \prod_{B_{i,j} \in B_i} (1 - \sigma(\mathbf{w}^T B_{i,j})).$$

Similarly, a negative bag contains no positive instance at all. Hence

$$p(y = 0|B_i) = \prod_{B_{i,j} \in B_i} (1 - \sigma(\mathbf{w}^T B_{i,j})).$$

For the training set B, the maximum likelihood estimate of $\mathbf{w}$ is

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} p(B|\mathbf{w}) = \arg\max_{\mathbf{w}} p(B_1^+, \ldots, B_1^-, \ldots|\mathbf{w}).$$

Assuming the statistical independence across all the bags, it is equivalent to maximize the log likelihood

$$\hat{\mathbf{w}} = \arg\max_{\mathbf{w}} \sum [y_i \log p(y_i = 1|B_i) + (1 - y_i) \log p(y_i = 0|B_i)].$$

The aforementioned logistic regression model for MIL was the main idea depicted in [21]. Similar logistic regression models exist, different in the modeling of bag labels: the result of a binomial process of at least one instance label being positive [21]; the result of the aggregation of the instance-level logistic regression using softmax function [22]; or the result of averaging the instance-level probabilities [23].

2.2.2. k nearest neighbours

The method citation k nearest neighbour (kNN) was proposed for MIL in [24]. It is a generalization of the standard kNN by introducing a bag-level distance metric that is defined as the shortest instance-level distance between two bags. Namely, for two bags B_i and B_j, the distance between them is

$$dist(B_i, B_j) = \min_{B_{i,n} \in B_i, B_{j,m} \in B_j} \|B_{i,n} - B_{j,m}\| .$$

This metric is also known as the minimum *Hausdorff* distance.

In addition to the specialized metric, the voting strategy is also adapted for predicting the labels of unseen bags in the training samples. The label of a query bag is predicted not only based on the majority voting of its k nearest bags but also based on the number of times that the bag is cited by them. As an instance-based MIL classifier, citation kNN neither constructs a classification boundary nor explores the structure of the target concept.

2.2.3. Support vector machines

A support vector machine (SVM) constructs a hyperplane that maximizes the between-class margin so as to achieve a minimized generalization error [25]. There are a few variants of SVMs for MIL, different in the way of modeling the labels.

Andrews et al. [14] suggested two adapted SVM models. In mi-SVM, the unknown instance labels are modeled as hidden variables. Let $l_{i,j}$ denote the unknown label for instance $B_{i,j}$, the following constraint can be appended so that the bag labels conform to the definition of MIL:

$$\begin{cases} \sum_j \frac{l_{i,j}+1}{2} \geq 1, & if \quad y_i = 1; \\ l_{i,j} = -1, & if \quad y_i = -1. \end{cases}$$

The maximization of the margin between positive and negative concepts subject to a joint constraint of a kernelized discriminant function and the unknown labels leads to an integer programming problem. The MI-SVM model implements the notion of the maximum margin at bag level. Inspired by the fact that the label of a bag follows the label of the "most positive" instance in that bag, a generalized functional margin is defined as

$$\gamma \equiv y_i \max_{B_{i,j} \in B_i} (\langle \mathbf{w}, B_{i,j} \rangle + b),$$

where $\mathbf{w}$ is the model parameters for a linear SVM. The instance $B_{i,j}$ resulting in the maximum $\langle \mathbf{w}, B_{i,j} \rangle + b$ is thus the "most positive" instance in a positive bag or the "least negative" instance in a negative bag.

Both mi-SVM and MI-SVM end up as an integer programming problem that can be solved heuristically. For mi-SVM, all instance labels are initialized to the corresponding bag labels and iteratively refined in accordance with the definition of MIL. MI-SVM chooses only the most positive instance in each bag to participate into the computation of model parameters.

A different consideration was taken in [26] where bag B_i is represented by a feature vector $s(B_i)$ summarized by operator $s(\cdot)$ from the instances inside the bag. Features are usually the statistics of the bag and the objective is constructed based on the notion of *set kernels*. As for the minimax kernel described in that paper, the feature vector consists of the component-wise minimum and maximum values that a bag extends to in the instance space. For a d-dimensional instance space, the feature vector is of length $2d$, where the i^{th} ($1 \leq i \leq d$) and the $2i^{th}$ features are the minimum and maximum values of the i^{th} feature across all the instances in the bag. It is evident that a large amount of information is lost while wrapping bags into feature vectors.

