Multiple Instance Learning: A Survey of Problem Characteristics and Applications

Marc-André Carbonneau[*]
marcandre.carbonneau@gmail.com

Veronika Cheplygina[†]
v.cheplygina@tudelft.nl

Eric Granger[*]
eric.granger@etsmtl.ca

Ghyslain Gagnon[‡]
ghyslain.gagnon@etsmtl.ca

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.

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 label 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 a growing labeling effort.

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

[*]Laboratoire d'imagerie, de vision et d'intelligence artificielle, École de technologie supérieure, Montreal, Canada

[†]Biomedical Imaging Group Rotterdam, Erasmus Medical Center, Rotterdam, The Netherlands and Pattern Recognition Laboratory, Delft University of Technology, Delft, The Netherlands

[‡]Laboratoire de communications et d'intégration de la microélectronique, École de technologie supérieure, Montreal, Canada

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 use 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. 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]. Algorithms are often better suited for only one of these two types of task [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 affects 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, and negative class modeling – 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. Section 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, two broad assumptions are considered: 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 Y be the label of a bag X, defined as a set of feature vectors $X = \{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N\}$. Each instance (i.e. feature vector) $\mathbf{x}_i$ corresponds to a label y_i. The label of the bag is given by:

$$Y = \begin{cases} +1 & \text{if} \quad \exists y_i : \ y_i = +1; \\ -1 & \text{if} \quad \forall y_i : \ y_i = -1. \end{cases} \tag{1}$$

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. This will be discussed in detail 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 interaction or the accumulation of several instances. A simple representative example given by Foulds and Frank [17] is the classification of desert, sea and beach images. Images of deserts will 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. 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, it takes many cars to create a traffic jam. In this survey, the *collective assumption* designates all assumptions in which more than one instance defines bag labels.

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 instance individually. As pointed out in [32], 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. [33, 34]) 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 [35]. 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 task 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 [36], or the best fit in a regression model [37]. Other methods work under the collective assumption and use the average or a weighted combination of the instances to represent bags as a single feature vector [38–40]. Alternatively, on can simply replace a bag-level classifier by a regressor [41].

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 [42] or at the instance level [43].

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 [44]). Alternatively, clustering can be performed at the instance level. For example, in [45], 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-level methods performed better in terms of bag classification accuracy, however, performance depends on the data and the distance function or the embedding method. Finally, very recently, a book on MIL has been published [46]. It discusses most of the tasks of Section 2.2 along with associated methods, as well as data reduction and imbalanced data.

Some papers study specific topics of MIL. For instance, Foulds and Frank reviewed the assumptions [17] 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, such as the Musk data set [3]. Sabato and Tishby [47] analyzed the of sample complexity in MIL, and they 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 algorithm performances.

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 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, these differences 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 can be 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-level classifier. They also state, as in [15], that if the instances provide partial information about the bag labels, it is preferable to use bag-level 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. [48] studied the ability to identify witnesses (positive instances) of several MIL methods. They found that depending 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 [49] 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 [50, 51]. They also found that methods using the collective assumption performed better for image classification. Vankatesan et al. [52] showed that simple lazy-learning techniques could be applied to some MIL problems to obtain results comparable to state-of-the-art techniques. Kandemir and Hamprecht [53] 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, 48].
- When it is necessary to model combinations of instances to infer bag labels, bag-level and embedding methods perform better [15, 21, 49].
- 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 [47].
- Similarity between the instances of a same bag affect classification performance [53].

