Pattern Recognition 77 (2018) 329–353

Contents lists available at ScienceDirect

Pattern Recognition

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

Multiple instance learning: A survey of problem characteristics and applications

Marc-André Carbonneau[a,*], Veronika Cheplygina[b,c], Eric Granger[a], Ghyslain Gagnon[a]

[a] École de technologie supérieure, Université du Québec, Montréal, Canada
[b] Department of Biomedical Engineering, Eindhoven University of Technology, Eindhoven, The Netherlands
[c] Biomedical Imaging Group Rotterdam, Erasmus Medical Center, Rotterdam, The Netherlands

ARTICLE INFO

Article history:
Received 17 January 2017
Revised 4 August 2017
Accepted 7 October 2017
Available online 13 October 2017

Keywords:
Multiple instance learning
Weakly supervised learning
Classification
Multi-instance learning
Computer vision
Computer aided diagnosis
Document classification
Drug activity prediction

ABSTRACT

Multiple instance learning (MIL) is a form of weakly supervised learning where training instances are arranged in sets, called bags, and a label is provided for the entire bag. This formulation is gaining interest because it naturally fits various problems and allows to leverage weakly labeled data. Consequently, it has been used in diverse application fields such as computer vision and document classification. However, learning from bags raises important challenges that are unique to MIL. This paper provides a comprehensive survey of the characteristics which define and differentiate the types of MIL problems. Until now, these problem characteristics have not been formally identified and described. As a result, the variations in performance of MIL algorithms from one data set to another are difficult to explain. In this paper, MIL problem characteristics are grouped into four broad categories: the composition of the bags, the types of data distribution, the ambiguity of instance labels, and the task to be performed. Methods specialized to address each category are reviewed. Then, the extent to which these characteristics manifest themselves in key MIL application areas are described. Finally, experiments are conducted to compare the performance of 16 state-of-the-art MIL methods on selected problem characteristics. This paper provides insight on how the problem characteristics affect MIL algorithms, recommendations for future benchmarking and promising avenues for research. Code is available on-line at https://github.com/macarbonneau/MILSurvey.

1. Introduction

Multiple instance learning (MIL) deals with training data arranged in sets, called bags. Supervision is provided only for entire sets, and the individual labels of the instances contained in the bags are not provided. This problem formulation has attracted much attention from the research community, especially in the recent years, where the amount of data needed to address large problems has increased exponentially. Large quantities of data necessitate considerable labeling efforts.

Weakly supervised methods, such as MIL, can alleviate this burden since weak supervision is generally obtained more efficiently. For example, object detectors can be trained with images collected from the web using their associated tags as weak supervision instead of locally-annotated data sets [1,2]. Computer-aided diagnosis algorithms can be trained with medical images for which only patient diagnoses are available instead of costly local annotations provided by an expert. Moreover, there are several types of problems that can naturally be formulated as MIL problems. For example, in the drug activity prediction problem [3], the objective is to predict if a molecule induces a given effect. A molecule can take many conformations which can either produce, or not, a desired effect. Observing the effect of individual conformations is unfeasible. Therefore, molecules must be observed as a group of conformations, hence the use of the MIL formulation. Because of these attractive properties, MIL has been increasingly used in many other application fields over the last 20 years, such as image and video classification [4–9], document classification [10,11] and sound classification [12].

Several comparative studies and meta-analyses have been published to better understand MIL [13–23]. All these papers observe that the performance of MIL algorithms depends on the characteristics of the problem. While some of these characteristics have been partially analyzed in the literature [10,11,24,25], a formal definition of key MIL problem characteristics has yet to be described.

A limited understanding of such fundamental problem characteristics affects the advancement of MIL research in many ways.

* Corresponding author.
E-mail addresses: marcandre.carbonneau@gmail.com (M.-A. Carbonneau), v.cheplygina@tue.nl (V. Cheplygina), eric.granger@etsmtl.ca (E. Granger), ghyslain.gagnon@etsmtl.ca (G. Gagnon).

https://doi.org/10.1016/j.patcog.2017.10.009
0031-3203/

Experimental results can be difficult to interpret, proposed algorithms are evaluated on inappropriate benchmark data sets and results on synthetic data often do not generalize to real-world data. Moreover, characteristics associated with MIL problems have been addressed under different names. For example, the scenario where the number of positive instances in a bag is low was referred to as either sparse bags [26,27] or low witness rate [24,28]. It is thus important for future research to formally identify and analyze what defines and differentiates MIL problems.

This paper provides a comprehensive survey of the characteristics inherent to MIL problems, and investigates their impact on the performance of MIL algorithms. These problem characteristics are all related to unique features of MIL: the ambiguity of instance labels and the grouping of data in bags. We propose to organize problem characteristics in four broad categories: *Prediction level, Bag composition, Label ambiguity* and *Data distribution*.

Each characteristic raises different challenges. When instances are grouped in bags, predictions can be performed at two levels: bags-level or instance-level [19]. These two tasks have different misclassification costs therefore algorithms are often better suited for only one of them [20,21]. Bag composition, such as the proportion of instances from each class and the relation between instances, also affects the performance of MIL methods. The source of ambiguity on instance labels is another important factor to consider. This ambiguity can be related to label noise as well as to instances not belonging to clearly defined classes [17]. Finally, the shape of positive and negative distributions affect MIL algorithms depending on their assumptions about the data.

As additional contributions, this paper reviews state-of-the-art methods which can address challenges of each problem characteristic. It also examines several applications of MIL, and in each case, identifies their main characteristics and challenges. For example, in computer vision, instances can be spatially related, but this relationship does not exist in most bioinformatics applications. Finally, experiments show the effects of selected problem characteristics – the instance classification task, witness rate, negative class modeling and label noise – with 16 representative MIL algorithms. This is the first time that algorithms are compared on the bag and instance classification tasks in the light of these specific challenges. Our findings indicate that these problem characteristics have a considerable impact on the performance of all MIL methods, and that each method is affected differently. Therefore, problem characterization cannot be ignored when proposing new MIL methods and conducting comparative experiments. Finally, this paper provides novel insights and direction to orient future research in this field from the problem characteristics point-of-view.

The rest of this paper is organized as follows. The next section describes MIL assumptions and the different learning tasks that can be performed using the MIL framework. Section 3 reviews previous surveys and general MIL studies. Sections 4 and 5 identify and analyze the key problem characteristics and applications, respectively. Experiments are presented in Section 6, followed by a discussion in Section 7.

2. Multiple instance learning

2.1. Assumptions

In this paper, we consider two broad assumptions: the standard and the collective assumption. For a more detailed review on the subject, the reader is referred to [17].

The *standard MIL assumption* states that all negative bags contain only negative instances, and that positive bags contain at least one positive instance. These positive instances are named witnesses in many papers and this designation is used in this survey. Let X be a bag defined as a set of feature vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N\}$. Each instance (i.e. feature vector) $\mathbf{x}_i$ in feature space $\mathcal{X}$ can be mapped to a class by some process $f : \mathcal{X} \rightarrow \{0, 1\}$, where the negative and positive classes correspond 0 and 1 respectively. The bag classifier $g(X)$ is defined by:

$$g(X) = \begin{cases} 1, & \text{if } \exists \mathbf{x} \in X : f(\mathbf{x}) = 1; \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

This is the working assumption of many of the early methods [3,6,29], as well as recent ones [30,31]. To correctly classify bags under the standard assumption, it is not necessary to identify all witnesses as long as at least one is found in each positive bag. A more detailed discussion will be presented in Section 4.1.

The standard MIL assumption can be relaxed to address problems where positive bags cannot be identified by a single instance, but by the distribution, interaction or accumulation of the instances it contains. Here, instances in a bag are no longer independent and bag classifiers can take many forms. Next, we give three representative examples.

In some problems, several positive instances are necessary to assign a positive label to a bag. For example, in traffic jam detection from images of a road, a car would be a positive instance. However, an image containing a few cars is not positive because it takes many cars to create a traffic jam. In this case a bag classifier can be given by:

$$g(X) = \begin{cases} 1, & \text{if } \theta \leq \sum_{\mathbf{x} \in X} f(\mathbf{x}); \\ 0, & \text{otherwise}, \end{cases} \tag{2}$$

where θ is the minimal number of witnesses in positive bags.

A more general case for the collective assumption is when the class of a bag is defined by instances belonging to more than one concept. Foulds and Frank [17] give a simple and representative example of this assumption by classifying images of desert, sea and beach. Images of deserts contain sand segments, while images of the sea contain water segments. However, images of beaches must contain both types of segments. To correctly classify beach images, the model must verify the presence of both types of witnesses, and thus, methods working under the standard MIL assumption would fail in this case. Some methods assign instances to a set of defined concepts ($\mathcal{C}$), and some of these concepts belong to the positive class ($\mathcal{C}^+ \subset \mathcal{C}$). In that case, the bag classifier $g(X)$ is defined by:

$$g(X) = \begin{cases} 1, & \text{if } \forall c \in \mathcal{C}^+ : \theta_c \leq \sum_{\mathbf{x} \in X} f_c(\mathbf{x}); \\ 0, & \text{otherwise}, \end{cases} \tag{3}$$

where $f_c(\mathbf{x})$ is a process that outputs 1 if $\mathbf{x}$ belongs to concept c and θ_c is the number of instances belonging to c required to observe a positive bag. There are different levels of generality for multiple concept assumptions of this type [32]. Alternatively, bags can be seen as distributions of instances. In [33], the bag space $\mathcal{B}$ is defined as the set of all probability distributions on the instance space ($\mathcal{P}(\mathcal{X})$). Each bag X is a probability distribution over instances $P(\mathbf{x}|X)$. In that case, a bag classifier is a process that maps a probability distribution to a label: $g(X) : \mathcal{P}(\mathcal{X}) \rightarrow \{0, 1\}$.

In this survey, the *collective assumption* designates all assumptions in which more than one instance is needed to identify a positive bag.

2.2. Tasks

Classification: Classification can be performed at two levels: bag and instance. Bag classification is the most common task for MIL algorithms. It consists in assigning a class label to a set of instances. The individual instance labels are not necessarily important depending on the type of algorithm and assumption. Instance classification is different from bag classification because while training is performed using data arranged in sets, the objective is

to classify instances individually. As pointed out in [34], the loss functions for the two tasks are different (see Section 4.1). When the goal is bag classification, misclassifying an instance does not necessarily affect the loss at bag-level. For example, in a positive bag, few true negative instances can be erroneously classified as positive and the bag label will remain unchanged. Thus, the structure of the problem, such as the number of instances in bags, plays an important role in the loss function [20]. As a result, the performance of an algorithm for bag classification is not representative of the performance obtained for instance classification. Moreover, many methods proposed for bag classification (e.g. [35,36]) do not reason in instance space, and thus, often cannot perform instance classification.

MIL classification is not limited to assigning a single label to instances or bags. Assigning multiple labels to bags is particularly relevant considering that they can contain instances representing different concepts. This idea has been the object of several publications [37–39]. Multi-label classification is subject to the same problem characteristics as single label classification, thus no distinction will be made between the two in the rest of this paper.

Regression: MIL regression consists in assigning a real value to a bag (or an instance) instead of a class label. The problem has been approached in different ways. Some methods assign the bag label based on a single instance. This instance may be the closest to a target concept [40], or the best fit in a regression model [41]. Other methods work under the collective assumption and use the average or a weighted combination of the instances to represent bags as single feature vectors [42–44]. Alternatively, on can simply replace a bag-level classifier by a regressor [45].

Ranking: Some methods have been proposed to rank bags or instances instead of assigning a class label or a score. The problem differs from regression because the goal is not to obtain an exact real valued label, but to compare the magnitude of scores to perform sorting. Ranking can be performed at the bag-level [46] or at the instance-level [47].

Clustering: This task consists in finding clusters or a structure among a set of unlabeled bags. The literature on the subject is limited. In some cases, clustering is performed in bag space using standard algorithms and set-based distance measures (e.g. *k*-Medoids and the Hausdorff distance [48]). Alternatively, clustering can be performed at the instance-level. For example, in [49], the algorithm identifies the most relevant instance of each bag, and performs maximum margin clustering on these instances.

Most of the discussion in the remainder of the paper will be articulated around classification, as it is the most studied task. However, challenges and conclusions related to problem characteristics are also applicable to the other tasks.

3. Studies on MIL

Because many problems can be formulated as MIL, there is a plethora of MIL algorithms in the literature. However, there is only a handful of general MIL studies and surveys. This section summarizes and interprets the broad conclusions from these general MIL papers.

The first survey on MIL is a technical report written in 2004 [13]. It describes several MIL algorithms, some applications and discusses learnability under the MIL framework. In 2008, Babenko published a report [14] containing an updated survey of the main families of MIL methods, and distinguished two types of ambiguity in MIL problems. The first type is polymorphism ambiguity, in which each instance is a distinct entity or a distinct version of an entity (e.g. conformations of a molecule). The second is part-whole ambiguity in which all instances are parts of the same object (e.g. segments of an image). In a more recent survey [15], Amores proposed a taxonomy in which MIL methods are divided in three broad categories following the representation space. Methods operating in the instance-space are grouped together, and the methods operating in bag-space are divided in two categories based on whether a bag embedding is performed or not. Several experiments were performed to compare bag classification accuracy in four application fields. Bag-space methods performed better in terms of bag classification accuracy, however, performance depends on the data and the distance function or the embedding method. Recently, a book on MIL has been published [50]. It discusses most of the tasks of Section 2.2 along with associated methods, as well as data reduction and imbalanced data. Finally, Quellec et al. [51] wrote a survey on MIL for medical imaging applications, for which MIL is a particularly attractive solution. They review how problems are formulated in this field of application and analyze results from various experiments. They conclude that, while being more convenient, MIL outperforms single instance learning because it can pick up on subtle global visual cues that cannot be properly segmented and used as single instances to train a classifier.

Some papers study specific topics of MIL. For instance, Foulds and Frank [17] reviewed the assumptions made by MIL algorithms. They stated that these assumptions influence how algorithms perform on different types of data sets. They found that algorithms working under the collective assumption also perform well with data sets corresponding to the standard MIL assumption. Sabato and Tishby [52] analyzed the sample complexity in MIL, and found that the statistical performance of MIL is only mildly dependent on the number of instances per bag. In [23] the similarities between MIL benchmark data sets were studied. The data sets were represented in two ways: by meta-features describing numbers of bags, instances and so forth, and by features based on performances of MIL algorithms. Both representations were embedded in a 2-D space and found to be dissimilar to each other. In other words, data sets often considered similar due to the application or size of data did not behave similarly, which suggest that some unobserved properties influence MIL algorithms performance.

Some papers compare MIL to other learning settings to better understand when to use MIL. Ray and Craven [18] compared the performance of MIL methods against supervised methods on MIL problems. They found that in many cases, supervised methods yield the most competitive results. They also noted that, while some methods systematically dominate others, the performance of the algorithms was application-dependent. In [19], the relationship between MIL and other settings, such as group-based classification and set classification, is explored. They state that MIL is applicable in two scenarios: the classification of bags and the classification of instances. Recently, the differences between these two scenarios were rigorously investigated [20]. It was shown analytically and experimentally that the correlation between classification performance at bag and instance level is relatively weak. Experiments showed that depending on the data set, the best algorithm for bag classification provides average, or even the worst performance for instance classification. They too observed that different MIL algorithms perform differently given the nature of the data.

The classification of instances is a task in itself, but can also be an intermediate step toward bag classification for instance-space methods [15]. Alpaydin et al. [21] compared instance-space and bag-space classifiers on synthetic and real-world data. They concluded that for datasets with few bags, it is preferable to use an instance-space classifier. They also state, as in [15], that if the instances provide partial information about bag labels, it is preferable to use a bag-space representation. In [22], Cheplygina et al. explored the stability of the instance labels assigned by MIL algorithms. They found that algorithms yielding best bag classification performance were not the algorithms providing the most consistent instance labels. Carbonneau et al. [53] studied the ability to identify witnesses of several MIL methods. They found that de-

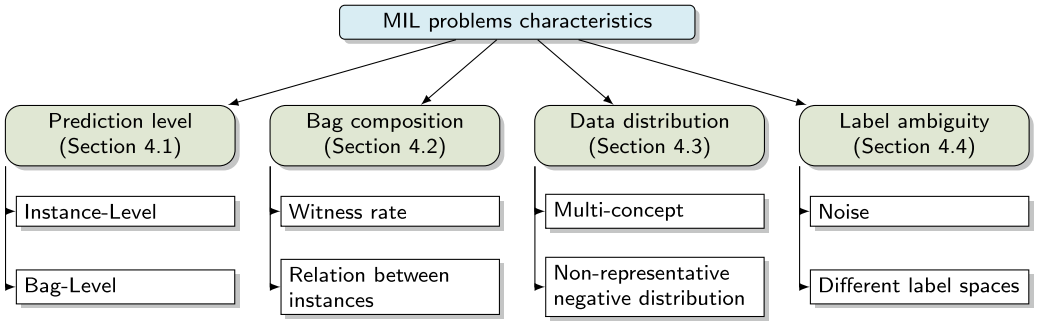

Fig. 1. Characteristics inherent to MIL problems.

pending on the nature of the data, some algorithms perform well while others would have difficulty learning.

Finally, some papers focus on specific classes of algorithms and applications. Doran and Ray [16] analyzed and compared several SVM-based MIL methods. They found that some methods perform better for instance classification than for bag classification, or vice-versa, depending on the method properties. Wei and Zhou [54] compared methods for generating bags of instances from images. They found that sampling instances densely leads to a higher accuracy than sampling instances at interest points or after segmentation. This agrees with other bag-of-words (BoW) empirical comparisons [55,56]. They also found that methods using the collective assumption performed better for image classification. Vankatesan et al. [57] showed that simple lazy-learning techniques can be applied to some MIL problems to obtain results comparable to state-of-the-art techniques. Kandemir and Hamprecht [58] compared several MIL algorithms in two computer-aided diagnosis (CAD) applications. They found that modeling intra-bag similarities was a good strategy for bag classification in this context.

The main conclusions of these studies are summarized as follows:

- The performance of MIL algorithms depends on several properties of the data set [15,18,20,21,23,53].
- When it is necessary to model combinations of instances to infer bag labels, bag-space and embedding methods perform better [15,21,51,54].
- The best bag-level classifier is rarely the best instance-level classifier, and vice versa [16,20].
- When the number of bags is low, it is preferable to use an instance-based method [21].
- Some MIL problems can also be solved using standard supervised methods [18].
- Performance of MIL is only mildly dependent on the number of instances per bag [52].
- Similarity between the instances of a same bag affect classification performance [58].

All of these conclusions are related to one or more characteristics that are unique to MIL problems. *Identifying these characteristics and gaining a better understanding of their impact on MIL algorithms is an important step towards the advancement of MIL research.* This survey paper mainly focuses on these characteristics and their implications for methods and applications. For a more general survey on MIL methods, we refer the interested reader to [15].

4. Characteristics of MIL problems

We identified four broad categories of key characteristics associated with MIL problems which directly impact on the behavior of MIL algorithms: *prediction level, bag composition, data distributions* and *label ambiguity* (as shown in Fig. 1). Each characteristic poses different challenges which must be addressed specifically.

In the remainder of this section, each of these characteristics will be discussed in more detail, along with representative specialized methods proposed in the literature to address them.