2.2.4. Decision trees

Building a decision tree involves recursively partitioning samples into subsets based on the information gain of attributes. For a learned tree, conjunctions of attributes at intermediate nodes lead to the class labels at leaves. Information gain is a commonly used criterion that guides the selection of variables as tree nodes. To adapt decision trees for MIL, the information gain is calculated based on the number of unlabelled instances rather than the labelled bags [27]. The consequence is that the instances in a bag might be split into different branches. Given an unseen bag not in the training samples, it is identified as positive if one of its instance reaches a positive leaf, or negative otherwise. To avoid producing complex trees, during the tree construction, once an instance is identified as positive all the rest of instances in that bag are neglected. Noteworthy is that every positive leaf leading a way to the root node gives a representation of the target concept. Thus, the degree of quantization of every feature becomes important that determines the scope of the target concept. A high granularity of the quantization gives a precise description of the target concept at the cost of the increased complexity of the tree and thus the increased computational time. Due to this reason, decision trees may be preferable for MIL classification.

RELIC [28] is another variant of decision trees for the MIL problem. The frequency of the occurrence of attribute values are calculated at instance level. The construction of the tree is exactly the same as in the single-instance scenario based on the C4.5 entropy measure.

An ensemble of decision trees contrasts the single-tree methodology in that a decision is made based on the overall voting of a group of participating trees. The adaption of the ensembled trees for MIL was proposed in [29]. As to preserve the diversity of a learned tree, labels of instances are recovered via a non-convex optimization process subject to the constraint that at least one instance in each bags comes from the target concept.

2.2.5. Artificial neural networks

An artificial neural network (ANN) is a structural representation that can express an arbitrarily complex non-linear function. In the classic feedforward ANN, the error for backpropagation learning is defined as

$$E = \sum_{i=1}^{N} (f(\mathbf{x}_i, \boldsymbol{\theta}) - y_i)^2,$$

where $f(\cdot)$ is the network function governed by a parameter vector $\boldsymbol{\theta}$ that maps an input $\mathbf{x}$ to the corresponding output y. A trained network has the minimum overall error on N samples with the optimized $\boldsymbol{\theta}$.

In the adaption of the neural network for MIL in [30], the error function is defined as

$$E = \sum_{i=1}^{N} (\max_{B_{i,j} \in B_i} f(B_{i,j}, \boldsymbol{\theta}) - y_i)^2.$$

Thus, at each iteration of the learning process, only the "most positive" instance in each positive bag and the "least negative" instance in each negative bag are involved in the updating process of the network parameters. This formulation coincides with the MI-SVM [14], where only the "most positive" instances participate in the determination of the maximum between-class margin. When a new bag is fed to the trained network, it is classified as positive if any of its instances is predicted as positive.

2.2.6. Boosting

In supervised learning, boosting is a meta-algorithm that ensembles a set of weak learners to create a strong learner [31]. The idea is inspired by the fact that every weak learner could produce correct outputs on portion of the data, and combining the correctness of a number of such weak learners can thus enhance the overall performance. In boosting, weights are associated with individual training samples to reflect the correctness of a classifier, and adjusted during the training process. Algorithms mainly differ in the strategy of adjusting weights based on the previously misclassified samples.

Implementations of boosting have been extended for MIL. MILBoost [11] was proposed based on the Anyboost framework [32]. The objective is a log likelihood function derived from the noisy-OR probabilistic model. A similar boosting method based on RealBoost framework [33] was proposed in [34].

3. Future directions

As a specialized supervised learning paradigm, MIL aims to discover the concepts underlying the collectively labelled data so as to restore the labelling mechanism. The concepts in

some applications are of paramount importance not only because they are abstracts of domain knowledge or experiences but also support predictive machine computing. For example, in image understanding related problems (e.g., semantic image segmentation and content-based image retrieval), a deep insight into the constitution of visual components of a particular object allows us to identify those discriminative features so as to design more discriminative models. In the drug activity testing problem, identifying the 3D molecular structure is critical to activate or deactivate the target molecules. Regarding MIL concept learning, questions arise in (a) how concepts to be modeled? (b) How to classify samples based on the learned concept? The existing MIL concept learners have attempted representing the concept with the smallest geometry or representative feature vector under different probabilistic frameworks. However, none of those concept modelings accommodate the uncertainty that a concept is usually encountered in practice. Characterizing the concept with a probability distribution may meet our needs for uncertainty analysis of the concept as well as a decision theory guided classification. Beyond that, our preliminary experimental results over various datasets have shown that discriminant analysis of the concept features is also important for concept-aided classification.