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.**

4 Characteristics of MIL Problems

We identified four broad categories of key characteristics associated with MIL problems which directly impacts on the behavior of MIL algorithms: *task*, *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. While these two tasks appear similar, there are key differences, and

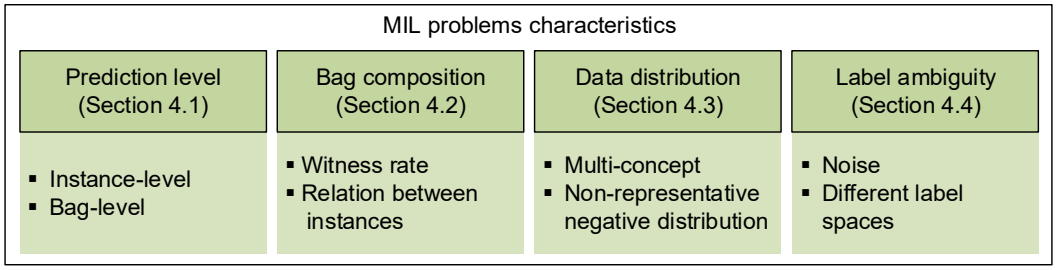

Figure 1: Characteristics inherent to MIL problems.

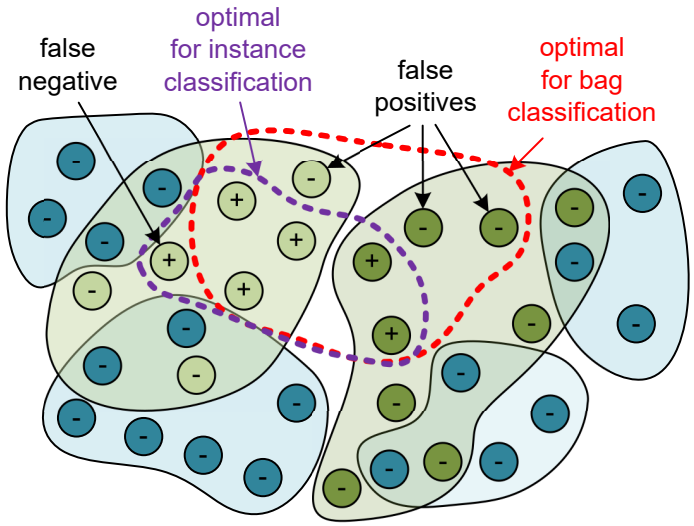

Figure 2: Illustration of two decisions boundaries on a fictive problem. While only the purple boundary correctly classifies all instances, both them achieve perfect bag classification. This is because, in that case, false positive and false negative instances do not impact on bag labels.

thus, the bag classification performance of a method often is not representative of its instance classification performance [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. This means that algorithms designed for bag classification are not optimal for instance classification. Most methods in the literature address the bag classification problem, and sometimes perform instance classification as a *side feature* (e.g. MILES [4]). One of the challenges for developing instance-level classification algorithm is the scarcity of benchmark data sets providing ground truth for instance labels.

The main difference between the two tasks is the misclassification cost of instances. 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. The individual labels of the instances are identified on each instance. 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). This is why MIL algorithms using bag accuracy as an optimization criterion (e.g. APR [3], MI-SVM [6], MIL-Boost [54], EM-DD [33], MILD [55]) can learn a suboptimal decision boundary for instance classification.

It has been proposed to consider negative and positive bags separately in the classifier loss function [56]. 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 [32]. In [57], a different weight is attributed to FP and FN during the optimization of an SVM. Some methods label all instances independently, like mi-SVM [6] and MissSVM [58]. These methods yield the best results in our experiments on instance-level classification (see Section 6.3).

4.2 Bag Composition

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 [59]. 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 [33] and APR [3] instances are considered to have the same label as their bag. When the WR is low, this is no longer reasonable and leads to lower performances. Methods which analyze instance distributions in bags [60–62] may also have problems dealing with low WR because distribution in positive and negative bags become similar. Also, some methods represent bags by the average of the instances they contain, like NSK-SVM [63], or by considering their contribution to the bag label equally [64]. 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.

Several authors studied low WR problems in recent years. For example, sparse transductive MIL (stMIL) [27] is an SVM formulation similar to NSK-SVM [63]. 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 [65], 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 [66] 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.