4.1. Prediction: instance-level vs. bag-level

In some applications, like object localization in images, the objective is not to classify bags, but to classify individual instances. In that case, problems are formulated with the implicit assumption that instances can be labeled as positive or negative. Following the notation of Section 2.1, for instance classification, the task is to learn $f(\mathbf{x})$ rather than $g(\mathbf{x})$. These two tasks are related in the sense that a perfect instance classifier $f^*(\mathbf{x})$ would result in a perfect bag classifier under the standard MIL assumption:

$$g^*(X) = \begin{cases} 1, & \text{if } \exists \mathbf{x} \in X : f^*(\mathbf{x}) = 1; \\ 0, & \text{otherwise,} \end{cases} \tag{4}$$

Inversely, a perfect bag classifier $g^*(X)$ achieves perfect instance classification since an instance can be viewed as a singleton bag, $S = \{\mathbf{x}\}$:

$$g^*(S) = f^*(\mathbf{x}). \tag{5}$$

In practice, none of these optimal classifiers are likely to be trained. More importantly, the relation between optimal classifiers for a given finite data set is no longer reciprocal. A perfect instance classifier still leads to an optimal bag classifier but the inverse is not true. For example, suppose all instances of a MIL data set are sampled from either one of two positive concepts (C_1 and C_2) or from a negative concept (C_-). Also, all positive bags contain positive instances from both positive concepts and from the negative concept: $X^+ = \{\mathbf{x}_1 \in C_1, \mathbf{x}_2 \in C_2, \mathbf{x}_3 \in C_-\}$. All negative bags contain instances sampled from the negative concept: $X^- = \{\mathbf{x}_1 \in C_-, \mathbf{x}_2 \in C_-, \ldots, \mathbf{x}_N \in C_-\}$. In this case, the following classifier achieves perfect bag classification:

$$\hat{g}^*(X) = \begin{cases} 1, & \text{if } \exists \mathbf{x} \in X : \hat{f}(\mathbf{x}) = 1; \\ 0, & \text{otherwise,} \end{cases} \tag{6}$$

where

$$\hat{f}(\mathbf{x}) = \begin{cases} 1, & \text{if } \mathbf{x} \in C_1; \\ 0, & \text{otherwise.} \end{cases} \tag{7}$$

While $\hat{g}^*(X)$ would correctly classify all bags in the data set, $\hat{f}(\mathbf{x})$ would misclassify half of the positive instances.

In MIL, training an instance classifier is non-trivial because instance labels are unavailable. This is why many methods use bag classification accuracy (e.g. APR [3], MI-SVM [6], MIL-Boost [59], EM-DD [35], MILD [60]) as a surrogate optimization objective to train an instance classifier in the hope that bag-level accuracy will

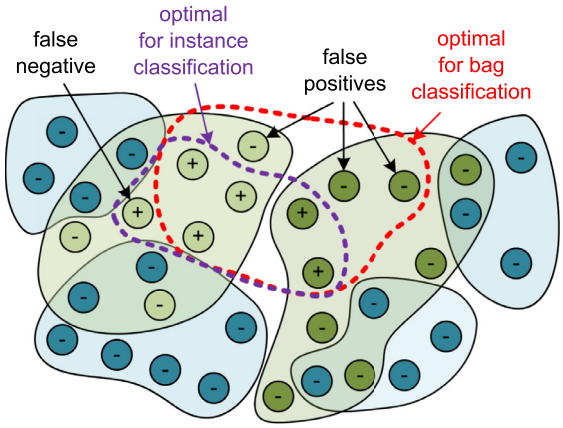

Fig. 2. Illustration of two decisions boundaries on a fictive problem. While only the purple boundary correctly classifies all instances, both of them achieve perfect bag classification. This is because, in that case, false positive and false negative instances do not impact on bag labels. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

be representative of instance-level accuracy. However, as will be discussed next, there are key differences in the cost function of the two tasks. These differences explain why the bag-level accuracy of a method does not reflect its accuracy at instance-level [16,20]. It was shown in analytic and empirical investigations [20] that the relationship between the accuracy at the two levels depends of the number of instances in bags, the class imbalance and the accuracy of the instance classifier. It follows that algorithms designed for bag classification are not optimal for instance classification.

Here we explain the difference between the instance misclassification cost for both classification levels. Under the standard MIL assumption, as soon as a witness is identified in a bag, it is labeled as positive and all other instance labels can be ignored. In that case, false positives (FP) and false negatives (FN) have no impact on the bag classification accuracy, but still count as classification errors at the instance level. In addition, when considering negative bags, a single FP causes a bag to be misclassified. This means that if 1% of the instances in each negative bag were misclassified, the accuracy on negative bags would be 0%, although the accuracy on negative instances would be 99%. This is illustrated in Fig. 2. The green ensembles represent positive bags, while negative bags correspond to blue ensembles. Each instance is identified with its true class. In this figure, both decision boundaries (dotted lines) are optimal for bag classification because they include at least one instance from all positive bags, while excluding all instances from negative bags. However, only one of the two boundaries achieves perfect instance classification (purple).

Most methods in the literature address the bag classification problem. These methods have been extensively surveyed in the past, and thus we refer the interested reader to [13–15]. A large proportion of methods proposed for instance classification measure bag classification accuracy to train an instance classifier. The predictions for all instances from a bag are aggregated, generally using the max function (or a differentiable approximation), and the loss is computed with respect to the bag label. This idea has been used to train a Boosting classifier in [61,62] and other types of model such as logistic regression [18] and deep neural networks [2]. The aforementioned methods were proposed for instance classification but are not different in spirit from most bag classification methods reasoning in the instance space like APR [3], EM-DD [35], MI-OptimalBall [63], MI-SVM [6] and SDB-MIL [31]. These methods classify instances individually before predicting bag labels which means they can directly be used for instance-level classification.

As explained above, using bag classification accuracy as a surrogate optimization objective is suboptimal. This is why it has been proposed to consider negative and positive bags separately in the classifier loss function [64]. The accuracy on positive bags is taken at bag level, but for negative bags, all instances are treated individually. This optimization criterion was proposed to adjust the decision threshold of bag classifiers for instance classification and improve their accuracy in [34]. In [65], a different weight is assigned to FP and FN during the optimization of an SVM. Virtually any bag-level classifier can classify instances if they are seen as singleton bags. This is the rationale behind Citation-kNN-ROI [66] which, however, does not perform well in practice (see Section 6.4). MILES [4] is a bag classification method based on prototype distance embedding and an SVM that can be used for instance classification. The method computes the contribution of each instance to the bag label assignation based on its distance to selected prototypes. Instances in positive bags for which the contribution is above a given threshold are identified as witnesses.

Some methods try to uncover the true label of the instances to train an instance classifier. One of the most well-known methods is mi-SVM [6]. After instance labels have been initialized, an SVM classifier is trained and used to update the label assignation. These two steps are performed iteratively until the label assignation remains unchanged. The resulting SVM classifier is used to predict the label of test instances. MissSVM [67] views the problem as semi-supervised learning where instances in positive bags are unlabeled. The algorithm is similar to mi-SVM except that it enforces the constraint that every positive bags contain a positive instance. KI-SVM [68] uses a multiple kernel approach in which a kernel encodes possible label assignations in the SVM constraints. In this method, it is assumed that there is the same number of positive instances in all positive bags. MILD [60] discovers a set of *true positive* instances. The probability that an instance is positive depends on the bag labels in its vicinity defined by a Gaussian kernel. The discovered true positive instances are used to train an SVM classifier. A similar idea is proposed in RSIS-EoSVM [30] where instances are projected in random subspaces and vicinity depends on cluster assignations. In that case, label assignation is probabilistic. Several training sets are sampled based on these probabilistic assignations to train an ensemble of SVM classifiers.

4.2. Bag composition

4.2.1. Witness rate

The witness rate (WR) is the proportion of positive instances in positive bags. When the WR is very high, positive bags contain only a few negative instances. In that case, the label of the instances can be assumed be the same as the label of their bag. The problem then reverts to a supervised problem with one-sided noise which can be solved in a regular supervised framework [69]. However, in some applications, WR can be arbitrarily small and hinder the performance of many algorithms. For example, in methods like Diverse Density (DD) [29], Citation-kNN [35] and APR [3] instances are considered to have the same label as their bag. When the WR is low, this is no longer reasonable which leads to lower performances. Methods which analyze instance distributions in bags [70–72] may also have problems dealing with low WR because distributions in positive and negative bags become similar. Also, some methods represent bags by the average of the instances they contain (e.g. NSK-SVM [73]), or consider their contribution to the bag label equally [74]. With very low WRs, the few positive instances have a limited effect after the pooling process. Finally, in instance classification problems, lower WRs mean serious class imbalance problems, which leads to bad performance for many methods.

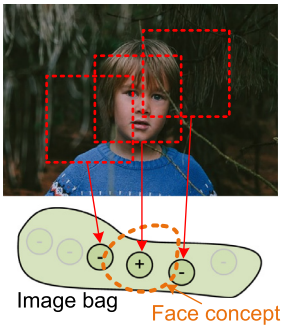

Fig. 3. Illustration of intra-bag similarity between instances: the patches are overlapping, and thus, share similarities with each other.

Several authors studied low WR problems in recent years. For example, sparse transductive MIL (stMIL) [27] is an SVM formulation similar to NSK-SVM [73]. However, to better deal with low WR bags, the optimization constraints of the SVM are modified to be satisfied when at least one witness is found in positive bags. This method performs well at low WR but is less efficient when it is higher. Sparse balanced MIL (sbMIL) [27] incorporates an estimation of the WR as a parameter in the optimization objective to solve this problem. WR estimation has also been successfully used in low WR problems by ALP-SVM [75], SVR-SVM [24] and the γ-rule [28]. One drawback of using the WR as a parameter is that the WR is assumed to be constant across all bags. Other methods, like CR-MILBoost [76] and RSIS [30], estimate the probability that each instance is positive before training an ensemble of classifiers. During training, the classifiers give more importance to the instances that are more likely to be witnesses. In miGraph [10], similar instances in a bag are grouped in cliques. The importance of each instance is inversely proportional to the size of its clique. Assuming positive and negative instances belong to different cliques, the WR has little impact. In miDoc [26], a graph represents the entire MIL problem, where bags are compared based on the connecting edges. Experiments show that the method performs well on very low WR problems.

4.2.2. Relations between instances

Most existing MIL methods assume, often not explicitly, that positive and negative instances are sampled independently from a positive and a negative distribution. However, this is rarely the case with real-world data. In many applications, the i.i.d. assumption is violated because a structure or correlations exist between instances and bags [10,77]. We make a distinction between three types of relation: intra-bag similarities, instance co-occurrences and structure.

Intra-bag similarities: In some problems, instances belonging to the same bag share similarities that instances from other bags do not. For instance, in the drug activity prediction problem [3], each bag contains many conformations of the same molecule. It is likely that instances of the same molecule are similar to some extent, while being different from other molecules [13]. One must ensure that the MIL algorithm learns to differentiate active from non-active conformations, instead of learning to classify molecules. In image-related applications, it is likely that all segments share some similarities related the capture conditions (e.g. illumination, noise, etc.). Alternatively, similarities between instances of a same bag may be related to the instance generation process. For example, some methods use densely extracted patches which overlap (Fig. 3). Since they share a certain number of pixels, they are likely to be correlated. Also, the background of a picture could be split in different segments which can be very similar (see Fig. 4).

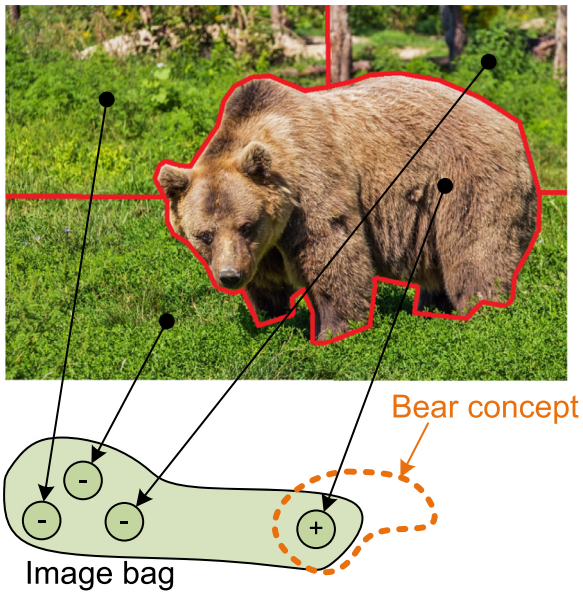

Fig. 4. Example of co-occurrence and similarity between instances: three segments contain grass and forest and are therefore very similar. Moreover, since this is an image of a bear, the background is more likely to be nature than a nuclear central control room.

Intra-bag similarities raise some challenges during learning. For instance, transductive algorithms (e.g. mi-SVM [6]) may not be able to infer instance labels if the nature of negative instances from positive and negative bags differ [18].

Very few methods were proposed explicitly to address this problem. To deal with similar instances, miGraph [10] builds a graph per bag and groups similar instances together to adjust their relative importance based on the group size. CCE [36] performs a clustering of the instance space. Bags are represented by a binary vector in which each bit corresponds to a cluster. A bit is set to 1 if at least one instance in the bag has been assigned to the corresponding cluster. A similar approach is used in [78] except bits are associated a pool of subgraphs patterns mined from the data set. Because features are binary, many instances can be assigned to the same cluster and the representation remains unaffected, which provides robustness to intra-bag similarity.

Instances are similar if they are close to each other in the metric space used by the classifier. Depending on the type of data, similarity or dissimilarity can be measured using different distance measures such as Euclidean [79], cosine [26] or χ^2 [80]. A good way to mitigate problems related to intra-class similarity is to define a new instance space in which distance are more related to class than bag membership. This new space can be obtained by selecting features which truly discriminate between classes (instead of bags) or by learning a representation in which class discriminant information is enhanced. In most cases, the new reduced instance space maximizes the distance between negative instances and the most positive instance of each positive bag. For example, Relief-MI [81] is an adaptation of the Relief [82] feature selection algorithm for MIL. For random bags, it identifies the nearest neighbors from each class under different versions of the Hausdorff distance. Then, it assigns a score to each feature based on the distance difference between the neighbor of the same class and the others under this feature. The most discriminant features are selected and the others are discarded. Other feature selection algorithms have been adapted for MIL in a similar fashion [83,84]. In B-M3IFW [79], a positive bag is represented by its most positive instance to form a pool of positive prototypes. Feature weights are obtained by maximizing a margin defined as the difference between two terms: the

distance between positive prototypes and negative instances and the distance between positive prototypes the mean of positive prototypes.

Several methods include built-in feature selection or weighting mechanisms. For instance, APR [3] searches for a subset of features in which a hyper-rectangle encompasses at least one instance from all positive bags while keeping negative instances outside. MIRVM [85] performs classification and feature selection at the same time in a Bayesian learning framework. It uses MILR [18] and perform optimal feature selection with the type-II maximum likelihood method. Diverse Density [29,35] scales the importance of each feature to define the optimal region encompassing the positive concept in the instance space. This scaling has also been used in [86] to increase the performance of BP-MIP [87].

Finally, feature learning methods project instances in a space of reduced dimensionality where class discrimination at bag level is enforced. Usually this means maximizing the distance between negative instances and the most positive instance of each positive bag in the projection space. This can be achieved using an MIL adaptation of discriminant analysis or other linear projection methods [88–91] where bag classification accuracy is maximized.

Instance co-occurrence: Instances co-occur in bags when they share a semantic relation. This type of correlation happens when the subject of a picture is more likely to be seen in some environment than in another, or when some objects are often found together (e.g. knife and fork). For example, the bear of Fig. 4 is more likely to be found in nature than in a nightclub. Thus, observing nature segments might help to decide if the image contains a cocktail or a bear [92]. In [93], it is shown that different birds are often heard in the same audio fragment, so a "negative" bird song could help to correctly classify the bird of interest. In these examples, co-occurrence represents an opportunity for better accuracy, however, in some cases it is a necessary condition for successful classification. Consider the example given by Foulds and Frank [17] where one must classify sea, desert and beach images. Both desert and beach images can contain sand instances, while water instances can be found in sea and beach images. However, both instances must co-occur in a beach image. Most methods working under the collective assumption [17] naturally leverage co-occurrence. Many of these methods, like BoW [70,94], miFV [72], FAMER [95] or PPMM [96] represent bags as instance distributions which indirectly account for co-occurrence. This has also been directly modeled in a tensor model [97] and in a multi-label framework [37].

While useful to classify bags, in instance classification problems, the co-occurrence of instances may confuse the learner. If a given positive instance often co-occurs with a given negative instance, the algorithm is more likely to consider the negative instance as positive, which in this context would lead to a higher false positive rate (FPR).

Instance and bag structure: In some problems, an underlying structure exists between instances of a same bag or even between bags [77]. Structure is more complex than simple co-occurrence in the sense that instances follow a certain order, or are related in a meaningful way. Capturing this structure may lead to better classification performance [10,80,98]. The structure may be spatial, temporal, relational or even causal. For example, when a bag represents a video sequence, all frames or patches are temporally and spatially ordered. For example, it is difficult to differentiate between a person taking or leaving a package without taking this temporal order into account. Alternatively, in web mining tasks [77] where websites are bags and pages linked by the websites are instances, there exists a semantic relation between two bags representing websites linked together.

Graph models were proposed to better capture the relations between the different entities in non-i.i.d. MIL problems. Structure can be exploited at many levels: graphs can be used to model the relations between bags, instances or both [26,77]. Graphs enforce that related objects belong to the same class. Alternatively, [99] represents bags as graphs capturing diverse relationships between objects. The objects are shared across all bags and all possible sub-graphs of the bag graph correspond to instances. In [78,100], complex objects such as web pages and scientific papers are represented as collections of graphs. Discriminative subgraph patterns are mined to create a dictionary. Graph collections are represented by binary feature vectors in which each bit corresponds a subgraph in the dictionary. A bit is set to 1 if the corresponding subgraph is part of the collection. In [101], spatial structure in the image is captured by a similarity matrix and a neighborhood consistency constraint is enforced in a 1-norm SVM formulation.

Temporal and spatial structure between instances can be modeled in different ways. In BoW models for computer vision, this can be achieved by dividing the images [102,103] or videos [80] into different spatial and/or temporal zones. Each zone is characterized individually, and the final representation is the concatenation of every zone feature vectors. For audio and video, sub-sequences of instances have been analyzed using traditional sequence modeling tools such as conditional random fields (CRF) [104] and hidden Markov model (HMM) [105]. Spatial dependency in images have also been modeled in with CRF in [37,106].

4.3. Data distributions

Many methods make implicit assumptions on the shape of the distributions, or on how well the negative distribution is represented by the training set. In this section, the challenges associated with the nature of the overall data distribution is studied.

4.3.1. Multimodal distributions of positive instances

Some MIL algorithms work under the assumption that the positive instances are located in a single cluster or region in feature space. This is the case for several early methods like APR [3], which searches for a hyper-rectangle that maximizes the inclusion of instances from positive bags while excluding instances from negative bags. Diverse Density (DD) [29] methods follow a similar idea. These methods locate the point in feature space closest to instances in positive bags, but far from instances in negative bags. This point is considered to be the positive concept. Some more recent methods also follow the single cluster assumption. CKMIL [107] locates the most positive instance in each bag based on its proximity to a single positive cluster center. In [31], the classifier is a sphere encompassing at least one positive instance from each positive bag while excluding instances from negative bags. The method in [104] employs a similar strategy.

The single cluster assumption is reasonable in some applications such as molecule classification, but problematic in many other contexts. In image classification, the target concept may correspond to many clusters. For example, Fig. 5, shows several pictures of ants. Ants can be black, red or yellow, they can have wings and different body shapes depending on species and castes. Their appearance also changes depending on the point-of-view. It is unlikely that a compact location in feature space encompasses all of these variations.