All of the previously discussed methods focus on the standard definition of MIL where instances in a set are assumed to be independent. However, it could be that what generates the positive labelling of a bag is the inclusion of a particular pattern among its instances, e.g., a subset of its instances following a particular spatial or configuration arrangement. In such situations, the mere discovery of common instances among positive bags may not solve the problem since what is common are not individual instances but rather multi-instance patterns. This observation has prompted us to think about the following question: if we remove the independence assumption, can we generalize the goal of MIL from the one of finding a description of the common instance(s) to that of finding a description of the common pattern(s) among instances? Similarly, can we generate bag classifiers that discriminate based on common pattern(s) among instances instead of just common instance(s)?

The previous question can be answered by determining whether or not there is an informed and efficient way of extracting within-bag structural information. Knowledge about the local structure of a bag, e.g., probability density or clustering, may be helpful. Notice that in the special case that bags are sparse of instances, the nearest neighbor in each bag provides critical information for the commonality analysis over the highly overlapped regions of different positive bags. However, when there is an abundance of instances in each bag, the analysis of within-bag structural information may provide a more accurate description of any existing pattern. For example, suppose that every bag contains enough instances that reflect the distribution of the target concept. In such case, we can conduct a preprocessing step to analyze the modes of the within-bag distribution. The extracted modes along with the corresponding density can be further utilized in a joint probabilistic modeling of the instance and bag labels. In this case, the local modes can be understood as the representatives of each bag and their density values are the corresponding weights. In two aspects, the methodology is superior to the logistic regression model [21] described in the previous section, where only one statistics (e.g., the mean) of each bag is used as the representative for that bag. On the one hand, bags are simplified as a few informative representatives are extracted by way of local structures analysis. As every bag is fairly loaded with instances, presumably, one of those representative may better reflect the target concept than the nearest neighbor or the mean of the bag. One the other hand, the weights associated with the representatives can be incorporated into the bag-level commonality analysis. Thus, the local modes and their weights are the two factors

that determine how much a positive bag overlaps with the other positive bags. The extra cost on the local analysis will be later compensated in the commonality analysis over the simplified representations of bags.

Researchers have taken MIL into a multi-label setting [35, 36, 37, 38, 39], under which, a concept unifies several sub-concepts. For problems such as object recognition and text categorization, it is very likely that a target concept has separate representations in the instance space. For example, for the human body detection problem, a body concept may consists of visually different representations such as hair, face cloth, pants, etc. And each sub-concept might correspond to a distribution of features in the instance space. Unfortunately, learning multiple sub-concepts from MIL data has not been sufficiently addressed. Current approaches to multi-label MIL are all MIL classifiers and decompose the multi-label MIL problem into either the single-label MIL problem or the multi-label single instance learning problem. Overlooking the possibility of multiple representations of a single concept may severely degrade the performance for both MIL classifiers and concept learners. Although probabilistic modeling of multiple sub-concepts using noisy-or model is inherently capable of handling the problem, it requires the number of sub-concepts to be specified in advance. In comparison, the most-likely-cause model is less likely to produce a reliable estimate because only the most likely instance is taken into account.

4. Conclusions

We presented a survey of previous research on Multiple Instance Learning and explained how it has focused on the standard definition of the problem where instances in a set are independent. The following question, and other related questions, was raised and discussed: if we remove the independence assumption, can we generalize the concept of finding a description of the common instance(s) to that of finding a description of the common pattern(s) among instances? Similarly, can we generate bag classifiers that discriminate based on common pattern(s) among instances instead of just common instance(s)? We plan to explore this question as part of our future work and also hope that other researchers will investigate this exciting direction.

References

[1] Dietterich, T. G., Lathrop, R. H., Pérez, T. L.: Solving the multiple instance problem with axis-parallel rectangles. Artificial Intelligence, 89(1-2), pp. 31–71, 1997. ISSN 0004-3702.

[2] Teramoto, R., Kashima, H.: Prediction of protein-ligand binding affinities using multiple instance learning. Journal of Molecular Graphics and Modelling, 29(3), pp. 492–497, 2010.

[3] Maron, O., Perez, T. L.: A framework for multiple-instance learning. In: Proceedings of the Advances in Neural Information Processing Systems, pp. 570–576. 1998.

[4] Zhang, Q., Goldman, S. A.: EM-DD: An Improved Multiple-Instance Learning Technique. In: Advances in Neural Information Processing Systems, volume 14, pp. 1073–1080. MIT Press, 2001.