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 structure or correlations exist between the instances and bags [10, 67]. We make a distinction between three types of relation: intra-bag similarities, instance co-occurrences and structure.

Intra-Bag Similarities: In some problems, the instances belonging to the same bag share similarities, that instances from other bags do not share. 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 thus 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 condition (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 (Figure 3). Since

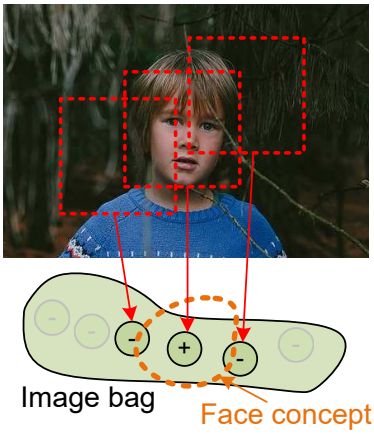

Figure 3: Illustration of intra-bag similarity between instances: The patches are overlapping, and thus, share similarities with each other.

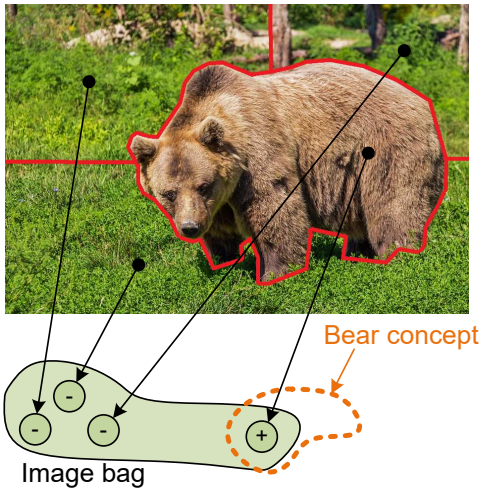

Figure 4: Example of co-occurrence and similarity between instances: Three segments contain grass and forest and are therefore very similar. Moreover, since this is an image of a bear, the background is more likely to be nature than a nuclear central control room.

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 Figure 4).

Intra-bag similarities raise some difficulties when learning. For instance, transductive algorithms (e.g. mi-SVM [6]) might not be able to infer instance labels if the negative instances from positive and negative bags differ in nature [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. In CCE [34], a binary vector represents the bags by encoding the assignation of at least one instance to a cluster. 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.

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 Figure 4 is more likely to be found in nature than in a nightclub. Thus, the observation of nature segments might help to decide if the image contains a cocktail or a bear [68].

In [69], 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 [60, 70], miFV [62], FAMER [71] or PPMM [72] represent bags as instance distributions which indirectly account for co-occurrence. This has also been directly modeled in a tensor model [73] and in a multi-label framework [74].

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, there exists an underlying structure between instances in bags or even between bags [67]. 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, 75, 76]. 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 [67] 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 to increase classification performance. Structure can be exploited at many levels: graphs can be used to model the relations between bags, instances or both [26, 67]. Graphs enforce that related objects belong to the same class. Alternatively, in [77] bags are represented by a graph capturing diverse relationships between objects. The objects are shared across all bags and all possible sub-graphs of the bag graph correspond to instances.

Temporal and spatial structure between instances can be modeled in different ways. In BoW models, this can be achieved by dividing the images [78, 79] or videos [75] 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) [80] and hidden Markov model (HMM) [81]. Spatial dependency in images have also been modeled in with CRF in [74, 82].

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.

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 follow the single cluster assumption. CKMIL [83] 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 [80] employs a similar strategy.