Many MIL methods can learn multimodal positive concepts, however, only few representative approaches will be mentioned due to space constraints. First, non-parametric methods based on distance between bags like Citation-kNN [108] and MInD [93] naturally deal with all shapes of distributions. Simple non-parametric methods often lead to competitive results in MIL problems [57]. Methods using distances to a set of prototypes as bag representation, like DD-SVM [109] and MILES [4], can model many positive

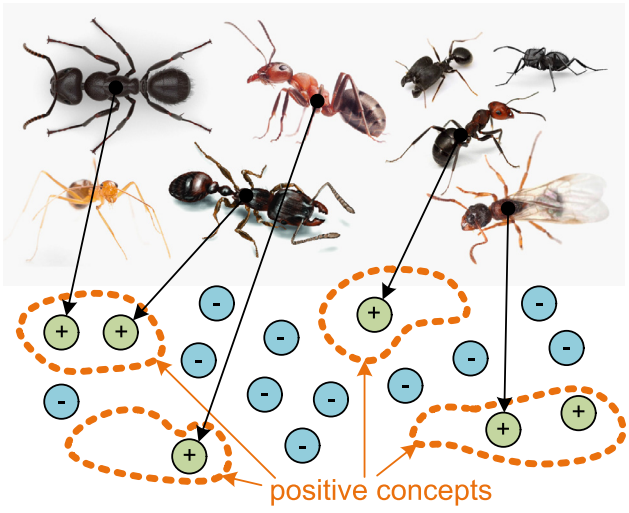

Fig. 5. For the same concept *ants*, there can be many data clusters (modes) in feature space corresponding to different poses, colors and castes. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

clusters, because each different cluster can be represented by a different prototype. Instance-space SVM-based methods like mi-SVM [6] can deal with disjoint regions of positive instances using a kernel. Also, methods modeling instance distributions in bags such as vocabulary-based [70] methods naturally deal with data sets containing multiple concepts/modes. The mixture-model in [110] naturally represents different positive clusters.

4.3.2. Non-representative negative distribution

In [33], it is stated that learnability of instance concepts requires that the distribution in test is identical to the training distribution. This is true for positive concepts, however, in some applications, the training data cannot entirely represent the negative instance distribution. For instance, provided sufficient training data, it is reasonable to expect that an algorithm learns a meaningful representation that captures the visual concept of a human person. However, since humans can be found in many different environments, ranging from jungle to spaceship, it is almost impossible to entirely model the negative class distribution. In contrast, in some applications like tumor identification in radiography, healthy tissue regions compose the negative class. These tissues possess a limited appearance range that can be modeled using a finite number of samples.

Several methods model only the positive class, and thus are well-equipped to deal with different negative distributions in test. In most cases, these methods search for a region encompassing the positive concept. In APR [3] this region is a hyper-rectangle, while in many others it is one, or a collection of, hyper-spheres/-ellipses [29,31,35,111]. These methods perform classification based on the distance to a point (concept) or a region in feature space. Everything that is far enough from the point, or outside the positive region, is considered negative. Therefore, the shape of the negative distribution is unimportant. A similar argument can be made for some non-parametric methods such as Citation-kNN [108]. These methods use the distance to positive instances, instead of positive concepts, and thus, offer the same advantage. Alternatively, the MIL problem can be seen as a one-class problem, where positive instances are the target class. Consequently, several methods using one-class SVM have been proposed [112–114].

Experiments in Section 6.6 compare reference MIL algorithms in contexts where the negative distribution is different in training and in test.

4.4. Label ambiguity

Label ambiguity is inherent to weak supervision. In MIL, this ambiguity can take different forms depending on the assumption under which the problem is formulated. Under the standard MIL assumption, there is no ambiguity on instance labels in negative bags. In that case, MIL can be viewed as a special kind of semi-supervised problem [67] where the labeled portion of the data belongs to only one class and the instance are structured in sets with label constraints. Under more relaxed MIL assumptions, there are supplementary sources of ambiguity such as noise on labels and different label spaces for instances and bags.

4.4.1. Label noise

Some MIL algorithms, especially those working under the standard MIL assumption, rely heavily on the correctness of bag labels. For instance, it was shown in [57] that DD is not tolerant to noise in the sense that a single negative instance in the neighborhood of the positive concept can hinder performance. A similar argument was made for APR [60] for which a negative bag mislabeled as positive, would lead to a high FPR.

In practice, there are many situations where positive instances may be found in negative bags. There are situations where labeling errors occur, but sometimes labeling noise is inherent to the data. For example, in computer vision applications, it is difficult to guarantee that negative images contain no positive patches: An image showing a house may contain flowers, but is unlikely to be annotated as a flower image [115]. Similar problems may arise in text classification, where a paragraph contains an analogy and uses words from another subject.

Methods working under the collective assumption can naturally deal with label noise. Positive instances found in negative bags have less impact because these methods do not assign label solely based on the presence of a single positive instance. Methods representing bags as distributions [70,71,116] can naturally deal with noisy instances because a single positive instance does not significantly change the distribution of a negative bag. Methods summarizing bags by averaging the instances like NSK-kernel [73] also provide robustness to noise in a similar manner. Another strategy to deal with noise is to count the number of positive instances in bags, and establish a threshold for positive classification. This is referred as the threshold-based MI Assumption in [17]. The method proposed [115] uses both the thresholding and the averaging strategies. The instances of a bag are ranked from most positive to less positive, and the bags are represented by the mean of the top-ranking instances and the mean of the bottom ranking instances. The averaging operation mitigates the effects of positive instance in negative bags. In [117], robustness to label noise is obtained by using dominant sets to perform clustering and select relevant instance prototypes in a bag-embedding algorithm similar to MILES [4].

Experiments in Section 6.7 compare the robustness to label noise of the reference methods.

4.4.2. Different label spaces

There are MIL problems in which the label space for instances is different from the label space for bags. In some cases, these spaces will correspond to different granularity levels. For example, a bag labeled as a car will contain instances labeled as wheel, windshield, headlights, etc. In other cases, instance labels might not have clear semantic meanings. Fig. 6 shows an example where the positive concept is zebra (represented by the region encompassed by the orange dotted line). This region contains several types of patches that can be extracted from a zebra picture. However, it is possible to extract patches from negative images that fall into this positive region. In this example, some patches extracted from the

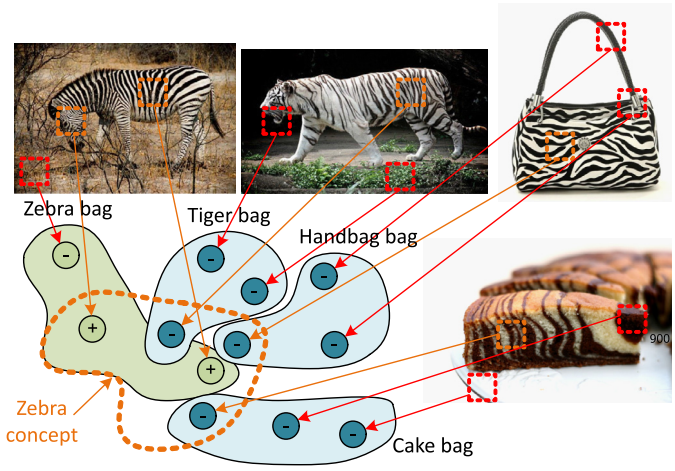

Fig. 6. This is an example of instances with ambiguous labels. *Zebra* is the target concept and instances relating to this concept should fall in the region delimited by the dotted line. However, negative images can also contain instances falling inside the zebra concept region. (For interpretation of the references to color in this figure, the reader is referred to the web version of this article.)

image of a white tiger, a purse and a marble cake fall into the zebra concept region. In that case the patches do not have semantic meaning easily understandable by humans.

When instances cannot be assigned to a specific class, methods operating under the standard MIL assumption, which must identify positive instances, are inadequate. Therefore, in those cases, using the collective assumption is necessary. Vocabulary-based methods [70] are particularly well adapted for this situation. They associate instances to words (e.g. prototypes or clusters) discovered from the instance distribution. Bags are represented by distributions over these words. Similarly, methods using embedding based on distance from selected prototype instances, such as MILES [4] and MILIS [118], can also deal with this type of problem.

All the characteristics presented in this section define a variety of MIL problems, which each must be addressed differently. The next section relates these characteristics to the prominent application fields of MIL.

5. Applications

MIL represents a powerful approach that is used in different application fields mostly (1) to solve problems where instances are naturally arranged in sets and (2) to leverage weakly annotated data.

This section surveys the main application fields of MIL. Each field is examined with respect to its different problem characteristics of Section 4 (summarized in Table 1).

5.1. Biology and chemistry

Problems in biology and chemistry can often be naturally formulated as MIL because of the inability to observe individual instance classes. For example, in the molecule classification task presented in the seminal paper by Dietterich et al. [3], the objective is to predict if a molecule will be binding to a musk receptor. Each molecule can take many conformations, with different binding strengths. It is not possible to observe the binding strength of a single conformation, but it is possible to observe it for groups of conformations, hence the MIL problem formulation.

Since then, MIL has found use in many drug design and biological applications. Usually, the approach is similar to Dietterich's: complex chemical or biological entities (compounds, molecules, genes, etc.) are modeled as bags. These entities are composed of

Table 1
Typical problem characteristics associated with MIL in literature for different application fields (Legend: ✓ likely to have a moderate impact, ✓✓ likely to have a large impact on performance).

Application Fields	Problem characteristics									
	Instance classification	Real-valued outputs	Low witness rate	Intra-bag similarities	Instance co-occurrence	Structure in bags	Multimodal positive distribution	Non-representative negative distribution	Label noise	Different label spaces
Drug activity prediction	✓	✓		✓✓			✓	✓		
DNA Protein identification	✓✓	✓		✓✓		✓✓	✓	✓		
Binding sites identification	✓✓	✓	✓	✓✓			✓	✓		
Image Retrieval			✓	✓	✓✓	✓✓	✓✓	✓✓	✓	✓✓
Object localization in image	✓✓		✓	✓	✓	✓	✓✓	✓✓	✓✓	✓
Object localization in video	✓✓		✓	✓	✓	✓✓	✓✓	✓✓	✓✓	✓
Computer aided diagnosis	✓	✓	✓	✓	✓	✓✓	✓	✓	✓✓	✓
Text classification	✓		✓		✓✓		✓✓	✓	✓	✓
Web mining	✓		✓	✓	✓	✓	✓	✓		✓
Sound classification	✓			✓	✓	✓✓	✓	✓	✓	
Activity recognition	✓				✓	✓✓	✓	✓	✓	✓

parts or regions that can induce an effect of interest. The goal is to classify unknown bags and sometimes to identify witnesses to better understand underlying mechanisms of the biological or chemical phenomenon. MIL has been used, among others, to predict a drug's bioavailability [46], predict the binding affinity of peptides to major histocompatibility complex molecules [45], discover binding sites governing gene expression [119,120] and predict gene functions [121].

The problems presented in this section are of various natures, but there are some characteristics that apply to a majority of them. For example, in most cases, the bags represent many arrangements or viewpoints of the same entity (e.g. drug, genes, etc.), which translate into high intra-bag similarities. Also, many applications call for quantification, using ranking or regression [40] (e.g. quantifying the binding strength of a molecule), which is more difficult and less documented than classification. Some characteristics apply only to a type of application. Objects like DNA sequences produce structured bags, while the many conformations of the same molecule do not. Finally, some problems require the identification of entities responsible for an effect (e.g. drug binding). This calls for methods with instance classification capabilities.

5.2. Computer vision

MIL is used in computer vision for two main reasons: to characterize complex visual concepts using sets of different sub-concepts, and to learn from weakly annotated data. The next subsections describe how MIL is used for content-based image retrieval (CBIR) and object localization. MIL is gaining momentum in the medical imaging community, and a subsection will also be devoted to this application field.

5.2.1. Content based image retrieval

CBIR is probably the single most popular application of MIL. The list of publications addressing this problem is long [4–7,112,122–125]. The task in CBIR is to categorize images based on the objects/concepts they contain. The exact localization of objects is not important, which means it is primarily a bag classification problem. Typically, images are partitioned into smaller parts or segments, which are then described by feature vectors. Each segment corresponds to an instance, while the whole image corresponds to a bag. Images can be partitioned in many ways, which are compared in [54]. For example, the image can be partitioned using a regular grid [123], key-points [94] or semantic regions [65,109]. In the latter case, the images are divided using state-of-the-art segmentation algorithms. This limits instance ambiguity since segments tend to contain only one object.

Visual data poses several challenges to MIL algorithms mainly because images are a good example of non-i.i.d. data. For one, some objects are more likely to co-occur in the same picture than others (e.g. bird and sky). Methods leveraging these co-occurrences tend to be more successful. Also, a bag can contain many similar instances, especially if the instances are obtained using dense grid sampling. Methods using segmentation algorithms are less subject to this problem since segments tend to correspond to single objects. Sometimes, the image is composed of several concepts, which means methods working under the collective assumption perform better. Moreover, working with images often means working with large intra-class variability. The same object can, for instance, appear considerably different depending on the points of view. Many types of object also come is a variety of shapes and colors. This means it is unlikely that a unimodal distribution adequately represents the entire class. Finally, backgrounds can vary considerably, making it difficult to learn a negative distribution that models every possible background object.

5.2.2. Object localization and segmentation

In MIL, the localization of objects in images (or videos) means learning from bags to classify instances. Typically, MIL is used to train visual object recognition systems on weakly labeled image data sets. In other words, labels are assigned to entire images based on the objects they contain. These objects do not have to be in the foreground, and an image may contain multiple objects. In contrast, in strongly supervised applications, bounding boxes indicating the location of each object are provided along with object labels. In other cases, pixel-wise annotations are provided instead. These bounding boxes, or pixel annotations, are often manually specified, and thus, necessitate considerable human effort. The computer vision community turned to MIL to leverage the large quantity of weakly annotated images found on the Internet to build object detectors. The weak supervision can come from description sentences [126–128], web search engine results [129], tags associated with similar images and words found on web pages associated with the images [2].

In several methods for object localization, bags are composed of many candidate bounding boxes corresponding to instances [1,59,130–132]. The best bounding box to encompass the target object is assumed to be the most positive instance in the bag. Efforts were dedicated to localize objects and segment them at pixel-level using traditional segmentation algorithms such as Constraint Parametric Min-Cuts [133], JSEG [37] or Multi-scale combinatorial grouping [134]. Alternatively, segmentation can be achieved by casting each pixel of the image as an instance [135].

Instance classification has also been applied in videos. It has been used to recognize complex events such as "attempting a board trick" or "birthday party" [8,136]. Several concepts compose these complex events. Evidences of these concepts sometimes last only for a short time, and can be difficult to observe in the total amount of information presented in the video. To deal with this problem, video sequences are divided in shorter sequences (instances) that are later classified individually. This problem formulation is also used in [137] to recognize scenes that are inappropriate for children. Also in videos, MIL methods were proposed to perform object tracking [61,138,139]. For example, in [61] a classifier is trained online to recognize and track an object of interest in a frame sequence. The tracker proposes candidate windows which compose a bag and are used to train the MIL classifier.

It can be difficult to manually select a finite set of classes to represent every object found in a set of images. Thus, it was proposed to perform the object localization alongside class discovery [129]. The method is akin to multiple instance clustering methods [48,49], but generates bags using a saliency detector, which remove background objects from positive bags to achieve higher cluster purity. A method based on multiple instance clustering was also proposed to discover a set of actons (sub-actions) from videos to create a mid-level representation of actions [140].

Object localization is susceptible to the same challenges as CBIR: instances in images are correlated, exhibit high similarity and spatial (and temporal for videos) structures exist in the bags. The objects can be deformable, have various appearances and be observed from different viewpoints. Therefore, a single concept is often represented by a multimodal distribution, and the negative distribution cannot be entirely captured by a training set. However, object localization is different from CBIR because it is an instance classification problem, which means that many bag-level algorithms are inapplicable. Several authors have also noted that in this context, MIL algorithms are sensitive to initialization [9,130].

5.2.3. Computer aided diagnosis and detection

MIL is gaining popularity in medical applications. Weak labels, such as the overall diagnosis of a subject, are typically easier to obtain than strong labels, such as outlines of abnormalities in a med-

ical scan. The MIL framework is appropriate in this situation given that patients have both abnormal and healthy regions in their medical scan, while healthy subjects have only healthy regions. The diseases and image modalities used are very diverse; applications include classification of cancer in histopathology images [141], diabetes in retinal images [142], dementia in brain MR [143], tuberculosis in X-ray images [144], classification of a chronic lung disease in CT [145] and others.

Like in other general computer vision tasks, there are two main goals in these applications: diagnosis (i.e. predicting labels for subjects), and detection or segmentation (i.e. predicting labels for a part of a scan). These parts can be pixels or voxels (3D pixel), image patches or regions of interest. Different applications pursue one or both goals, and have different reasons for doing so.

When the focus is on classifying bags, MIL classifiers benefit from using information about co-occurrence and structure of instances. For example, in [144], a MIL classifier trained only with X-ray images labeled as healthy or as containing tuberculosis, outperforms its supervised version, trained on outlines of tuberculosis lesions. Similar results are observed on the task of classification of chronic obstructive pulmonary disease (COPD) from chest computed tomography images [145].

Literature that is focused on classifying instances is somewhat less common, which may be a consequence of the lack of instance-labeled datasets. However, the lack of instance labels is often the motivation for using MIL in the first place, which means instance-level evaluation is necessary if these classifiers are to be translated into clinical practice. Some papers do not perform instance-level evaluation because the classifier does not provide such output [143], but state that this would be a useful extension of the method in the future. Others provide instance labels but do not have access to ground truth, thus resorting to more qualitative evaluation. For example, [145] examines whether the instances classified as "most positive" by the classifier have similar intensity distributions to what is already known in the literature. Finally, when instance-level labels are available, the classifier can be evaluated quantitatively and/or qualitatively. Quantitative evaluation is performed in [58,142,144]. In addition, the output of the classifier can be displayed in the image, which is an interpretable way of visualizing the results. In [144], the mi-SVM classifier provides local real-valued tuberculosis abnormality scores for each pixel in the image, which are then visualized as a heatmap on top of the X-ray image.

CAD shares many key challenges with other less constrained computer vision tasks. Depending on the sampling – which can be done on a dense grid [58,144], randomly [145] or according to constraints [143] – the instances can display varying degrees of similarity. In many pathologies, abnormalities are likely to include different subtypes, which have different appearance resulting in multimodal concept distributions. Moreover, differences between patients, such as age, sex and weight, as well as differences in acquisition of the images can also lead to large intra-class variability. On the other hand, the negative distribution (healthy tissue) is more constrained than in computer vision applications. This means that it is reasonable to attempt to capture and model the negative distribution, which is very difficult in unconstrained image recognition problems. Another particularity of CAD problems is that they are naturally suitable to have real-valued outputs, because diseases can have different stages, although this is often not considered when off-the-shelf algorithms are applied. For example, the chronic lung disease COPD has 4 different stages, but [145] treats them all as the positive class. During evaluation, the mild stage is most often misclassified as healthy. Tong et al. [143] considers binary classification tasks out of four possible classes (healthy, two types of mild cognitive impairment, and Alzheimer's), while these could be considered as a continuous scale. Lastly, CAD can be formulated as an instance and a bag classification task.

5.3. Document classification and web mining

Considering the Bag-of-Words (BoW) model is a MIL model working under the collective assumption, document classification is one of the earliest (1954) applications of MIL [146]. BoW represents texts as frequency histograms quantifying the occurrence of each word in the text. In this context, texts and web pages are multi-part entities that require the MIL classification framework.

Texts often contain several topics and are easily modeled as bags. Text classification problems can be formulated as MIL at different levels. At the lowest level, instances are words like in the BoW model. Alternatively, instances can be sentences [44,147], passages [6,148] or paragraphs [18]. In [6], bags are text documents, which are divided in overlapping passages corresponding to instances. The passages are represented by a binary vector in which each element is a medical term. The task is to categorize the texts. In [149], instances are short posts from different newsgroups. A bag is a collection of posts and the task is to determine if a group of posts contains a reference to a subject of interest. In [18], the task consists of identifying texts that contain a passage which links a protein to a particular component, process or function. In this case, paragraphs are instances while entire texts are bags. The paragraphs are represented by a BoW alongside distances from protein names and key terms. In [150], the content of emails is analyzed to detect spam. A common approach to elude spam filters is to include words that are not associated with spam in the message. Representing emails as bags of passages proved to be an efficient way to deal with these attacks. In [44,147,151,152], MIL was used to infer the sentiment expressed in individual sentences based on labels provided for entire user reviews. MIL has also been used to discover relations between named entities [11]. In this case, bags are collections of sentences containing two words that may or may not express a target relation (e.g. "Rick Astley" lives in "Montreal"). If the two words are related in the specified way, some of the sentences in the bag will express this relation. If that is not the case, none of the sentences will indicate the relation, hence the MIL formulation.