[5] Wang, C., Zhang, L., Zhang, H.-J.: Graph-based multiple-instance learning for object-based image retrieval. In: Proceedings of the 1st ACM international conference on Multimedia information retrieval, MIR '08, pp. 156–163. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-312-9.

[6] Li, F., Liu, R.: Multi-graph multi-instance learning for object-based image and video retrieval. In: Proceedings of the 2nd ACM International Conference on Multimedia Retrieval, ICMR '12, pp. 35:1–35:8. ACM, New York, NY, USA, 2012. ISBN 978-1-4503-1329-2.

[7] Rahmani, R., Goldman, S. A., Zhang, H., Krettek, J., Fritts, J. E.: Localized content based image retrieval. In: Proceedings of the 7th ACM SIGMM international workshop on Multimedia information retrieval, MIR '05, pp. 227–236. ACM, New York, NY, USA, 2005. ISBN 1-59593-244-5.

[8] Gondra, I., Xu, T.: A multiple instance learning based framework for semantic image segmentation. Multimedia Tools Appl., 48(2), pp. 339–365, 2010. ISSN 1380-7501.

[9] Qi, Z., Xu, Y., Wang, L., Song, Y.: Online multiple instance boosting for object detection. Neurocomput., 74(10), pp. 1769–1775, 2011. ISSN 0925-2312.

[10] Dollár, P., Babenko, B., Belongie, S., Perona, P., Tu, Z.: Multiple Component Learning for Object Detection. In: Proceedings of the 10th European Conference on Computer Vision: Part II, ECCV '08, pp. 211–224. Springer-Verlag, Berlin, Heidelberg, 2008. ISBN 978-3-540-88685-3.

[11] Viola, P., Platt, J. C., Zhang, C.: Multiple Instance Boosting for Object Detection. In: Advances in Neural Information Processing Systems 18, pp. 1417–1426. 2007.

[12] Babenko, B., Yang, M.-H., Belongie, S.: Robust Object Tracking with Online Multiple Instance Learning. IEEE Trans. Pattern Anal. Mach. Intell., 33(8), pp. 1619–1632, 2011. ISSN 0162-8828.

[13] Zhou, Z.-H., Jiang, K., Li, M.: Multi-Instance Learning Based Web Mining. Applied Intelligence, 22(2), pp. 135–147, 2005. ISSN 0924-669X.

[14] Andrews, S., Tsochantaridis, I., Hofmann, T.: Support vector machines for multiple-instance learning. In: Advances in Neural Information Processing Systems, volume 15, pp. 561–568. 2003.

[15] Zhou, X., Ruan, J., Zhang, W.: Promoter prediction based on a multiple instance learning scheme. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology, BCB '10, pp. 295–301. ACM, New York, NY, USA, 2010. ISBN 978-1-4503-0438-2.

[16] Mei, S., Fei, W.: Structural Domain Based Multiple Instance Learning for Predicting Gram-Positive Bacterial Protein Subcellular Localization. In: IJCBS'09, pp. 195–200. 2009.

[17] Long, P. M., Tan, L.: PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. In: Proceedings of the 9th Annual Conference on Computational learning Theory, pp. 228–234. 1996. ISBN 0-89791-811-8.

[18] Blum, A., Kalai, A.: A Note on Learning from Multiple-Instance Examples. Machine Learning, 30(1), pp. 23–29, 1998. ISSN 0885-6125.

[19] Auer, P.: On learning from multi-instance examples: empirical evaluation of a theoretical approach. In: Proceedings of the 4th International Conference on Machine Learning, pp. 21–29. 1997. ISBN 1-55860-486-3.

[20] Xu, T., Chiu, D., Gondra, I.: Constructing target concept in multiple instance learning using maximum partial entropy. In: Proceedings of the 8th international conference on Machine Learning and Data Mining in Pattern Recognition, MLDM'12, pp. 169–182. Springer-Verlag, Berlin, Heidelberg, 2012. ISBN 978-3-642-31536-7.

[21] Raykar, V. C., Krishnapuram, B., Bi, J., Dundar, M., Rao, R. B.: Bayesian multiple instance learning: automatic feature selection and inductive transfer. In: Proceedings of the 25th international conference on Machine learning, ICML '08, pp. 808–815. ACM, New York, NY, USA, 2008. ISBN 978-1-60558-205-4.