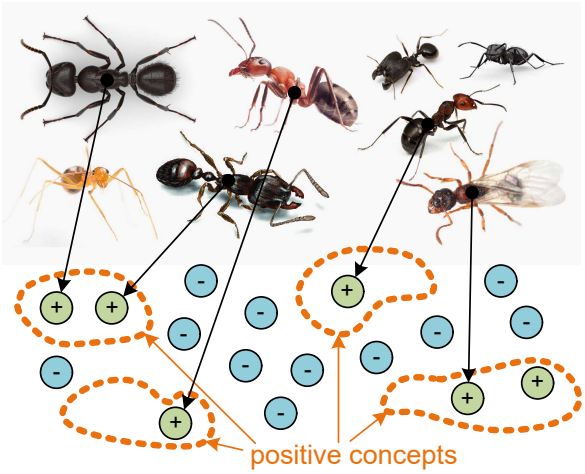

Figure 5: For the same concept *ants*, there can be many data clusters (modes) in feature space corresponding to different poses, colors and castes.

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 the 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 [84] and MInD [69] naturally deal with all shapes of distributions. Simple non-parametric methods often lead to competitive results in MIL problems [52]. Methods using distances to a set of prototypes as bag representation, like DD-SVM [85] and MILES [4], can model many positive clusters, because each different cluster can be represented by a different prototype. Instance-level 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 [60] methods naturally deal with data sets containing multiple concepts/modes. The mixture-model in [86] naturally represents different positive clusters. In [30] instances are grouped in clusters and the composition of the clusters are analyzed to compute the probability that instances are positive.

Non-Representative Negative Distribution

In [87], it is stated that learnability of instance concept 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 spaceships, 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] the region is a hyper-rectangle, while in many others it is one, or a collection of, hyper-spheres/-ellipses [29, 31, 33, 88]. 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 [84]. 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 [89–91].

Experiments in Section 6.5 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. However, there are supplementary sources of ambiguity such as noise on labels and instance labels different from bag labels.

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 [52] that DD is not tolerant to noise in the sense that a single negative instance in the neighborhood of the positive concept can hinder performances. A similar argument was made for APR [55] 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 [92]. Similar problems may arise in text classification, where a paragraph contains an analogy and thus, 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. The methods representing bags as distributions [60, 61, 93] 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 [63] 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 [92] 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 [94], robustness to label noise is obtained by using dominant sets to perform clustering and select relevant instance prototype in a bag-embedding algorithm similar to MILES [4].

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, instances 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 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 [60] 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

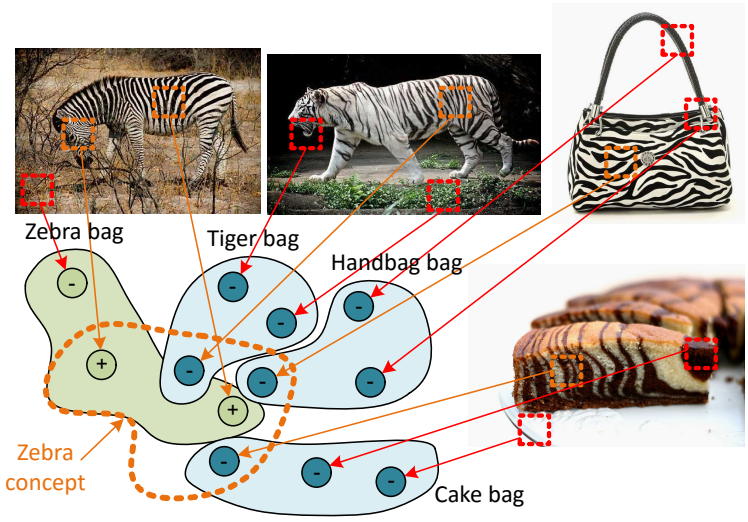

Figure 6: This is an example of instances with ambiguous labels. *Zebra* is the target concept and instances relating to this concept should fall in the region delimited by the dotted line. However, negative images can also contain instances falling inside the zebra concept region.

using embedding based on distance from selected prototype instance, such as MILES [4] and MILIS [95], can also deal with this type of problem.