Web pages can also be naturally modeled using the MIL framework. Just like texts, web pages often contain many topics. For instance, a news channel website contains several articles on a diversity of subjects. MIL has been used for web index-page recommendations based on a user browsing history [153,154]. A web index page contains links, titles and sometimes short description of web pages. In this context, a web index page is a bag, and the linked web pages are the instances. Following the standard MIL assumption, it is hypothesized that if a web index page is marked as favorite, the user is interested in a least one of the pages linked to it. Web pages are represented by the set of the most frequent terms they contain. In contextual web advertisement, advertisers prefer to avoid certain pages containing sensitive content like war or pornography. In [147], a MIL classifier assesses sections of web pages to identify suitable web pages for advertisement.

Text data poses particular challenges for MIL. Most of the time, instances are non-i.i.d. Words may have different meanings depending on the context and thus, co-occurrence is important in this type of application. While BoW methods are successful to some degree, structure is an important component of sentences which convey important semantic information. Often, only small passages or specific words indicate the class of the document, which means WR can be quite low. Depending on the task and the formulation of the problem, bag and instance classification can be performed. In addition, text classification can present an additional difficulty compared to other applications. When texts are

represented by word frequency features (e.g. BoW) the data is very sparse and high-dimensional [6]. This type of data is often difficult to handle by classifiers using Euclidean-like distance measures. These distributions are highly multimodal and it is difficult to adequately represent the distribution of negative data.

5.4. Other applications

The MIL formulation has found its way to various other application fields. In this section, we present some less common applications for MIL along with their respective formulation.

Reinforcement learning (RL) shares some similarities with MIL. In both cases, only a weak supervision is provided for the instances. In RL, a reward, the weak supervision, is assigned to a state/action pair. The reward obtained for the state/action pair is not necessarily directly related to it, but might be related to preceding actions and states. Consider a RL agent learning how to play chess. The agent obtains a reward (or punishment) only at the end of the game. In other words, a label is given for a collection (bag) of action/state pairs (instances). This correspondence has motivated the use of MIL to accelerate RL by the discovery of subgoals in a task [99]. The action/state pairs leading to the achievement of these sub-goals are, in fact, the positive instances in the successful episodes. The main challenge for RL tasks is to consider the structure in bags and the label noise since good actions can be found in bad sequences.

Just like for images, some sound classification tasks can be cast as MIL. In [155], the objective is to automatically determine the genre of musical excerpts. In training, labels are provided for entire albums or artists, but not for each excerpt. The bags are collections of excerpts from the same artist or album. It is possible to find different genres of music on the same album or from the same artist, therefore the bags may contain positive and negative instances. In [12], MIL is used to identify bird songs in recordings made by unattended microphones in the wild. Sound sequences contain several types of birds and other noises. The objective is to identify each birdsong individually while training only on weakly labeled sound files.

Some methods represent audio signals as spectrograms and use image recognition techniques to perform recognition [156]. This idea has been used for bird song recognition [157] with histograms of gradients. In [158], personality traits are inferred from speech signals represented as spectrograms in a BoW framework. In that case, entire speech signals are bags and small parts of the spectrogram are instances. The BoW framework has been used in a similar fashion in [159], however, in that case instances are cepstrum feature vectors representing 1 s-long audio segments. Audio classification poses different challenges depending on how sounds are represented. For example, when a sound signal is represented as a time series, capturing structure is important. However, in a BoW framework, the co-occurrence of different markers will be more important. In many cases, the background noise related to capture conditions leads to high intra-bag similarity.

Time series are found in several applications other than audio classification. For instance, in [105,160] MIL is used to recognize human activities from wearable body sensors. The weak supervision comes from the users stating which activities were performed in a given time period. Typically, activities do not span across entire periods and each period may contain different activities. In this setup, instances are sub-periods, while the entire periods are bags. A similar model is used for the prediction of hard drive failure [161]. In this case, time series are a set of measurements on hard drives taken at regular intervals. The goal is to predict when a product is about to fail. Time series imply structure in bags that should not be ignored.

In [162,163], MIL classifiers detect buried landmines from ground-penetrating radar signals. When a detection occurs at a given GPS coordinate, measures are taken at various depths in the soil. Each detection location is a bag containing feature vectors for different depths.

In [29], MIL is used to select stocks. Positive bags are created by pooling the 100 best-performing stocks each month, while negative bags contain the 5 worst performing stocks. An instance classifier selects the best stocks based on these bags.

In [99], a method learning relational structure in data predicts which movies will be nominated for an award. A movie is represented by a graph that models its relations to actors, studios, genre, release date, etc. The MIL algorithm identifies which subgraph explains the nomination to infer the success of test cases. This type of structural relation between bags and instance is akin to web page classification problems.

6. Experiments

In this section, 16 reference methods are compared using data sets that allow to shed light on some of the problem characteristics discussed in Section 4. These experiments are conducted to show how problem characteristics influence the behavior of MIL algorithms, and demonstrate that these characteristics cannot be neglected when designing or comparing MIL algorithms. Four characteristics were selected, each from a different category, to represent the spectrum of characteristics. Algorithms are compared on the instance classification task, under different WR, with an unobservable negative distribution and with different degrees of label noise. These characteristics were chosen because their effect can be isolated and easily parametrized. The reference methods used in the experiments were chosen because they represent most families of approaches and include most widely used reference methods. All experiments have been conducted using Matlab and some implementations from the MIL toolbox [164] and the LAMDA website.[1]

Next we describe the reference methods used in the experiments. The methods are grouped based on the representation space following a taxonomy similar to [15]. Instance-space methods classify each instance individually and combine the instance labels to assign a bag to a class. Bag-space methods do not classify, explicitly at least, instances individually. Bag-space methods employ one of two strategies: either compare distance between bags using an appropriate distance measure for sets or distributions, or encode the content of the bags to obtain a summarizing representation used in a supervised learning setting.

6.1. Instance-space methods

SI-SVM, SI-SVM-TH and SI-kNN: These are not MIL methods *per se*, but this type of approaches has been used as a reference point in many papers [18,21,27] to give an indication on the pertinence of using MIL methods instead of regular supervised algorithms. In these algorithms, each instance is assigned the label of its bag, and bag information is discarded. In test, the classifier assigns a label to each instance, and a bag is positive if it contains at least one positive instance. For SI-SVM-TH the number of positive instances detected is compared to a threshold that is optimized on the training data.

MI-SVM and mi-SVM [6]: These algorithms are transductive SVMs. Instances inherit their bag label. The SVM is trained and classifies each instance in the data set. It is then retrained using the new label assignments. This procedure is repeated until the labels remain stable. The resulting classifier is used to classify test

[1] http://lamda.nju.edu.cn/.

instances. MI-SVM uses only the most positive instance of each bag for training, while mi-SVM uses all instances.

EM-DD [35]: DD [29] measures the probability that a point in feature space belongs to the positive class given the class proportion of instances in the neighborhood. EM-DD uses the Expectation-Maximization algorithm to locate the maximum of the DD function. Classification is based on the distance from this maximum point.

RSIS [30]: This method probabilistically identifies the witnesses in positive bags using a procedure based on random subspacing and clustering introduced in [53]. Training subsets are sampled using the probabilistic labels of the instances to train an ensemble of SVMs.

MIL-Boost [59]: The MIL-Boost algorithm used in this paper is a generalization of the algorithm presented in [62]. The method is essentially the same as gradient boosting [165] except that the loss function is based on bag classification error. The instances are classified individually, and their labels are combined to obtain bag labels.

6.2. Bag-space methods

C-kNN [108]: This is an adaptation of kNN to MIL problems. The distance between two bags is measured using the minimal Hausdorff distance. C-kNN relies on a two-level voting scheme inspired from the notion of citations and references in research papers. The algorithm was adapted in [66] to perform instance classification.

MInD [93]: With this method, each bag is encoded by a vector whose fields are dissimilarities to the other bags in the training data set. A regular supervised classifier, an SVM in this case, classifies these feature vectors. Many dissimilarity measures are proposed in the paper, but the *meanmin* offered the best overall performance and will be used in this paper.

CCE [36]: This algorithm is based on clustering and classifier ensembles. At first, the feature space is clustered using a fixed number of clusters. The bags are represented as binary vectors in which each bit corresponds to a cluster. A bit is set to 1 when at least one instance in a bag is assigned to its cluster. The binary codes are used to train one of the classifiers in the ensemble. Diversity is created in the ensemble by using a different number of clusters each time.

MILES [4]: In Multiple Instance Learning via Embedded instance Selection (MILES) an SVM classifies bags represented by feature vectors containing maximal similarities to selected prototypes. The prototypes are instances from the training data selected by a 1-norm SVM. Instance classification relies on a score representing the instance contribution to the bag label.

NSK-SVM [73]: The normalized set kernel (NSK) basically averages the distances between all instances contained in two bags. The kernel is used in an SVM framework to perform bag classification.

miGraph [10]: This method represents each bag by a graph in which instances correspond to nodes. Cliques are identified in the graph to adjust the instance weights. Instances belonging to large cliques have lower weights so each concept present in the bag is equally represented when instances are averaged. A graph kernel captures similarity between bags and is used in an SVM.

BoW-SVM: Creating a dictionary of representative words is the first step when using a BoW method. This is achieved with BoW-SVM by performing k-means clustering on all the training instances [15]. Next, instances are represented by the most similar word contained in the dictionary. Bags are represented by frequency histograms of the words. Histograms are classified by an SVM using a kernel suitable for histogram comparison (exponential χ^2 in this case).

EMD-SVM: The Earth Mover distance (EMD) [116] is a measure of the dissimilarity between two distributions. Each bag is a distribution of instances and the EMD is used to create a kernel used in an SVM.

6.3. Data sets

Spatially Independent, Variable Area, and Lighting (SIVAL) [166]: This data set contains 500 images each segmented and manually labeled by [149]. It contains 25 classes of complex objects photographed from different viewpoints in various environments. Each bag is an image partitioned in approximately 30 segments. A 30-dimensional feature vector encodes the color, texture and neighbor information of each segment. There are 60 images in each class, which are in turn considered as the positive class. 5 randomly selected images from each of the 24 other classes yield 120 negative bags. The WR is 25.5% in average but ranges from 3.1% to 90.6%. In this data set, unlike in other image data sets, co-occurrence information between the objects of interest and the background is nonexistent because all 25 objects are photographed in the same environments.

Birds [12]: The bags of this data set correspond to 10 s recordings of bird songs from one or more species. The recording is segmented temporally to create instances, which belong to a particular bird or to background noise. These 10,232 instances are represented by 38-dimensional feature vectors. Readers should refer to the original paper for details on the features. There are 13 types of bird in the data set, each in turn considered as the positive class. Therefore 13 problems can be generated from this data set. In this data set, low WR poses a challenge, especially since it is not constant across bags. Moreover, bag classes are sometimes severely imbalanced.

Newsgroups [149]: The newsgroups data set was derived from the *20 Newsgroups* [167] data set corpus. It contains posts from newsgroups on 20 subjects. Each post is represented by 200-term frequency-inverse document frequency (TFIDF) features. This representation generally yields sparse vectors, in which each element is representative of a word frequency in the text scaled by its frequency in the entire corpus. When one of the subjects is selected as the positive class, all 19 other subjects are used as the negative class. The bags are collections of posts from different subjects. The positive bags contain an average of 3.7% of positive instances. This problem is semi-synthetic and does not correspond to a real-world application. There is thus no exploitable co-occurrence information, intra-bag similarities or bag structure. However, the representation yields sparse data, which is different from the two previous data sets, and is representative of text applications.

HEPMASS [168]: The instances of this data set come from the HEPMASS Data Set.[2] It contains more than 10M instances which are simulation of particle collisions. The positive class corresponds to collisions that produce exotic particles, while the negative class is background noise. Each instance is represented by a 27-dimensional feature vector containing low-level kinematic measurements and their combination to create higher level mass features (see original paper for more details). For each WR value, 10 versions of the MIL data are randomly generated. For each version, the training and a test sets contain 50 positive bags and 50 negative bags composed of 100 instances.

Letters [169]: This semi-synthetic MIL data set uses instances from the Letter Recognition data set.[3] It contains a total of 20k instances representing each of the 26 letters in the English alphabet. Each of these letters can be seen as a concept and used to cre-

[2] http://archive.ics.uci.edu/ml/datasets/HEPMASS.

[3] https://archive.ics.uci.edu/ml/datasets/Letter+Recognition.

ate different positive and negative distributions. Each letter is encoded by a 16-dimensional feature vector that has been standardized. The reader is referred to the original paper for more details. In WR experiments, for each WR value, 10 versions of the MIL data sets are randomly generated. Each version has a training and a test set. Both sets contain 50 positive bags and 50 negative bags each containing 20 instances. In the positive bags, witness are sampled from 3 letters randomly selected to represent positive concepts. All other letters are considered as negative concepts. For the experiments on negative class modeling, the data set is divided in train and test partitions each containing 200 bags. Each bag contains 20 instances. The bag classes are equally proportioned and the WR is 20%. Like before, the positive instances are samples from 3 randomly selected letters. Half of the remaining letters constitute the initial negative distribution and the other half constitutes the unknown negative distribution.

Gaussian toy data: In this synthetic data set, the positive instances are drawn from a 20-dimensional multivariate Gaussian distribution ($\mathcal{G}(\mu, \mathbf{\Sigma})$) that represents the positive concept. The values of $\boldsymbol{\mu}$ are drawn from $\mathcal{U}(-3, 3)$. The covariance matrix ($\mathbf{\Sigma}$) is a randomly generated semi-definite positive matrix in which the diagonal values are scaled to [0,0.1]. The negative instances are sampled from a randomly generated mixture of 10 similar Gaussian distributions. This distribution is gradually replaced by another randomly generated mixture. The data set is standardized after generation. The test and training partitions both contain 100 bags. There are 20 instances in each bag and the WR is 20%.

6.4. Instance-level classification

In this section, the reference methods with instance classification capabilities are compared on three benchmark data sets: SIVAL, Birds and Newsgroups. These data sets are selected because they represent three different application fields and because instance labels are provided, which is somewhat uncommon with MIL benchmark data sets. There already exist several comparative studies for bag-level classification, we refer interested reader to [15,58].

The experiments were conducted using a nested cross-fold validation protocol [170]. It consists of two cross-validation loops. An outer loop assesses the performance of the algorithm in test, and an inner loop is used to optimize the algorithm hyper-parameters. This means that for each test fold of the outer loop, hyper-parameters optimization is performed via grid-search. Average performance is reported on results for the outer loop test folds.

Instance classification problems often exhibit class imbalance, especially when the WR is small. In these cases, comparing algorithm in terms of accuracy can be misleading. In this section, algorithms are compared in terms of unweighted average recall (UAR) and F_1-score. The UAR is the average of the accuracy for each class. The F_1-score is the harmonic mean between precision and recall. The 3 data sets translate into 58 different problems. For easy comparison, Figs. 7 and 8 present the results in the form of critical difference diagrams [171] with a significance level of 1%.

Results indicate that a successful strategy for instance classification is to discard bag information. With both metrics, the best algorithms are mi-SVM and SI-SVM, which assign the bag label to each instance and then treat them as atomic elements. This is consistent to the results obtained in [58]. These two methods are closely related because SI-SVM corresponds to the first iteration of mi-SVM. SI-kNN also yields competitive results and uses the same strategy. Even if the Birds and the Newsgroups data sets both possess low WR, it would seem that supervised methods are better suited for this task than MIL methods which use bag accuracy as an optimization objective (MILES, EMDD and MIL Boost). MI-SVM and RSIS rely on the identification of the most positive instances

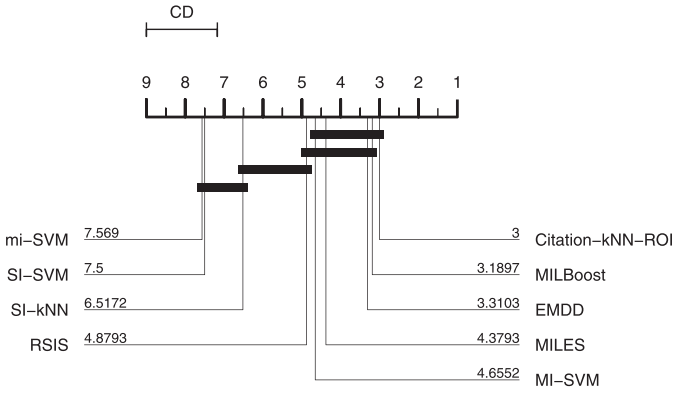

Fig. 7. Critical difference diagram for UAR on instance classification ($\alpha = 0.01$). Higher numbers are better.

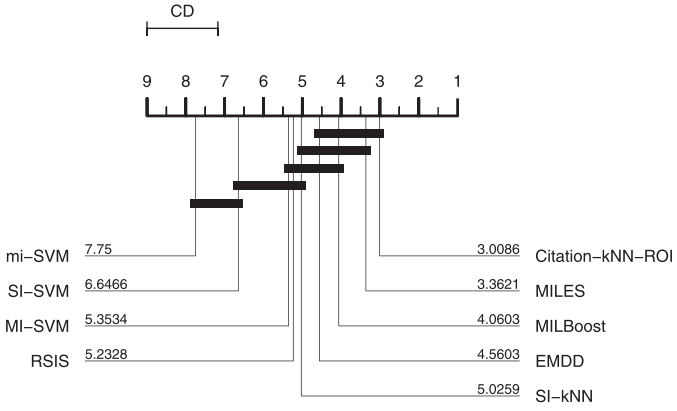

Fig. 8. Critical difference diagram for the F_1-score on instance classification ($\alpha = 0.01$). Higher numbers are better.

in each bag. This strategy seems successful to some degree, but is prone to ignore more ambiguous positive instances that are dominated by the others in the same bag. These conclusions have also been observed in the results obtained on the individual data sets.

6.5. Bag composition: witness rate

These experiments study the effects of the WR on performance. Two semi-synthetic data sets were created to allow control over the WR, and observe the behavior of the reference methods in greater detail: Letters and HEPMASS. These data sets are created from supervised problems that were artificially arranged in bags. This has the advantage of eliminating any structure and co-occurrence in the data, and thus better isolate the effect of WR. The original data sets must possess a high number of instances to emulate low WR. In the Letters data set, the positive class contains three concepts while in HEPMASS there is only one concept, which has an impact for some algorithms.

All hyper-parameters were optimized for each version of the data sets, and for each WR value using grid search and cross-validation. The results reported in Figs. 9–12 are the average results obtained on the test data for each of the 10 generated versions. Performance are compared using AUC and the UAR.

There are several things that can be concluded by examining the experiment results. Firstly, *for all methods, lower WR translates into lower accuracy*. However, Fig. 9 shows that *for the instance classification task, higher WR does not necessarily means higher accuracy* for all methods. In fact, for the Letters data set, three different letters are used to create positive instances which makes the positive distribution multimodal. As discussed in Section 6.4, some meth-

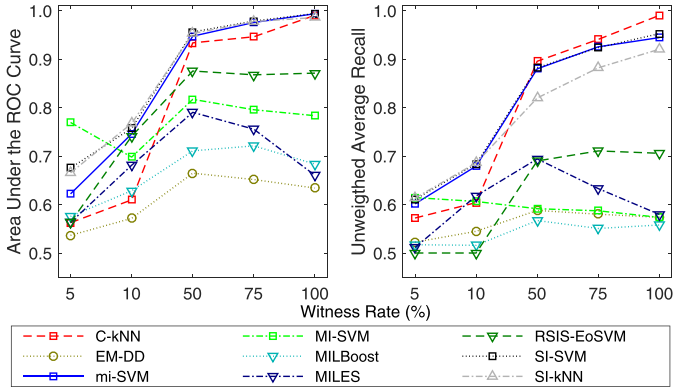

Fig. 9. Average performance of the MIL algorithms for instance classification on the Letters data set as the witness rate increases.