[22] Ray, S., Craven, M.: Supervised versus multiple instance learning: an empirical comparison. In: Proceedings of the 22nd international conference on Machine learning, ICML '05, pp. 697–704. ACM, New York, NY, USA, 2005. ISBN 1-59593-180-5.

[23] Xu, X., Frank, E.: Logistic Regression and Boosting for Labeled Bags of Instances. In: Proc. of the PacificAsia Conf. on Knowledge Discovery and Data Mining, pp. 272–281. Springer-Verlag, 2004.

[24] Wang, J., Zucker, J. D.: Solving the Multiple-Instance Problem: A Lazy Learning Approach. In:

Proceedings of the 17th International Conference on Machine Learning, pp. 1119–1126. 2000. ISBN 1-55860-707-2.

[25] Boser, B. E., Guyon, I. M., Vapnik, V. N.: A training algorithm for optimal margin classifiers. In: Proceedings of the fifth annual workshop on Computational learning theory, COLT '92, pp. 144–152. ACM, New York, NY, USA, 1992. ISBN 0-89791-497-X.

[26] Gärtner, T., Flach, P. A., Kowalczyk, A., Smola, A. J.: Multi-Instance Kernels. In: In Proc. 19th International Conf. on Machine Learning, pp. 179–186. Morgan Kaufmann, 2002.

[27] Chevaleyre, Y., Zucker, J. D.: Solving Multiple-Instance and Multiple-Part Learning Problems with Decision Trees and Rule Sets. Application to the Mutagenesis Problem. In: Proceedings of the 14th Biennial Conference of the Canadian Society on Computational Studies of Intelligence, pp. 204–214. 2001.

[28] Ruffo, G.: Learning Single and Multiple Instance Decision Trees for Computer Security Applications. Ph.D. thesis, Universita di Torino, Italy, 2001.

[29] Leistner, C., Saffari, A., Bischof, H.: MIForests: multiple-instance learning with randomized trees. In: Proceedings of the 11th European conference on Computer vision: Part VI, ECCV'10, pp. 29–42. Springer-Verlag, Berlin, Heidelberg, 2010. ISBN 3-642-15566-9, 978-3-642-15566-6.

[30] Zhang, M. L., Zhou, Z. H.: Adapting RBF Neural Networks to Multi-Instance Learning. Neural Process. Lett., 23(1), pp. 1–26, 2006.

[31] Schapire, R. E.: The Strength of Weak Learnability. Mach. Learn., 5(2), pp. 197–227, 1990. ISSN 0885-6125.

[32] Mason, L., Baxter, J., Bartlett, P., Frean, M.: Boosting Algorithms as Gradient Descent. In: In Advances in Neural Information Processing Systems 12, pp. 512–518. MIT Press, 2000.

[33] Friedman, J., Hastie, T., Tibshirani, R.: Additive Logistic Regression: a Statistical View of Boosting. The Annals of Statistics, 38(2), 2000.

[34] Hajimirsadeghi, H., Mori, G.: Multiple Instance Real Boosting with Aggregation Functions. In: International Conference on Pattern Recognition, ICPR. 2012.

[35] Zhang, M.-L., Zhou, Z.-H.: M3MIML: A Maximum Margin Method for Multi-instance Multi-label Learning. In: Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, ICDM '08, pp. 688–697. IEEE Computer Society, Washington, DC, USA, 2008. ISBN 978-0-7695-3502-9.

[36] Xu, X.-S., Xue, X., Zhou, Z.-H.: Ensemble multi-instance multi-label learning approach for video annotation task. In: Proceedings of the 19th ACM international conference on Multimedia, MM '11, pp. 1153–1156. ACM, New York, NY, USA, 2011. ISBN 978-1-4503-0616-4.

[37] Li, Y.-X., Ji, S., Kumar, S., Ye, J., Zhou, Z.-H.: Drosophila Gene Expression Pattern Annotation through Multi-Instance Multi-Label Learning. IEEE/ACM Trans. Comput. Biol. Bioinformatics, 9(1), pp. 98–112, 2012. ISSN 1545-5963.

[38] Yakhnenko, O., Honavar, V.: Multi-Instance Multi-Label Learning for Image Classification with Large Vocabularies. In: Procedings of the British Machine Vision Conference 2011, pp. 59.1–59.12. British Machine Vision Association, 2011. ISBN 1-901725-43-X.

[39] Zhou, Z.-H., Zhang, M.-L.: Multi-Instance Multi-Label Learning with Application to Scene Classification. In: NIPS'06, pp. 1609–1616. 2006.