All the characteristics presented in this section define a variety of MIL problem, 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 their different problem characteristics of Section 4 (summarized in Table 1).

5.1 Biology and Chemistry

The problems in biology and chemistry can often be naturally formulated as MIL problems because of the inability to observe individual instance classes. For instance, 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 parts or regions that can induce an effect of interest. The goal is to classify unknown bags and sometimes to identify witness to better understand underlying mechanisms of the biological or chemical phenomenon. MIL has been used, among others, to predict a drug's bioavailability [42], predict the binding affinity of peptides to major histocompatibility complex molecules [41], discover binding sites governing gene expression [96, 97] and predict gene functions [98].

The problems presented in this section are of various natures and it is difficult to identify key characteristics applying to all cases. However, in most cases, the bags represent many arrangements or view-points of the same entity, which translate into high intra-bag similarities. Some objects like

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)

	Problem Characteristics									
Application Fields	Instance classification	Real-valued outputs	Low witness rate	Intra-bag similarities	Instance co-occurence	Structure in bags	Multimodal positive distribution	Non-modelable 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	✓				✓	✓✓	✓	✓	✓	✓

DNA sequences produce structured bags, while the many conformations of the same molecule do not. In some problems, the objective is to identify instances responsible for an effect (e.g. drug binding). Also, many applications call for quantification, using ranking or regression, instead of classification [36] (e.g. quantifying the binding strength of a molecule), which is more difficult, or at least less documented.

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.

Content Based Image Retrieval

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, 89, 99–102]. The task in CBIR is to categorize images based on the objects/concepts they contain. The exact localization of the 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 [49]. For example, the image can be partitioned using a regular grid [100], key-points [70] or semantic regions [57, 85]. 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.

This task is subject to most of the key-challenges associated with the problem characteristics in Section 4. Images are a good example of non-i.i.d. data. 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. Some

objects are more likely to co-occur in the same picture (e.g. bird and sky). Methods leveraging these co-occurrences tend to be more successful. Sometimes the subject of a picture is a composition of several concepts, which means methods working under the collective MIL assumption perform better. Working with images often means working with large intra-class variability. For instance, the same object can appear considerably different depending on the points of view. Also, many types of object can have different shapes and colors. This means it is unlikely that a unimodal distribution adequately represents the entire class. Furthermore, backgrounds can vary a lot, making it difficult to learn a negative distribution that models every possible background object.

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. The 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 [103–105], web search engine results [106], 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, 54, 107–109]. 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 [110], JSEG [74] or Multi-scale combinatorial grouping [111]. Alternatively, segmentation can be achieved by casting each pixel of the image as an instance [112].

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, 113]. Several concepts compose these complex events. Evidence of these concepts sometimes lasts 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 [114] to recognize scenes that are inappropriate for children. Also in videos, MIL methods were proposed to perform object tracking [115–117]. For example, in [115] 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 [106]. The method is akin to multiple instance clustering methods [44, 45], 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 [118].

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, take various appearances and be seen from different viewpoints. This means that a single concept is often represented by a multimodal distribution, and the negative distribution cannot be entirely captured by a training set. Object localization is different from CBIR because it is an instance classification problem, which means that many bag-level algorithms are inapplicable. Also, several authors noted that in this context, MIL algorithms are sensitive to initialization [9, 107].

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 medical scan. The MIL framework is appropriate in this situation given that patients have both ab-

normal 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 [119], diabetes in retinal images [120], dementia in brain MR [121], tuberculosis in X-ray images [122], classification of a chronic lung disease in CT [123] 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), an image patch or a region 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 [122], 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 [123].

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 what 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 [121], 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, [123] 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 [53, 120, 122]. In addition, the output of the classifier can be displayed in the image, which is an interpretable way of visualizing the results. In [122], 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.