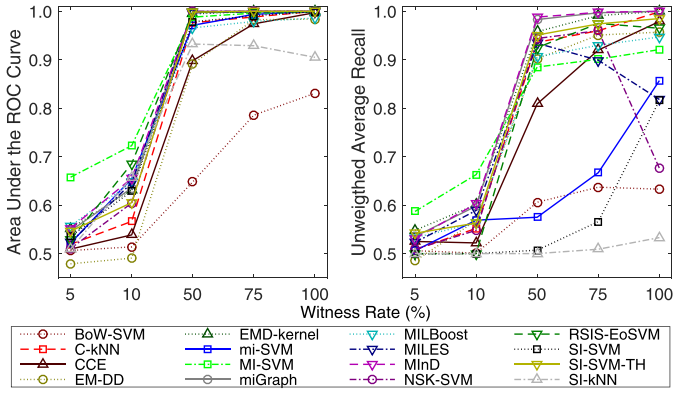

Fig. 10. Average performance of the MIL algorithms for bag classification on the Letters data set as the witness rate increases.

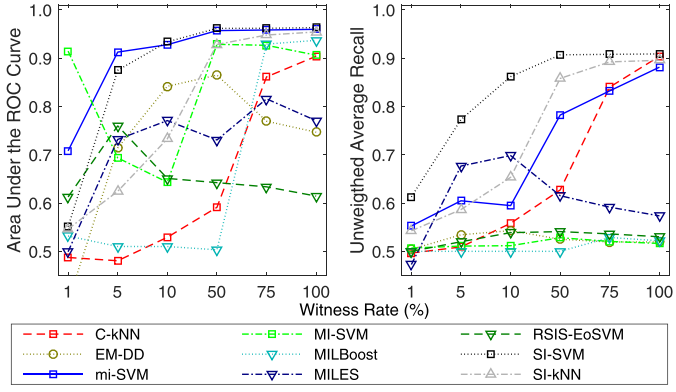

Fig. 11. Average performance of the MIL algorithms for instance classification on the HEPMASS data set as the witness rate increases.

ods are optimized for bag classification (EM-DD, MI-SVM, MILES, MILBoost, RSIS-EoSVM). In those cases, once a letter is assigned to the positive class in a positive bag, the bag is correctly classified. The remaining positive letters can be ignored and the algorithm still achieves perfect bag classification. This can be observed by comparing Figs. 9 and 11 with Figs. 10 and 12, where the methods optimized for bag classification deliver lower accuracy for instance classification, but their accuracy is comparable to other instance-based methods when classifying bags. This explains in part the observation [16,20] that an algorithm performance for one task is not always representative of the performance in the other.

The results in Figs. 9 and 11 suggest that *supervised classifiers are as effective for instance classification as the best MIL classifiers*

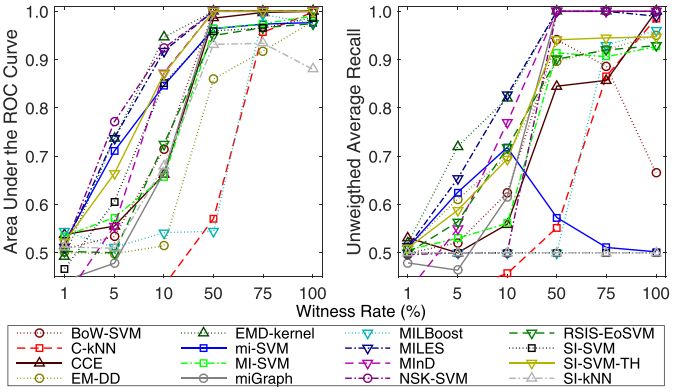

Fig. 12. Average performance of the MIL algorithms for bag classification on the HEPMASS data set as the witness rate increases.

Table 2
Ranking of instance-based methods vs. bag-based methods for the bag classification task.

Metric	Method type	WR	
		$<50\%$	$\geq 50\%$
Mean rank (AUC)	Instance-based	9.3	11.3
	Bag-based	7.7	5.7
Mean rank (UAR)	Instance-based	10.0	11.0
	Bag-based	7.0	6.0

when the WR is over 50%. In this case, the mislabeled negative instances are just noise in the training set, which is easily dealt with by the SVM or the voting scheme of the SI-kNN. Even when WR is lower than 50% supervised methods perform better than some of their MIL counterparts. MI-SVM has higher AUC performance when the WR is at its lowest compared to the other method. This is explained by the fact that positive bags are represented by their single most positive instance. When the WR is at its minimum, there is only one witness per bag which coincides with this representation.

The results for bag classification are reported in Figs. 10 and 12. For an easier comparison between instance- and bag-based methods, mean ranks for all experiments are reported in Table 2. These results show that, *in general, bag-space methods outperform their instance-space counterparts at higher WR* ($\geq 50\%$). At lower WR (5–10%), the difference between both approaches is lower. However, in the Letters experiment, MI-SVM outperform all other methods by a significant margin, while in the HEPMASS experiment, EMD-SVM and NSK-SVM perform better. This suggests that *at lower WRs, there are other factors to consider when selecting a method*, such as the shape of the positive and negative distributions and the consistency of the WR across positive bags.

6.6. Data distribution: non-representative negative distribution

In some applications, the negative instance distribution cannot be entirely represented by the training data set. The experiments in this section measure the ability of MIL algorithms to deal with a negative distribution different in test and training. We use two data sets in these experiments: the Letters data set and the synthetic Gaussian toy data set created specially for this experiment. Using these two data sets makes it possible to control factors to measure the effect of a changing negative distribution in isolation from other problem characteristics. In each experiment, there are two different negative instance distributions. The first one is used to generate the training data. For the test data sets, at first, the negative instances are also sampled from this same distribution, but are gradually replaced by instances from the second distribu-

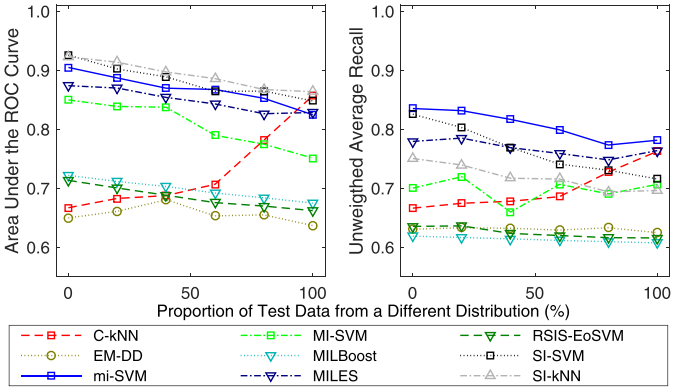

Fig. 13. Average performance of the MIL algorithms for instance classification on the Letters data as the test negative instance distribution increasingly differs from the training distribution.

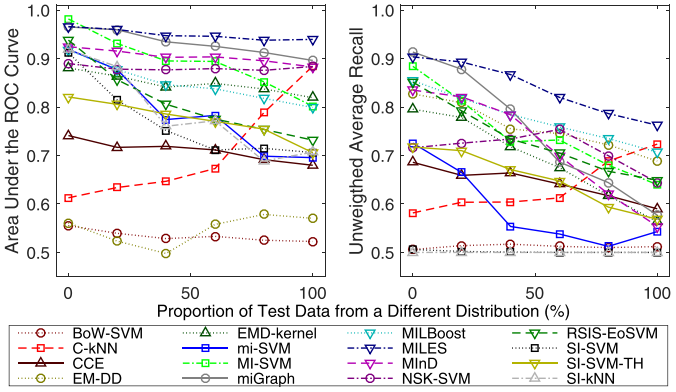

Fig. 14. Average performance of the MIL algorithms for bag classification on the Letters data as the test negative instance distribution increasingly differs from the training distribution.

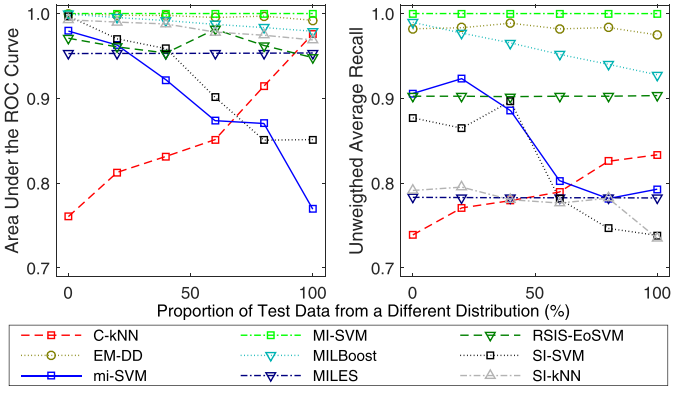

Fig. 15. Average performance of the MIL algorithms for instance classification on Gaussian toy data as the test negative instance distribution increasingly differs from the training distribution.

tion. The positive instances are sampled from the same distribution in both the training and test sets. For instance, using the Letters data set, this means that in the training data set the letters A, B and C are used as negative instances. Gradually, the instance from A, B and C are replaced by instances from D, E and F.

The results of the experiments, illustrated in Figs. 13–16, show that *most algorithms have decreasing performance when the test negative distribution differs from the training distribution*. However, C-kNN exhibits a contrasting behavior. The more the test instances differ from training instances, the better are performances. This is because C-kNN uses the minimal Hausdorff distance as a similarity

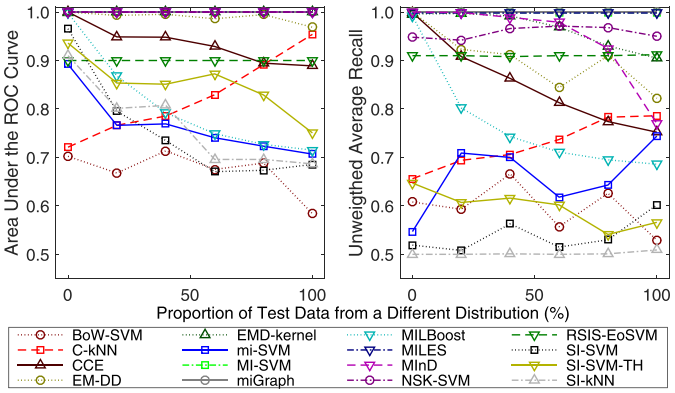

Fig. 16. Average performance of the MIL algorithms for bag classification on Gaussian toy data as the test negative instance distribution increasingly differs from the training distribution.

metric between bags. This is the distance between the two closest instances from each bag. If the negative instances come from the same distribution in all the bags, it is likely that the closest instances are both from the negative distribution, even if the bags are positive. If the bags have different labels, this leads to misclassification. If the negative test instances are different from those in the training set, the distance between two negative instances is likely to be greater than the distance between two positive instances, which are from the same distribution in both sets. Thus, positive bags are found to be closer to other positive bags leading to a higher accuracy.

The results for both data sets suggest that *bag-space methods are better for dealing with new negative distributions*. This may contribute to their success in computer vision applications where negative distributions are difficult to model. In Fig. 14 the AUC for bag classification is stable for most method while their accuracy decreases. This suggest that the score functions learned by the algorithms are still suitable for the new distribution, but the thresholds should be adjusted. This observation motivates the use of adaptive methods in practice to adjust the decision threshold as new data arrives.

6.7. Label ambiguity: label noise

It is generally assumed that the weak supervision provided by bag labels is accurate. However, as explained in Section 4.4, this is not always the case. Here, we measure the ability of reference algorithms to deal with noisy labels. Experiments are conducted on the Letters and SIVAL datasets. In these experiments, an increasing proportion of bag labels in the training set are inverted. When 50% of the labels are inverted, both classes contain an equal proportion of true positive and negative bags. After, 50% of the labels are inverted, the problem can be seen as the same classification problem where the negative class is considered as the positive class.

For bag classification, the experiments reveal that label noise robustness relates to the decision space used by MIL classifiers. *Bag-space methods using an embedding strategy (e.g. EMD-kernel, miGraph, MInD) are the most robust to label noise.* The results for these methods are reported in Figs. 19 and 20. The symmetry in their performance curves suggests that these *embedding methods make no distinction between the positive and the negative class*, and thus their label can be interchanged seamlessly. Embedding algorithms encode bags in a single feature vector and view the bag classification problem as a supervised problem. In that regard, the robustness of the method depends on the type of classifier used by a given method.

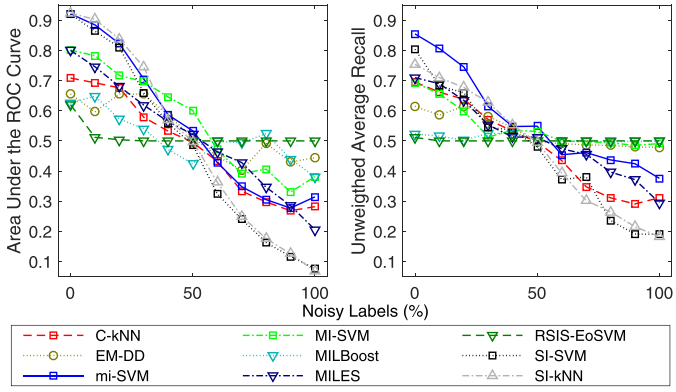

Fig. 17. Average performance of the MIL algorithms for instance classification on the Letters data with increasing label noise.

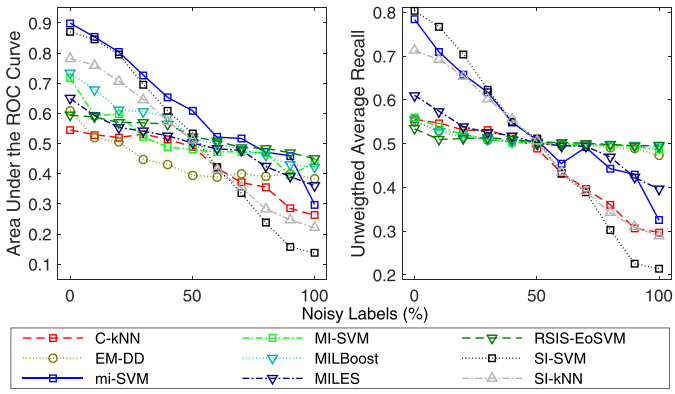

Fig. 18. Average performance of the MIL algorithms for instance classification on the SIVAL data with increasing label noise.

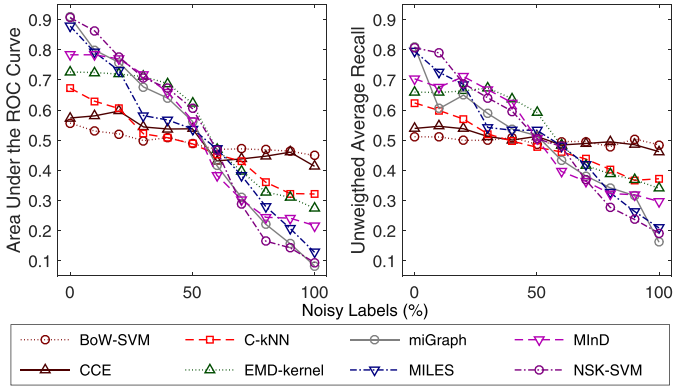

Fig. 19. Average performance of the bag-space MIL algorithms for bag classification on the Letters data with increasing label noise.

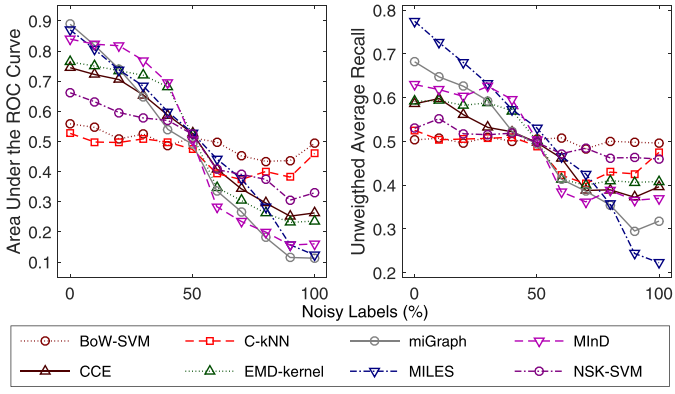

Fig. 20. Average performance of the bag-space MIL algorithms for bag classification on the SIVAL data with increasing label noise.

All methods in this experiment use an SVM which is known to be vulnerable to label noise [172]. Since all classifiers are SVMs, it is easier to compare embedding techniques. The performance curve shapes show which type of embedding is the most noise resistant. MInD and EMD-kernel both maintain their level of performance until there is 30% of mislabeled bags, while the performance of MILES, NSK-SVM and miGraph steadily decreases as the noise increases. MInD and EMD-kernel describe bags as distances between the other bags in a kernel. EMD-kernel computes the distance between distribution of instances, and MInD averages the minimal distance between all instances, which can also be seen as a distance between the two distributions. CCE also represents instance distributions in bags and exhibits a similar noise resistance in the experiments on SIVAL. Based on these observations, it would seem that *characterizing bags as instance distributions is a successful strategy to deal with label noise.*

While embedding methods characterize the distribution of instances in bags, MIL methods working under the standard MIL assumption (e.g. mi-SVM, MILBoost and MI-SVM) use a different approach. These *instance-space methods learn to identify witnesses as a step toward bag classification. In that case, the positive and the negative class are not equivalent.* This is shown by the asymmetry of the performance curves in Figs. 21 and 22. For most of these methods, when a majority of labels are inverted performance tends towards random classification. For instance-space methods, positive concepts must be cohesive and shared between positive bags while excluded from negative bags. When positive bags are mislabeled, positive instances are found in negative bags which makes the identification of the positive concept difficult. This is why *instance-space methods are more vulnerable to noise*. As shown in Figs. 21 and 22, the performance of all methods steadily degrades if the label noise level is over 10%. This is related to the instance classification performance degradation observed in Figs. 17 and 18. The experiments did not reveal a strategy that is more noise resistant than the others for instance classification.

In a nutshell, bag-space and instance-space methods differ in their dependency on the identification of positive concept. This identification process highly relies on the correctness of the bag labels which hinders the performance of instance-space methods in noisy problems.

7. Discussion

The problem characteristics identified in this paper allow for a discussion on validation procedures of MIL algorithms. The discussion is also based on the observations from the experiments in the previous secti on. After discussing practical considerations for MIL like available softwares and the complexity of MIL methods, we identify interesting research avenues for MIL.

7.1. Benchmarks data sets

Several characteristics inherent to MIL problems were discussed in this paper. It has been established that algorithms perform differently depending on these characteristics. However, even to this day, many approaches are validated only with the Musk and Tiger/Elephant/Fox (TEF) data sets. There are several problems with these benchmark data sets. First, they pose only some of the challenges discussed earlier. For example, the WR of these data sets is high. Since the instance labels are not supplied, the real

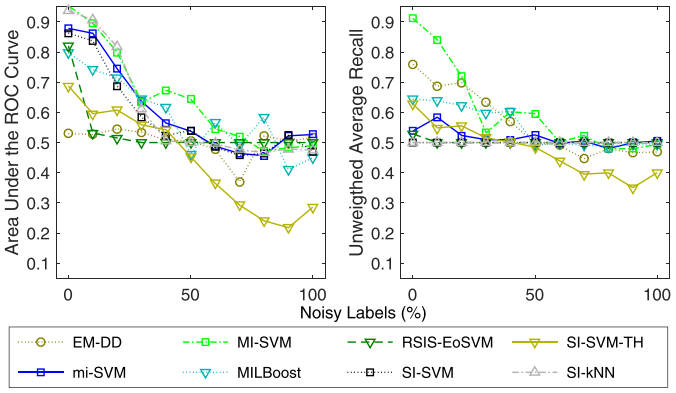

Fig. 21. Average performance of the instance-space MIL algorithms for bag classification on the Letters data with increasing label noise.

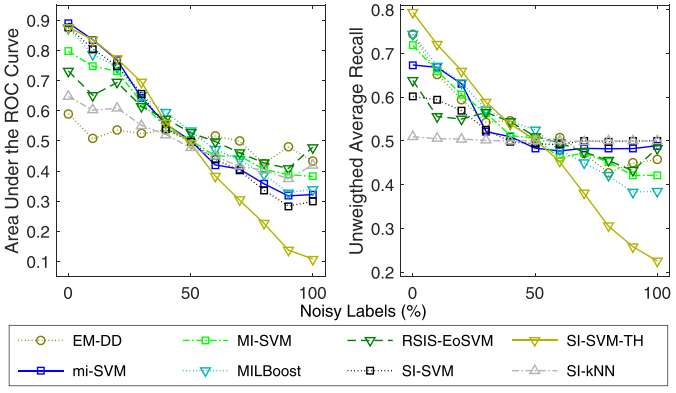

Fig. 22. Average performance of the instance-space MIL algorithms for bag classification on the SIVAL data with increasing label noise.

WR is unknown. However, it has been estimated in some papers [24,28,75] which reported 82–100% for Musk1, 23–90% for Musk2 and 38–100% for TEF. Moreover, in the Musk data sets, there is no explicit structure to be exploited. In the TEF data sets, the instances are represented by 230-dimensional feature vectors characterizing by color, texture and shape descriptors. No further details are given on these features, except that this representation is sub-optimal and should be further investigated [6]. It is possible that the theoretical Bayesian error has already been reached for this feature representation and that better results are obtained on account of protocol related technicality, such as fold partitions. Also, since the annotations at instance level are not available, it is difficult to assess if the fox classifier really identifies foxes, or if it identifies background elements related to foxes such as forest segments. This would explain the high WR estimated in [24,28,75]. For all these reasons, in our opinion, while the Musk and TEF data sets are representative of some problems, using more diverse benchmarks would provide a more meaningful comparison of MIL algorithms.

Because of the aforementioned TEF shortcomings, researchers should use more appropriate benchmark data for computer vision tasks. For example, several methods have been compared on the SIVAL data set. It contains different objects captured in the same environments, and provides labels for instances. In each image, the objects of interest are segmented into several parts. The algorithms ability to leverage co-occurrence can thus be measured, and since the objects are all captured in the same environments, the background instances do not interfere in the classification process. However, it would be more beneficial for the MIL community to use other existing strongly annotated computer vision data sets (e.g. Pascal VOC [173] or ImageNet [174]) as benchmarks. These types of data set provide bounding boxes or even pixel-level annotations that can be used to create instance labels in MIL problems. MIL algorithms could be compared to other types of techniques, which is almost never done in the MIL literature. Also, supplying the position of instances in images for these new computer vision MIL benchmarks would help to develop and compare methods that leverage spatial structure in bags.

In application fields other than computer vision, there are relatively few publicly available real-world data sets. From these few data sets, to our knowledge, there is only one (Birds [12]) that supplies instance labels and is non-artificial. This is understandable since MIL is often used to avoid the labor-intensive instance labeling process. Nevertheless, real-world MIL data needs to be created to measure the instance labeling capability of different MIL methods, as it is an increasingly important task. Also, to our knowledge, there is no publicly available benchmark data set for MIL regression, which would surely stimulate research on this subject.

Finally, several methods are validated using semi-artificial data sets. These data sets are useful to isolate one parameter of MIL problems, but are generally not representative of real-world data. In these data sets, instances are usually i.i.d. which almost never happens in real problems. Authors should justify the use of this type of data, clearly mention what assumptions are made and how the data sets are different from real data. As a start, Table 3 compiles the characteristics which are believed to be associated with some of the most widely used benchmark data sets, based on parameter estimation and data descriptions found in literature. These are believed to be true but would benefit from rigorous investigation in the future.

In short, whenever only the Musk and the TEF data sets are used to validate a new method, it is difficult to predict how the method will perform in different MIL problems. Moreover, because researchers are encouraged to evaluate their methods on these data sets, promising models may be dismissed too early because they do not outperform the best performing methods optimized for these benchmark data sets. We argue that a better understanding of the characteristics of the MIL data sets should be promoted, and that the community should use other data sets to compare MIL algorithms in regard of the challenges and properties of MIL problems.

7.2. Accuracy vs. AUC

While benchmark data is of paramount importance, the proper selection of performance metrics is equally important to avoid hasty conclusions. In all experiments, some algorithms have obtained contrasting performance when comparing AUC to accuracy and UAR. This has also been observed in other experiments [30]. This is an important factor that must be taken into consideration when comparing MIL algorithms.

Some algorithms (e.g. mi-SVM, SI-kNN, SI-SVM, miGraph, MILES) obtain high AUC that does not translate into high accuracy. There may be many reasons for this. Some algorithms optimize the decision thresholds based on bag accuracy, while others infer individual instance labels. In the first case, the algorithm is more prone to FN, while the latter is more prone to FP because of the asymmetric misclassification costs discussed in Section 6.4. Figs. 14 and 16 in Section 6.6 clearly illustrate this. As the negative distribution changes, the AUC remains stable for many algorithms, while accuracy decreases (e.g. miGraph, MILES, BoW-SVM). This means that the score function was still suitable for classification, but the decision threshold was no longer optimal. Considering the right end of the AUC curves in Fig. 14, where negative instances are completely sampled from a new distribution, one could conclude that miGraph performs better than RSIS-EoSVM. However, when comparing with

Table 3
Table compiling the characteristics of MIL benchmark data sets based on statement in the literature.

Benchmark MIL Data Sets	Instance labels	Low witness rate	Intra-bag similarities	Instance co-occurrence	Structure in bags	Multimodal positive distribution	Non-representative negative distribution	Label noise	Semi-Artificial
Musk [3]			√			√	√		
Tiger, Fox, Elephant [6]			√	√		√	√		
SIVAL [149]	√					√			
Birds [12]	√	√		√					
Newsgroups [149]	√	√				√			√
Corel [109]			√	√		√	√		
Messidor Diabetic Retinopathy [58]		√				√			
UCSB Breast [175]		√				√			
Biocreative [18]			√	√		√			

UAR, the inverse can be concluded. One could argue the AUC is a sufficient performance metric assuming that the decision threshold is optimized on a validation set, however, in many problems, the amount of available data is too limited for this assumption to hold. Also, in the case of instance classification, instance labels are unknown, therefore, it is not possible to perform such optimization.

In our opinion, the algorithms ability to accurately set this threshold is an important characteristic that should be measured, as well as the ability to learn a suitable score function. Therefore, an accuracy measure (e.g. accuracy, F_1-score, etc.) should always be reported alongside AUC.

7.3. Open source toolboxes

We think it is a good practice to report results from original papers because each method has been optimized by its own author for maximal performance. If these results are not available, some authors have published their code to allow fellow researchers to conduct more extensive experiments on other data sets. There are already several methods available from author websites [20,30,58,75,109,149]. The website of the LAMDA[4] lab is worth mentioning as it contains several implementations of MIL methods for Matlab. Other Matlab implementations of reference MIL methods can be found in the MIL toolbox [164]. There are also machine learning and data mining software packages such as Weka [176], KEEL [177] and JCLEC [178] for which MIL modules exist. Finally, the Python implementations of SVM-based MIL algorithms used in [16] are also available on-line. The wide variety of MIL problems calls for more comparative studies which will be facilitated by the availability of readily usable code. In that spirit, the code we used in our experiments has been made available on-line.[5]

7.4. Computational complexity

It has been noted by several authors that many MIL algorithms are too computationally expensive to be used with large data sets [15,179]. This represents a serious problem since one of the advantages of MIL is to increase the quantity of data available for training by leveraging weakly labeled data.

Many algorithms in literature do not scale well to big data sets. For example, the computational complexity of an SVM is between $\mathcal{O}(n^2)$ and $\mathcal{O}(n^3)$ when using traditional QP and LP solvers [180], where n is the number of instances. Thus, many methods using SVM and SVM-like algorithms [4,6,27,179,181,182] rapidly become impractical as the number of instances increases [46]. To address this problem, in [46], a bundle algorithm [183] is used to solve the SVM optimization problem in linear time ($\mathcal{O}(n)$). Alternatively, it has been proposed to use gradient descent with logistic regressions in a MILES like algorithm [184]. Gradient descent algorithms is more appropriate for large data sets than QP.

Methods computing distance between bags also become impractical as the data set size increases [15]. Obtaining the distance between two bags often means computing the distance between each pair of instances, which implies a classification cost of $\mathcal{O}(b^2m^2d)$, where b is the number of bags, m is the average number of instances per bag and d the dimensionality of the data. This becomes to $\mathcal{O}(b^2m^3d)$ when using the earth mover's distance (EMD) to compare the distributions in the two bags. Moreover, these methods must store the entire data set in memory which can also be problematic. To avoid these costs when comparing bags, it is preferable to use bag embedding techniques [72]. Representing bags as a single feature vector greatly reduces the number of training examples fed to the classifier, when compared to instance based methods. However, not all embedding methods possess the same scalability. For instance, methods representing bags as distance to instance prototypes (e.g. MILES [4]) or other bags [93] can produce very high dimensional representation with large data sets [118]. This can be avoided altogether by representing bags using a vocabulary-like encoding as proposed in [70,72]. In [95,185], hash functions have been used to accelerate the bag encoding process. Alternatively, bags can be represented by statistics on the instances as done in the Statistic Kernel (STK) [73].

While embedding methods decrease the computational cost, they generally do not allow for instance classification. In that case some methods have been proposed to reduce the size of the data set using instance selection. For example, [186] uses instance selection algorithms inspired by the immune system to reduce the size of the data set before using MIL algorithms. MILIS [118] has been proposed to reduce the complexity of MILES by selecting only one instance per bag instead of using a 1-norm SVM to perform the selection of prototypes.

Finally, parallelization can be employed to reduce computation time, like in [187], where a parallelized version of the G3P-MI [83] algorithm have been proposed to leverage the power of GPUs, and thus deal with large quantities of data.

7.5. Future direction

Based on the literature review of this survey, we identify several MIL topics that are interesting avenues for future research.

First, tasks like regression and clustering are not extensively studied when compared to classification. This might be because there are less applications for these tasks, and because there are no publicly available data sets. A good place to start exploration on MIL regression could be in affective computing applications, where the objective is to quantify abstract concepts, such as emotions and

[4] http://lamda.nju.edu.cn.
[5] https://github.com/macarbonneau/MILSurvey.

personality traits. In these applications, real-valued labels express the appreciation of human judges for speech or video sequences (bags). The sequences are represented by an ensemble of observations (instances), and it is unclear which observations contributed to the appreciation level. In this light, these problems perfectly fit in the MIL framework. Better regression algorithms would also be useful in CAD to assess the progression stage of a pathology instead of only classifying subjects as diseased or healthy.

Also, it is fairly recent that the difference between instance and bag classification is throughly investigated. It is demonstrated in [16,20], in Section 4.1 and our experiments that these tasks are different. It is showed in this paper and [34] that many instance-space methods proposed for bag classification are sub-optimal for instance classification. There is a need for MIL algorithms primarily addressing instance classification, instead of performing it as a side feature. Based on the results Section 6.4 approaches discarding or only minimally using the bag arrangement information appears to be better suited for this task. We believe that this bag arrangement could be better leveraged than how it is done by existing methods, which often seek to maximize bag-level accuracy. To further stimulate research on this topic, more instance-annotated MIL data sets are needed.

In some applications, the training data contains only positive and unlabeled data. For example, in recommender systems, the history of a user contains a list of consulted products that can be modeled as bags. If the user bought a product, it is considered as a positive bag. The other consulted products may or may not be interesting to the customer and therefore remain unlabeled. This type of problem is well studied in single instance learning [188], but requires more exploration in the MIL context. As explained before, and observed in the experiments, many MIL methods depend on the characterization of the negative distribution and the correctness of bag labels to identify positive concepts. In this case, learning from positive and unlabeled bags becomes a difficult problem for MIL. So far, only a handful of papers are dedicated to this subject [189–191].

While tasks outside bag classification would benefit from more exploration, there are also problem characteristics that necessitate the attention of the MIL community. For instance, intra-bag similarities have never been identified as a challenge, and thus, directly addressed. A possible approach could be to perform some sort of normalization or calibration in each bag to remove what is common to each instance and specific to the bag. In computer vision, this is usually done in a preliminary normalizing step. However, in other tasks such as molecule classification, this type of procedure could be helpful. For example, in the Musk data, the instances in the bag are conformations of the same molecule. Discarding the information related the “base” shape of the molecule could help to infer what more subtle particularity of the configurations is responsible for the effect when comparing to other molecules.

There are only a few methods that leverage the structure in bags. This is an important topic that has been addressed in some BoW methods, but was never thoroughly studied in other types of MIL methods, except for some methods using graphs [10,26,77,78,99]. Some of these methods represent similarities between instances or represent whole bags as graphs. Methods that create an intermediate graph representation in which some instances are grouped in sub-graphs could be an interesting way to leverage the inner structure of bags. In that case, the witnesses would correspond to ordered arrangements of instances. With this type of representation, complex objects could be identified more reliably in complex environments.

In many problems, the numbers of negative and positive instances are severely imbalanced, and yet, the existing learning methods for imbalanced data set have not studied extensively in MIL. There exist many methods to deal with imbalanced data [192]. There are external methods like SMOTE [193] and RUSBoost [194] that necessitate accurate labels to perform over or under sampling. To be adapted to MIL these methods could use some kind of probabilistic label function. Internal methods [195,196] adjust the misclassification cost independently for each class. These schemes could be used in algorithms such as mi-SVM which require the training of an SVM with high class imbalance when the WR is low. Class imbalance has also been identified in [50] as an important topic for future research.

When working with MIL, one must deal with uncertainty. It would be beneficial in many applications to use active learning to train better classifiers by querying humans about most uncertain parts of the feature space. For example, in CAD, after preliminary image classification, the algorithm would determine which are the most critical instances and prompt the clinician to provide a label. These critical instances would be the most ambiguous or the ones that would most help the classifier. This would necessitate research to assert degrees of confidence in regions of feature space. Existing literature on this subject is rather limited [149,197–199]. Alternatively, the algorithm should be able to evaluate the information gain that each instance label would provide. As a related topic, new methods should be proposed to incorporate knowledge from external and reliable sources. Intuitively, the information obtained with strong labels should have more importance in the MIL algorithm’s learning and decision process than instances with weak labels.

Except for a few papers, MIL methods always focus on the classification/regression stage, and features are considered as immutable parameters of the problem. Recently, methods for representation learning [200] have gained in popularity because they usually yield a high level of accuracy. Some of these methods learn features in a supervised manner to obtain a more discriminative representation [201], or, in deep learning, a supervised training phase is often used to fine tune the features learned in an unsupervised manner [202]. This cannot be done directly in MIL because of the uncertainty on the labels. The adaptation of discriminative feature learning methods would be beneficial to MIL. Also, it has be shown that mid-level representation help to bridge the semantic gap between low-level features and concepts [203–205]. These methods obtain a mid-level representation using supervised learning on images or videos annotated with bounding boxes. Learning techniques for these mid-level representations should also be proposed for MIL. This is an area where multiple instance clustering would be useful. There are already a few papers on this promising subject [129,140]. However, there are still a lot of open questions and limitations to overcome, such as dealing with multiple objects in a single image or the dependency to a saliency detector.

In some applications, like emotion or complex event recognition from videos, objects are represented using different modalities. For example, the voice and facial expression of a subject can be used to analyze its behavior or emotional state [206]. Alternatively, events in videos can be represented, among others, by frame, texture and motion descriptors [207,208]. In both cases, a video sequence is represented by a feature vector collection corresponding to a bag. The difference with existing MIL problems is that these instances belong to different feature spaces. This is analogous to multi-view MIL which has been studied in a few papers [209–212]. This interesting problem necessitates more research from the MIL community, and will find applications in many areas, such as multimedia analysis or problems related to the Internet-of-things, which necessitate the fusion of diverse information sources. By their nature these applications imply large quantity of data, and thus MIL would allow exploiting all this information and reduce the burden of annotation. Several fusion strategies should be explored. Instances could be mapped to the same semantic space to be compared directly, graph model could be used to aggregate several het-

erogeneous descriptors or instances could be combined in pairs to create new spaces for comparison similarly to [213].

8. Conclusion

In this paper, the characteristics and challenges of MIL problems were surveyed with applications in mind. We identified four types of characteristics which define MIL problems and dictate the behavior of MIL algorithms on data sets. It is an important topic in MIL because a better knowledge of these MIL characteristics helps interpreting experimental results and may lead to the proposal of improved methods in the future.

We conducted experiments using 16 methods which represent a broad spectrum of approaches. The experiments showed that these characteristics have an important impact on performance. It was also shown that each method behaves differently given the problem characteristics. Therefore, careful characterization of problems should not be neglected when experimenting and proposing new methods. More specific conclusions have also been drawn from the experiments:

- For instance classification tasks, when the WR is relatively high, there is no need for MIL algorithms. The problem can be cast as a regular supervised problem with one-sided noise.
- For instance classification tasks, the best approaches do not use bag information (or only very lightly). Also, methods optimized using bag classification accuracy as an objective have a higher false negative rate (as the WR increases), which limits their performance for this task.
- Bag-space methods and methods assuming instances inherit their bag label yield better classification performance especially when the WR is high.
- Bag-space methods are more robust than instance-space methods in problems where the negative distribution cannot be completely represented by the training data. This was particularly true when using the minimal Hausdorff distance.
- Embedding-space methods are generally robust to label noise, while instance-space methods are not.
- Measuring performance only in terms of AUC is misleading. Some algorithms learn an accurate score function, but fail to optimize the decision threshold used to obtain hard labels, and thus, yield low accuracy.

After observing how problem characteristics impact MIL algorithms, we discussed the necessity of using more benchmark data sets than the Musks and Tiger, Elephant and Fox data sets to compare proposed MIL algorithms. It became evident that appropriate benchmark data sets should be selected based on the characteristics of the problem to be solved. We then identified promising research avenues to explore in MIL. For example, we found that only few papers address MIL regression and clustering, which is useful in emerging applications such as affective computing. Also, more methods leveraging structure among instances should be proposed. These methods are in high demand in the era of the Internet of things, where large quantities of time series data are generated. Finally, methods dealing efficiently with large amounts of data, multiple modalities and class imbalance require further investigation.

Acknowledgments

This work was supported by the Natural Sciences and Engineering Research Council of Canada (NSERC).

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.patcog.2017.10.009.

References

[1] J. Hoffman, D. Pathak, T. Darrell, K. Saenko, Detector discovery in the wild: joint multiple instance and representation learning, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[2] J. Wu, Y. Yu, C. Huang, K. Yu, Deep multiple instance learning for image classification and auto-annotation, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[3] T.G. Dietterich, R.H. Lathrop, T. Lozano-Pérez, Solving the multiple instance problem with axis-parallel rectangles, Artif. Intell. 89 (1–2) (1997) 31–71.
[4] Y. Chen, J. Bi, J.Z. Wang, MILES: multiple-instance learning via embedded instance selection, IEEE Trans. Pattern Anal. Mach. Intell. 28 (12) (2006) 1931–1947.
[5] R. Rahmani, S.A. Goldman, MISSL: multiple-instance semi-supervised learning, in: Proceedings of International Conference on Machine Learning, ICML, 2006.
[6] S. Andrews, I. Tsochantaridis, T. Hofmann, Support vector machines for multiple-instance learning, in: Proceedings of Conference on Neural Information Processing Systems, NIPS, 2002.
[7] Q. Zhang, S.A. Goldman, W. Yu, J. Fritts, Content-based image retrieval using multiple-instance learning, in: Proceedings of International Conference on Machine Learning, ICML, 2002.
[8] S. Phan, D.-D. Le, S. Satoh, Multimedia event detection using event-driven multiple instance learning, in: Proceedings of ACM International Conference on Multimedia, ACMMM, 2015.
[9] R.G. Cinbis, J. Verbeek, C. Schmid, Weakly supervised object localization with multi-fold multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 39 (1) (2017) 189–203, doi:10.1109/TPAMI.2016.2535231.
[10] Z.-H. Zhou, Y.-Y. Sun, Y.-F. Li, Multi-instance learning by treating instances as non-I.I.D. samples, in: Proceedings of International Conference on Machine Learning, ICML, 2009.
[11] R. Bunescu, R. Mooney, Learning to extract relations from the web using minimal supervision, in: Proceedings of Association for Computational Linguistics, ACL, 2007.
[12] F. Briggs, X.Z. Fern, R. Raich, Rank-loss support instance machines for MIML instance annotation, in: Proceedings of Conference on Knowledge Discovery and Data Mining, KDD, 2012.
[13] Z.-h. Zhou, Multi-Instance Learning: A Survey, Technical Report, 2004.
[14] B. Babenko, Multiple Instance Learning: Algorithms and Applications, Technical Report, San Diego, USA, 2008.
[15] J. Amores, Multiple instance classification: review, taxonomy and comparative study, Artif. Intell. 201 (2013) 81–105.
[16] G. Doran, S. Ray, A theoretical and empirical analysis of support vector machine methods for multiple-Instance classification, Mach. Learn. 97 (1–2) (2014) 79–102.
[17] J. Foulds, E. Frank, A review of multi-instance learning assumptions, Knowl. Eng. Rev. 25 (1) (2010) 1–25.
[18] S. Ray, M. Craven, Supervised versus multiple instance learning: an empirical comparison, in: Proceedings of International Conference on Machine Learning, ICML, 2005.
[19] V. Cheplygina, D.M. Tax, M. Loog, On classification with bags, groups and sets, Pattern Recognit. Lett. 59 (2015) 11–17.
[20] G. Vanwinckelen, V. Tragante do O, D. Fierens, et al., Instance-level accuracy versus bag-level accuracy in multi-instance learning, Data Min. Knowl. Discov. 30 (2) (2016) 313–341.
[21] E. Alpaydin, V. Cheplygina, M. Loog, D.M. Tax, Single- vs. multiple-instance classification, Pattern Recognit. 48 (9) (2015) 2831–2838.
[22] V. Cheplygina, L. Sørensen, D.M.J. Tax, M. Bruijne, M. Loog, Label stability in multiple instance learning, in: Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI, 2015.
[23] V. Cheplygina, D.M.J. Tax, Characterizing multiple instance datasets, in: Proceedings of International Workshop on Similarity-Based Pattern Recognition, SIMBAD, 2015.
[24] F. Li, C. Sminchisescu, Convex multiple-instance learning by estimating likelihood ratio, in: Proceedings of Conference on Neural Information Processing Systems, NIPS, 2010.
[25] Y. Han, Q. Tao, J. Wang, Avoiding false positive in multi-instance learning, in: Proceedings of Conference on Neural Information Processing Systems, NIPS, 2010.
[26] S. Yan, X. Zhu, G. Liu, et al., Sparse multiple instance learning as document classification, Multimed. Tools Appl. 76 (3) (2017) 4553–4570.
[27] R.C. Bunescu, R.J. Mooney, Multiple instance learning for sparse positive bags, in: International Conference on Machine Learning, ICML, 2007.
[28] Y. Li, D.M. Tax, R.P. Duin, M. Loog, Multiple-instance learning as a classifier combining problem, Pattern Recognit. 46 (3) (2013) 865–874.
[29] O. Maron, T. Lozano-Pérez, A framework for multiple-instance learning, in: Proceedings of Conference on Neural Information Processing Systems, NIPS, 1998.
[30] M.-A. Carbonneau, E. Granger, A.J. Raymond, G. Gagnon, Robust multiple-instance learning ensembles using random subspace instance selection, Pattern Recognit. 58 (2016) 83–99.
[31] Y. Xiao, B. Liu, Z. Hao, A sphere-description-based approach for multiple-instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 39 (2) (2017) 242–257, doi:10.1109/TPAMI.2016.2539952.
[32] N. Weidmann, E. Frank, B. Pfahringer, A two-level learning method for generalized multi-instance problems, in: Proceedings of European Conference on Machine Learning, ECML, 2003.

[33] G. Doran, Multiple Instance Learning from Distributions, Case Western Reserve University, 2015 Ph.D. thesis.
[34] M.-A. Carbonneau, E. Granger, G. Gagnon, Decision threshold adjustment strategies for increased accuracy in multiple instance learning, in: Proceedings of International Conference on Image Processing Theory, Tools and Applications, IPTA, 2016.
[35] Q. Zhang, S.A. Goldman, EM-DD: an improved multiple-instance learning technique, in: Proceedings of Conference on Neural Information Processing Systems, NIPS, 2001.
[36] Z.-H. Zhou, M.-L. Zhang, Solving multi-instance problems with classifier ensemble based on constructive clustering, Knowl. Inf. Syst. 11 (2) (2007) 155–170.
[37] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, Z. Wang, Joint multi-label multi-instance learning for image classification, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2008.
[38] Z.-H. Zhou, M.-L. Zhang, S.-J. Huang, Y.-F. Li, Multi-instance multi-label learning, Artif. Intell. 176 (1) (2012) 2291–2320.
[39] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, S. Vluymans, Multiple Instance Multiple Label Learning, Springer, pp. 209–230.
[40] D.R. Dooly, Q. Zhang, S.A. Goldman, R.A. Amar, Multiple instance learning of real valued data, J. Mach. Learn. Res. 3 (2003) 651–678.
[41] S. Ray, D. Page, Multiple instance regression, in: Proceedings of International Conference on Machine Learning, ICML, 2001.
[42] Z. Wang, V. Radosavljevic, B. Han, Z. Obradovic, S. Vucetic, Aerosol optical depth prediction from satellite observations by multiple instance regression, in: Proceedings of SIAM International Conference on Data Mining, SDM, 2008.
[43] K.L. Wagstaff, T. Lane, Salience assignment for multiple-instance regression, in: Proceedings of International Conference on Machine Learning, ICML, 2007.
[44] N. Pappas, A. Popescu-Belis, Explaining the stars: weighted multiple-instance learning for aspect-based sentiment analysis, in: Proceedings of Conference on Empirical Methods in Natural Language Processing, EMNLP, 2014.
[45] Y. EL-Manzalawy, D. Dobbs, V. Honavar, Predicting MHC-II binding affinity using multiple instance regression, IEEE/ACM Trans. Comput. Biol. Bioinform. 8 (4) (2011) 1067–1079.
[46] C. Bergeron, G. Moore, J. Zaretzki, C.M. Breneman, K.P. Bennett, Fast bundle algorithm for multiple-instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 34 (6) (2012) 1068–1079.
[47] Y. Hu, M. Li, N. Yu, Multiple-instance ranking: learning to rank images for image retrieval, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2008.
[48] M.-L. Zhang, Z.-H. Zhou, Multi-instance clustering with applications to multi-instance prediction, Appl. Intell. 31 (1) (2009) 47–68.
[49] D. Zhang, F. Wang, L. Si, T. Li, Maximum margin multiple instance clustering with applications to image and text clustering, IEEE Trans. Neural Netw. 22 (5) (2011) 739–751.
[50] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, S. Vluymans, Multiple Instance Learning: Foundation and Algorithms, Springer, 2016.
[51] G. Quellec, G. Cazuguel, B. Cochener, M. Lamard, Multiple-instance learning for medical image and video analysis, IEEE Rev. Biomed. Eng. PP (99) (2017) 1–1, doi:10.1109/RBME.2017.2651164.
[52] S. Sabato, N. Tishby, Multi-instance learning with any hypothesis class, J. Mach. Learn. Res. 13 (1) (2012) 2999–3039.
[53] M.-A. Carbonneau, E. Granger, G. Gagnon, Witness identification in multiple instance learning using random subspaces, in: Proceedings of International Conference on Pattern Recognition, ICPR, 2016.
[54] X.S. Wei, Z.H. Zhou, An empirical study on image bag generators for multi-instance learning, Mach. Learn. 105 (2) (2016) 155–198, doi:10.1007/s10994-016-5560-1.
[55] E. Nowak, F. Jurie, B. Triggs, Sampling strategies for bag-of-features image classification, in: Proceedings of European Conference on Computer Vision, Proceedings of European Conference on Computer Vision, ECCV, 2006.
[56] H. Wang, M.M. Ullah, A. Klaser, I. Laptev, C. Schmid, Evaluation of local spatio-temporal features for action recognition, in: Proceedings of British Machine Vision Conference, BMVC, 2009.
[57] R. Venkatesan, P. Chandakkar, B. Li, Simpler non-parametric methods provide as good or better results to multiple-instance learning, in: Proceedings of International Conference on Computer Vision, ICCV, 2015.
[58] M. Kandemir, F.A. Hamprecht, Computer-aided diagnosis from weak supervision: a benchmarking study., Comput. Med. Imaging Graph. 42 (2015) 44–50.
[59] B. Babenko, P. Dollár, Z. Tu, S. Belongie, Simultaneous learning and alignment: multi-instance and multi-pose learning, in: Proceedings of European Conference on Computer Vision, ECCV, 2008.
[60] W.J. Li, D.Y. Yeung, MILD: multiple-instance learning via disambiguation, IEEE Trans. Knowl. Data Eng. 22 (1) (2010) 76–89.
[61] B. Babenko, M.-H. Yang, S. Belongie, Robust object tracking with online multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1619–1632.
[62] P. Viola, J.C. Platt, C. Zhang, Multiple instance boosting for object detection, in: Proceedings of Conference on Neural Information Processing Systems, Proceedings of Conference on Neural Information Processing Systems, NIPS, 2006.
[63] P. Auer, R. Ortner, A Boosting Approach to Multiple Instance Learning.
[64] Y. Jia, C. Zhang, Instance-level semisupervised multiple instance learning, in: Proceedings of Conference on Artificial Intelligence, AAAI, 2008.
[65] C. Yang, M. Dong, J. Hua, Region-based image annotation using asymmetrical support vector machine-based multiple-instance learning, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2006.
[66] Z.-H. Zhou, X.-B. Xue, Y. Jiang, Locating regions of interest in CBIR with multi-instance learning techniques, in: Proceedings of Australian Joint Conference on Artificial Intelligence, AUS-AI, 2005.
[67] Z.-H. Zhou, J.-M. Xu, On the relation between multi-instance learning and semi-supervised learning, in: Proceedings of International Conference on Machine Learning, ICML, 2007.
[68] Y.-F. Li, J.T. Kwok, I.W. Tsang, Z.-H. Zhou, A convex method for locating regions of interest with multi-instance learning, in: Proceedings of European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases, ECML PKDD, Berlin, Heidelberg, 2009.
[69] A. Blum, A. Kalai, A note on learning from multiple-instance examples, Mach. Learn. 30 (1) (1998) 23–29.
[70] J. Amores, Vocabulary-based approaches for multiple-instance data: a comparative study, in: Proceedings of International Conference on Pattern Recognition, ICPR, 2010.
[71] G. Doran, S. Ray, Learning instance concepts from multiple-instance data with bags as distributions, in: Proceedings of Conference on Artificial Intelligence, AAAI, 2014.
[72] X.S. Wei, J. Wu, Z.H. Zhou, Scalable multi-instance learning, in: Proceedings of International Conference on Data Mining, ICD, 2014.
[73] T. Gärtner, P.A. Flach, A. Kowalczyk, A.J. Smola, Multi-instance kernels, in: Proceedings of International Conference on Machine Learning, ICML, 2002.
[74] X. Xu, E. Frank, Logistic regression and boosting for labeled bags of instances, in: Proceedings of Conference on Pacific Asia Knowledge Discovery and Data Mining, PAKDD, 2004.
[75] P. Gehler, O. Chapelle, Deterministic annealing for multiple-instance learning, in: Proceedings of International Conference on Artificial Intelligence and Statistics, AISTATS, 2007.
[76] K. Ali, K. Saenko, Confidence-rated multiple instance boosting for object detection, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2014.
[77] D. Zhang, Y. Liu, L. Si, J. Zhang, R.D. Lawrence, Multiple instance learning on structured data, in: Proceedings of Conference on Neural Information Processing Systems, NIPS, 2011.
[78] J. Wu, X. Zhu, C. Zhang, P.S. Yu, Bag constrained structure pattern mining for multi-graph classification, IEEE Trans. Knowl. Data Eng. 26 (10) (2014) 2382–2396.
[79] J. Chai, H. Chen, L. Huang, F. Shang, Maximum margin multiple-instance feature weighting, Pattern Recognit. 47 (6) (2014) 2091–2103.
[80] I. Laptev, M. Marszalek, C. Schmid, B. Rozenfeld, Learning realistic human actions from movies, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2008.
[81] A. Zafra, M. Pechenizkiy, S. Ventura, ReliefF-MI: an extension of relieff to multiple instance learning, Neurocomputing 75 (1) (2012) 210–218.
[82] I. Kononenko, Estimating Attributes: Analysis and Extensions of RELIEF, pp. 171–182.
[83] A. Zafra, S. Ventura, G3P-MI: a genetic programming algorithm for multiple instance learning, Inf. Sci. 180 (23) (2010) 4496–4513.
[84] A. Zafra, M. Pechenizkiy, S. Ventura, HyDR-MI: a hybrid algorithm to reduce dimensionality in multiple instance learning, Inf. Sci. 222 (2013) 282–301.
[85] V.C. Raykar, B. Krishnapuram, J. Bi, M. Dundar, R.B. Rao, Bayesian multiple instance learning: automatic feature selection and inductive transfer, in: Proceedings of International Conference on Machine Learning, ICML, 2008.
[86] M.-L. Zhang, Z.-H. Zhou, Improve multi-instance neural networks through feature selection, Neural Process. Lett. 19 (1) (2004) 1–10.
[87] Z.-H. Zhou, M.-L. Zhang, Neural networks for multi-instance learning, in: Proceedings of International Conference on Intelligent Information Technologies, ICIIT, 2002.
[88] W. Ping, Y. Xu, K. Ren, C.-H. Chi, F. Shen, Non-I.I.D. multi-instance dimensionality reduction by learning a maximum bag margin subspace, in: Proceedings of Conference on Artificial Intelligence, AAAI, 2010.
[89] S. Kim, S. Choi, Local dimensionality reduction for multiple instance learning, in: Proceedings of Conference on Machine Learning for Signal Processing, MLSP, 2010.
[90] J. Chai, X. Ding, H. Chen, T. Li, Multiple-instance discriminant analysis, Pattern Recognit. 47 (7) (2014) 2517–2531.
[91] Y.-Y. Sun, M.K. Ng, Z.-H. Zhou, Multi-instance dimensionality reduction, in: Proceedings of Conference on Artificial Intelligence, AAAI, 2010, pp. 587–592.
[92] F. Kang, R. Jin, R. Sukthankar, Correlated label propagation with application to multi-label learning, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2006.
[93] V. Cheplygina, D.M. Tax, M. Loog, Multiple instance learning with bag dissimilarities, Pattern Recognit. 48 (1) (2015) 264–275.
[94] G. Csurka, C.R. Dance, L. Fan, J. Willamowski, C. Bray, Visual categorization with bags of keypoints, in: Proceedings of European Conference on Computer Vision, ECCV, 2004.
[95] W. Ping, Y. Xu, J. Wang, X.-S. Hua, FAMER: making multi-instance learning better and faster, in: Proceedings of SIAM International Conference on Data Mining, SDM, 2011.
[96] H.-Y. Wang, Q. Yang, H. Zha, Adaptive P-posterior mixture-model kernels for multiple instance learning, in: Proceedings of International Conference on Machine Learning, ICML, 2008.

[97] G.J. Qi, X.S. Hua, Y. Rui, T. Mei, J. Tang, H.J. Zhang, Concurrent multiple instance learning for image categorization, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2007.
[98] M.S. Ryoo, J.K. Aggarwal, Spatio-temporal relationship match: video structure comparison for recognition of complex human activities, in: Proceedings of International Conference on Computer Vision, ICCV, 2009.
[99] A. Mcgovern, D. Jensen, Identifying predictive structures in relational data using multiple instance learning, in: Proceedings of International Conference on Machine Learning, ICML, 2003.
[100] J. Wu, S. Pan, X. Zhu, Z. Cai, Boosting for multi-graph classification, IEEE Trans. Cybern. 45 (3) (2015) 416–429.
[101] J. Bi, J. Liang, Multiple instance learning of pulmonary embolism detection with geodesic distance along vascular structure, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2007, pp. 1–8.
[102] K. Grauman, T. Darrell, The pyramid match kernel: discriminative classification with sets of image features, in: Proceedings of International Conference on Computer Vision, ICCV, 2005.
[103] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2006.
[104] D.M.J. Tax, E. Hendriks, M.F. Valstar, M. Pantic, The detection of concept frames using clustering multi-instance learning, in: Proceedings of International Conference on Pattern Recognition, ICPR, 2010.
[105] X. Guan, R. Raich, W.-K. Wong, Efficient multi-instance learning for activity recognition from time series data using an auto-regressive hidden markov model, in: Proceedings of International Conference on Machine Learning, ICML, 2016.
[106] J. Warrell, P.H.S. Torr, Multiple-instance learning with structured bag models, in: Proceedings of International Conference on Energy Minimization Methods in Computer Vision and Pattern Recognition, EMMCVPR, 2011.
[107] Z. Li, G.-H. Geng, J. Feng, J.-y. Peng, C. Wen, J.-l. Liang, Multiple instance learning based on positive instance selection and bag structure construction, Pattern Recognit. Lett. 40 (2014) 19–26.
[108] J. Wang, J.-D. Zucker, Solving the multiple-instance problem: a lazy learning approach, in: Proceedings of International Conference on Machine Learning, ICML, 2000.
[109] Y. Chen, J.Z. Wang, Image categorization by learning and reasoning with regions, J. Mach. Learn. Res. 5 (2004) 913–939.
[110] Q. Wang, L. Si, D. Zhang, A discriminative data-dependent mixture-model approach for multiple instance learning in image classification, in: Proceedings of European Conference on Computer Vision, ECCV, 2012.
[111] D.M. Tax, R.P. Duin, Learning curves for the analysis of multiple instance classifiers, in: Proceedings of International Association for Pattern Recognition, IAPR, 2008.
[112] C. Zhang, X. Chen, M. Chen, S.-C. Chen, M.-L. Shyu, A multiple instance learning approach for content based image retrieval using one-class support vector machine, in: Proceedings of International Congress on Mathematical Education, ICME, 2005.
[113] R.-S. Wu, W.-H. Chung, Ensemble one-class support vector machines for content-based image retrieval, Expert Syst. Appl. 36 (3) (2009) 4451–4459.
[114] Z. Wang, Z. Zhao, C. Zhang, Learning with only multiple instance positive bags, in: Proceedings of International Joint Conference on Neural Networks, IJCNN, 2016.
[115] W. Li, N. Vasconcelos, Multiple instance learning for soft bags via top instances, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[116] Y. Rubner, C. Tomasi, L.J. Guibas, The Earth nover's distance as a metric for image retrieval, Int. J. Comput. Vis. 40 (2) (2000) 99–121.
[117] A. Erdem, E. Erdem, Multiple-instance learning with instance selection via dominant sets, in: Proceedings of International Workshop on Similarity-Based Pattern Recognition, SIMBAD, 2011.
[118] Z. Fu, A. Robles-Kelly, J. Zhou, MILIS: multiple instance learning with instance selection, IEEE Trans. Pattern Anal. Mach. Intell. 33 (5) (2011) 958–977.
[119] S. Bandyopadhyay, D. Ghosh, R. Mitra, Z. Zhao, MBSTAR: multiple instance learning for predicting specific functional binding sites in microrna targets, Sci. Rep. 5 (2015) 8004.
[120] D. Palachanis, Using the Multiple Instance Learning Framework to Address Differential Regulation, Delft University of Technology, 2014 Master.
[121] R. Eksi, H.-D. Li, R. Menon, Y. Wen, G.S. Omenn, M. Kretzler, Y. Guan, Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data., PLoS Comput. Biol. 9 (11) (2013) 1–16.
[122] S. Vijayanarasimhan, K. Grauman, Keywords to visual categories: multiple-instance learning for weakly supervised object categorization, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2008.
[123] O. Maron, A.L. Ratan, Multiple-instance learning for natural scene classification, in: Proceedings of International Conference on Machine Learning, ICML, 1998.
[124] C. Leistner, A. Saffari, H. Bischof, MIForests: multiple-instance learning with randomized trees, in: Proceedings of European Conference on Computer Vision, ECCV, 2010.
[125] X. Song, L. Jiao, S. Yang, X. Zhang, F. Shang, Sparse coding and classifier ensemble based multi-Instance learning for image categorization, Signal Process. 93 (1) (2013) 1–11.
[126] H. Xu, S. Venugopalan, V. Ramanishka, M. Rohrbach, K. Saenko, A multi-scale multiple instance video description network, CoRR abs/1505.0 (2016) 1–14.
[127] A. Karpathy, L. Fei-Fei, Deep visual-semantic alignments for generating image descriptions, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[128] H. Fang, S. Gupta, F. Iandola, R.K. Srivastava, L. Deng, P. Dollar, J. Gao, X. He, M. Mitchell, J.C. Platt, C. Lawrence Zitnick, G. Zweig, From captions to visual concepts and back, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2015.
[129] J.Y. Zhu, J. Wu, Y. Xu, E. Chang, Z. Tu, Unsupervised object class discovery via saliency-Guided multiple class learning, IEEE Trans. Pattern Anal. Mach. Intell. 37 (4) (2015) 862–875.
[130] H.O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, T. Darrell, On learning to localize objects with minimal supervision, in: Proceedings of International Conference on Machine Learning, ICML, 2014.
[131] B. Babenko, M.-H. Yang, S. Belongie, Robust object tracking with online multiple instance learning, IEEE Trans. Pattern Anal. Mach. Intell. 33 (8) (2011) 1619–1632.
[132] M. Sapienza, F. Cuzzolin, P.H.S. Torr, Learning discriminative space–time action parts from weakly labelled videos, Int. J. Comput. Vis. 110 (1) (2014) 30–47.
[133] A. Müller, S. Behnke, Multi-instance methods for partially supervised image segmentation, in: Proceedings of International Association for Pattern Recognition, IAPR, 2012, pp. 110–119.
[134] B. Hariharan, P. Arbeláez, R. Girshick, J. Malik, Simultaneous detection and segmentation, in: Proceedings of European Conference on Computer Vision, ECCV, 2014.
[135] A. Vezhnevets, J.M. Buhmann, Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2010.
[136] K.T. Lai, F.X. Yu, M.S. Chen, S.F. Chang, Video event detection by inferring temporal instance labels, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2014.
[137] J. Wang, B. Li, W. Hu, O. Wu, Horror video scene recognition via multiple-instance learning, in: Proceedings of International Conference on Acoustics, Speech and Signal Processing, ICASSP, 2011.
[138] K. Zhang, H. Song, Real-time visual tracking via online weighted multiple instance learning, Pattern Recognit. 46 (1) (2013) 397–411.
[139] H. Lu, Q. Zhou, D. Wang, R. Xiang, A co-training framework for visual tracking with multiple instance learning, in: Proceedings of International Conference on Automatic Face & Gesture Recognition and Workshops, FG'11, 2011.
[140] J. Zhu, B. Wang, X. Yang, W. Zhang, Z. Tu, Action recognition with actons, in: Proceedings of International Conference on Computer Vision, ICCV, 2013.
[141] Y. Xu, et al., Weakly supervised histopathology cancer image segmentation and classification, MedIA 18 (3) (2014) 591–604.
[142] G. Quellec, et al., A multiple-instance learning framework for diabetic retinopathy screening, MedIA 16 (6) (2012) 1228–1240.
[143] T. Tong, R. Wolz, Q. Gao, R. Guerrero, J.V. Hajnal, D. Rueckert, A.D.N. Initiative, et al., Multiple instance learning for classification of dementia in brain mri, Med. Image Anal. 18 (5) (2014) 808–818.
[144] J. Melendez, et al., A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays, Trans. Med. Imaging 31 (1) (2014) 179–192.
[145] V. Cheplygina, L. Sørensen, D.M.J. Tax, J.H. Pedersen, M. Loog, M. de Bruijne, Classification of COPD with multiple instance learning, in: Proceedings of International Conference on Pattern Recognition, ICPR, 2014.
[146] Z.S. Harris, Distributional structure., Word 10 (1954) 146–162.
[147] Y. Zhang, A.C. Surendran, J.C. Platt, M. Narasimhan, Learning from multi-topic web documents for contextual advertisement, in: Proceedings of Conference on Knowledge Discovery and Data Mining, KDD, 2008.
[148] D. Zhang, J. He, R. Lawrence, Mi2ls: multi-instance learning from multiple informationsources, in: Proceedings of Conference on Knowledge Discovery and Data Mining, KDD, 2013.
[149] B. Settles, M. Craven, S. Ray, Multiple-instance active learning, in: Proceedings of Conference on Neural Information Processing Systems, NIPS, 2008.
[150] Z. Jorgensen, Y. Zhou, M. Inge, A multiple instance learning strategy for combating good word attacks on spam filters, J. Mach. Learn. Res. 9 (2008) 1115–1146.
[151] D. Kotzias, M. Denil, P. Blunsom, N. de Freitas, Deep multi-instance transfer learning, CoRR abs/1411.3 (2014) 1–9.
[152] D. Kotzias, M. Denil, N. de Freitas, P. Smyth, From group to individual labels using deep features, in: Proceedings of Conference on Knowledge Discovery and Data Mining, KDD, 2015.
[153] Z.-H. Zhou, K. Jiang, M. Li, Multi-instance learning based web mining, Appl. Intell. 22 (2) (2005) 135–147.
[154] A. Zafra, S. Ventura, E. Herrera-Viedma, C. Romero, Multiple instance learning with genetic programming for web mining, Comput. Ambient Intell. 4507 (2007) 919–927.
[155] M.I. Mandel, D.P.W. Ellis, Multiple-Instance Learning for Music information Retrieval, 2008.
[156] R.F. Lyon, Machine hearing: an emerging field [exploratory DSP], Signal Process. Mag. IEEE 27 (5) (2010) 131–139.
[157] J.F. Ruiz-Muñoz, M. Orozco-Alzate, G. Castellanos-Dominguez, Multiple instance learning-based birdsong classification using unsupervised recording segmentation, in: Proceedings of International Joint Conference on Artificial Intelligence, IJCAI, 2015.
[158] M.-A. Carbonneau, E. Granger, Y. Attabi, G. Gagnon, Feature learning from spectrograms for assessment of personality traits, IEEE Trans. Affective Comput. PP (99) (2017) 1–10, doi:10.1109/TAFFC.2017.2763132.

[159] A. Kumar, B. Raj, Weakly supervised scalable audio content analysis, 2016 IEEE International Conference on Multimedia and Expo (ICME), Seattle, WA, 2016, pp. 1–6, doi:10.1109/ICME.2016.7552989.
[160] M. Stikic, D. Larlus, S. Ebert, B. Schiele, Weakly supervised recognition of daily life activities with wearable sensors, IEEE Trans. Pattern Anal. Mach. Intell. 33 (12) (2011) 2521–2537.
[161] J.F. Murray, G.F. Hughes, K. Kreutz-Delgado, Machine learning methods for predicting failures in hard drives: A Multiple-Instance application, J. Mach. Learn. Res. 6 (2005) 783–816.
[162] A. Manandhar, K.D. Morton, L.M. Collins, P.A. Torrione, Multiple instance learning for landmine detection using ground penetrating radar, in: Proceedings of SPIE, 2012.
[163] A. Karem, H. Frigui, A multiple instance learning approach for landmine detection using ground penetrating radar, in: Proceedings of International Geoscience and Remote Sensing Symposium, IGARSS, 2011.
[164] D. Tax, V. Cheplygina, MIL, A Matlab Toolbox for Multiple Instance Learning, 2015, Version 1.1.0. https://prlab.tudelft.nl/david-tax/mil.html.
[165] J.H. Friedman, Greedy function approximation: a gradient boosting machine, Ann. Stat. 29 (5) (2001) 1189–1232.
[166] R. Rahmani, S.A. Goldman, H. Zhang, J. Krettek, J.E. Fritts, Localized content based image retrieval, in: Proceedings of Conference of the Special Interest Group on Multimedia, SIGMM, 2005.
[167] K. Lang, Newsweeder: learning to filter netnews, in: Proceedings of International Conference on Machine Learning, ICML, 1995.
[168] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, D. Whiteson, Parameterized machine learning for high-energy physics, (2016) 1–6, doi:10.1140/epjc/s10052-016-4099-4.
[169] P.W. Frey, D.J. Slate, Letter recognition using holland-style adaptive classifiers, Mach. Learn. 6 (2) (1991) 161–182.
[170] M. Stone, Cross-validatory choice and assessment of statistical predictions, J. R. Stat. Soc. Ser. B (Methodol.) 36 (2) (1974) 111–147.
[171] J. Demsar, Statistical comparisons of classifiers over multiple data sets, J. Mach. Learn. Res. 7 (2006) 1–30.
[172] B. Frenay, M. Verleysen, Classification in the presence of label noise: a survey, IEEE Trans. Neural Networks Learn. Syst. 25 (5) (2014) 845–869.
[173] M. Everingham, L. Van Gool, C.K. Williams, J. Winn, A. Zisserman, The PASCAL visual object classes (VOC) challenge, Int. J. Comput. Vis. 88 (2) (2010) 303–338.
[174] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A.C. Berg, L. Fei-Fei, Imagenet large scale visual recognition challenge, Int. J. Comput. Vis. 115 (3) (2015) 211–252.
[175] M. Kandemir, C. Zhang, F.A. Hamprecht, Empowering multiple instance histopathology cancer diagnosis by cell graphs, Proceedings of International Conference on Medical Image Computing and Computer-Assisted Intervention, MICCAI (2014).
[176] M. Hall, E. Frank, G. Holmes,, B. Pfahringer, P. Reutemann, I.H. Witten, The WEKA Data Mining Software: An Update, SIGKDD Explor. Newsl. 11 (1) (2009) 10–18.
[177] J. Alcala-Fdez, A. Fernandez, J. Luengo, J. Derrac, S. Garcia, L. Sanchez, F. Herrera, KEEL Data-mining software tool: data set repository, integration of algorithms and experimental analysis framework, J. Mult. Log. Soft Comput. 17 (2–3) (2011) 255–287.
[178] S. Ventura, C. Romero, A. Zafra, J.A. Delgado, C. Hervas, Jclec: a java framework for evolutionary computation, Soft Comput. 12 (4) (2008) 381–392.
[179] G.M. Fung, M. Dundar, B. Krishnapuram, R.B. Rao, Multiple instance learning for computer aided diagnosis, in: Proceedings of Conference on Neural Information Processing Systems, NIPS, 2007.
[180] L. Bottou, O. Chapelle, D. DeCoste, J. Weston, Support Vector Machine Solvers, MIT Press, pp. 1–27.
[181] C. Bergeron, J. Zaretzki, C. Breneman, K.P. Bennett, Multiple instance ranking, in: Proceedings of International Conference on Machine Learning, ICML, 2008.
[182] O.L. Mangasarian, E.W. Wild, Multiple instance classification via successive linear programming, J. Optim. Theory Appl. 137 (3) (2008) 555–568.
[183] A. Fuduli, M. Gaudioso, G. Giallombardo, Minimizing nonconvex nonsmooth functions via cutting planes and proximity control, SIAM J. Optim. 14 (3) (2003) 743–756.
[184] Z. Fu, A. Robles-Kelly, Fast multiple instance learning via L1,2 logistic regression, in: Proceedings of International Conference on Pattern Recognition, ICPR, 2008, pp. 1–4.
[185] D. Xu, J. Wu, D. Li, Y. Tian, X. Zhu, X. Wu, SALE: self-adaptive LSH encoding for multi-instance learning, Pattern Recognit. 71 (2017) 460–482, doi:10.1016/j.patcog.2017.04.029.
[186] L. Yuan, J. Liu, X. Tang, Combining example selection with instance selection to speed up multiple-instance learning, Neurocomputing 129 (2014) 504–515.
[187] A. Cano, A. Zafra, S. Ventura, Speeding up multiple instance learning classification rules on GPUs, Knowl. Inf. Syst. 44 (1) (2015) 127–145.
[188] B. Zhang, W. Zuo, Learning from positive and unlabeled examples: a survey, in: Proceedings of International Symposiums on Information Processing, ISIP, 2008.
[189] J. Wu, X. Zhu, C. Zhang, Z. Cai, Multi-instance learning from positive and unlabeled bags, in: Proceedings of Pacific-Asia Conference on Advances in knowledge Discovery and Data Mining, PAKDD, 2014.
[190] H. Bao, T. Sakai, I. Sato, M. Sugiyama, Risk minimization framework for multiple instance learning from positive and unlabeled bags, CoRR abs/1704.06767 (2017). arXiv preprint arxiv.org/abs/1704.06767.
[191] J. Wu, S. Pan, X. Zhu, C. Zhang, X. Wu, Positive and unlabeled multi-graph learning, IEEE Trans. Cybern. 47 (4) (2017) 818–829.
[192] P. Branco, L. Torgo, R.P. Ribeiro, A survey of predictive modeling on imbalanced domains, ACM Comput. Surv. 49 (2) (2016) 31:1–31:50.
[193] N.V. Chawla, K.W. Bowyer, L.O. Hall, W.P. Kegelmeyer, SMOTE: synthetic minority over-sampling technique, J. Artif. Int. Res. 16 (1) (2002) 321–357.
[194] C. Seiffert, T.M. Khoshgoftaar, J. Van Hulse, A. Napolitano, RUSBoost: a hybrid approach to alleviating class imbalance, IEEE Trans. Syst. Man Cybern. Part A Syst. Humans 40 (1) (2010) 185–197.
[195] T. Imam, K.M. Ting, J. Kamruzzaman, z-SVM: an SVM for improved classification of imbalanced data, in: Proceedings of Australasian Joint Conference on Artificial Intelligence, AJCAI, 2006.
[196] K. Veropoulos, C. Campbell, N. Cristianini, Controlling the sensitivity of support vector machines, in: Proceedings of International Joint Conference on Artificial Intelligence, IJCAI, 1999.
[197] J. Meessen, X. Desurmont, J.F. Delaigle, C.D. Vleeschouwer, B. Macq, Progressive learning for interactive surveillance scenes retrieval, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2007.
[198] J. Melendez, B. van Ginneken, P. Maduskar, R.H.H.M. Philipsen, H. Ayles, C.I. Sánchez, On combining multiple-instance learning and active learning for computer-aided detection of tuberculosis, IEEE Trans. Med. Imaging 35 (4) (2016) 1013–1024.
[199] D. Zhang, F. Wang, Z. Shi, C. Zhang, Interactive localized content based image retrieval with multiple-instance active learning, Pattern Recognit. 43 (2) (2010) 478–484.
[200] Y. Bengio, A. Courville, P. Vincent, Representation learning: a review and new perspectives, IEEE Trans. Pattern Anal. Mach. Intell. 35 (8) (2013) 1798–1828.
[201] J. Mairal, F. Bach, J. Ponce, G. Sapiro, A. Zisserman, Discriminative learned dictionaries for local image analysis, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2008.
[202] H. Larochelle, Y. Bengio, J. Louradour, P. Lamblin, Exploring strategies for training deep neural networks, J. Mach. Learn. Res. 10 (2009) 1–40.
[203] A. Hauptmann, R. Yan, W.H. Lin, M. Christel, H. Wactlar, Can high-level concepts fill the semantic gap in video retrieval? A case study with broadcast news, IEEE Trans. Multimed. 9 (5) (2007) 958–966.
[204] L.-j. Li, H. Su, L. Fei-fei, E.P. Xing, Object bank: a high-level image representation for scene classification & semantic feature sparsification, in: Proceedings of Conference on Neural Information Processing Systems, NIPS, 2010.
[205] S. Sadanand, J.J. Corso, Action bank: a high-level representation of activity in video, in: Proceedings of Conference on Computer Vision and Pattern Recognition, CVPR, 2012.
[206] F. Ringeval, A. Sonderegger, J. Sauer, D. Lalanne, Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions, in: Proceedings of International Conference on Automatic Face & Gesture Recognition and Workshops, FG'13, 2013.
[207] M. Merler, B. Huang, L. Xie, G. Hua, A. Natsev, Semantic model vectors for complex video event recognition, IEEE Trans. Multimed. 14 (1) (2012) 88–101.
[208] K. Tang, B. Yao, L. Fei-Fei, D. Koller, Combining the right features for complex event recognition, in: Proceedings of International Conference on Computer Vision, ICCV, 2013.
[209] J. Wu, X. Zhu, C. Zhang, Z. Cai, Multi-instance multi-graph dual embedding learning, in: Proceedings of International Conference on Data Mining, ICDM, 2013.
[210] J. Wu, Z. Hong, S. Pan, X. Zhu, Z. Cai, C. Zhang, Exploring features for complicated objects: cross-view feature selection for multi-instance learning, in: Proceedings of International Conference on Information and Knowledge Management, CIKM, 2014.
[211] B. Wu, E. Zhong, A. Horner, Q. Yang, Music emotion recognition by multi-label multi-layer multi-instance multi-view learning, in: Proceedings of International Conference on Mulitmedia, ICMM, 2014.
[212] C.-T. Nguyen, D.-C. Zhan, Z.-H. Zhou, Multi-modal image annotation with multi-instance multi-label LDA, in: Proceedings of International Joint Conference on Artificial Intelligence, IJCAI, 2013.
[213] H. Daumé III, Frustratingly easy domain adaptation, CoRR abs/0907.1815 (2009). arXiv preprint arxiv.org/abs/0907.1815.

Marc-André Carbonneau received a B.Eng. degree in electrical engineering in 2010 from École de technologie supérieure (Université du Québec), Montreal, where he is currently working towards his Ph.D. degree. His research interests include machine learning, multiple instance learning, reinforcement learning, computer vision, action recognition and signal processing.

Veronika Cheplygina received the Ph.D. degree from the Delft University of Technology, the Netherlands in 2015 for her thesis entitled "Dissimilarity-Based Multiple Instance Learning". She is now a postdoc at the Biomedical Imaging Group Rotterdam, Erasmus Medical Center, The Netherlands, where she works on applying machine learning algorithms to medical image analysis problems. Her research interests are centered around learning with less labels, such as multiple instance learning, transfer learning, and crowdsourcing.

Eric Granger earned Ph.D. in EE from École Polytechnique de Montréal in 2001, and worked as a Defense Scientist at DRDC-Ottawa (1999–2001), and in R&D with Mitel Networks (2001–04). He joined the École de technologie supérieure (Université du Québec), Montreal, in 2004, where he is presently Full Professor and director of LIVIA, a research laboratory focused on computer vision and artificial intelligence. His main research interests are adaptive pattern recognition, machine learning, computer vision and computational intelligence, with applications in biometrics, face recognition and analysis, video surveillance, and computer/network security.

Ghyslain Gagnon received the Ph.D. degree in electrical engineering from Carleton University, Canada in 2008. He is now an Associate Professor at École de technologie supérieure, Montreal, Canada. He is an executive committee member of ReSMiQ and Director of research laboratory LACIME, a group of 10 Professors and nearly 100 highly-dedicated students and researchers in microelectronics, digital signal processing and wireless communications. Highly inclined towards research partnerships with industry, his research aims at digital signal processing and machine learning with various applications, from media art to building energy management.