Like other computer vision tasks, CAD is subject to most of the key-challenges discussed in Section 4. Depending on the sampling - which can be done on a densely-sampled grid [53, 122], randomly [123], or according to constraints [121] – 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 also can lead to large intra-class variability. On the other hand, the negative distribution (healthy tissue) is more constrained than in computer vision applications. CAD problems 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 [123] treats them all as the positive class. During evaluation, the mild stage is most often misclassified as healthy. [121] 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, as for other applications, the difference between bag-level and instance-level classification presents an important challenge.

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 [124]. 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 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 [40, 125], passages [6, 126] 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 [127], 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 the protein names and key terms. In [128], 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 [40, 125, 129, 130], MIL was used to infer the sentiment expressed in individual sentences based on the 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 [131, 132]. 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 [125], a MIL classifier assesses sections of web pages to identify suitable web pages for advertisement.

The classification of web and text documents is subject to most of the difficulties associated with MIL problem characteristics. Depending on the task and the formulation of the problem, bag and instance classification can be performed. Often only small passages or specific words indicate the class of the document, which means WR can be quite low. Words may have different meanings depending on the context and thus, co-occurrence is important in this type of application. While structure is an important component of sentences, most of the existing MIL methods discard it. In addition, text classification can present an additional difficulty compared to other applications. When texts are represented by a 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 negative distribution.

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 sub-goals in a task [77]. These sub-goals are, in fact, the positive instances in the successful episodes. The main challenge for RL task 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 [133], 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 collection 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 an unattended microphone 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 [134]. This idea has been used for bird song recognition [135] with histograms

of gradients. In [136], 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 [137], however, in that case instances are cepstrum feature vectors representing 1 second-long audio segments. In general, audio classification is subject to the same challenges as image classification applications.

Time series are found in several applications other than audio classification. For instance, in [81,138] 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 [139]. 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 [140, 141], 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 [77], 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 sub-graph 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 allows to shed in 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. Three 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 and with an unobservable negative distribution. 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 a most families of approaches and include most of the most widely used reference methods.

6.1 Reference Methods

Instance Space Methods

SI-SVM, SI-SVM-TH and SI-*k*NN: These are not a MIL method *per se*, but 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. The classifier assign 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 classify 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 instances. MI-SVM uses only the most positive instance of each bag for training, while mi-SVM uses all instances.

EM-DD [33]: DD [29] measure 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 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 [48]. Training subsets are sampled using the probabilistic labels of the instance to train an ensemble of SVM.

MIL-Boost [54]: The MIL-Boost algorithm used in this paper is a generalization of the algorithm presented in [142]. The method is essentially the same as gradient boosting [143] 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.

Bag Space Methods

C-kNN [84]: 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 [144] to perform instance classification.

MInD [69]: 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 [34]: 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 a 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 [63]: 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.

mi-Graph [10]: This method represents each bag by a graph in which instances correspond to nodes. Cliques are identified in the graph to adjust the instances weights. Instances belonging to large cliques have lower weight so that every 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. 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) [93] 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.

6.2 Data Sets

Spatially Independent, Variable Area, and Lighting (SIVAL) [145]: This data set contains 500 images each segmented and manually labeled by [127]. 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 data sets are generated 5 times. 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 environment.

Birds [12]: The bags of this data set correspond to 10 seconds 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 noises. These 10232 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 are 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 [127]: The newsgroups data set was derived from the *20 Newsgroups* [146] 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 [147]: The instances of this data set come from the HEPMASS Data Set[1]. It contains more than 10M instances which are simulation of particle collisions. The positive class correspond 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 [148]: This semi-synthetic MIL data set uses instances from the Letter Recognition data set[2]. 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 create 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 are 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 μ 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%.

6.3 Instance-Level Classification

In this section, the reference methods with instance classification capabilities will be compared on three benchmark data sets: SIVAL, Birds and Newsgroups. These data sets are selected because

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

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

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, 53].

The experiments were conducted using a nested cross-fold validation protocol [149]. 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.

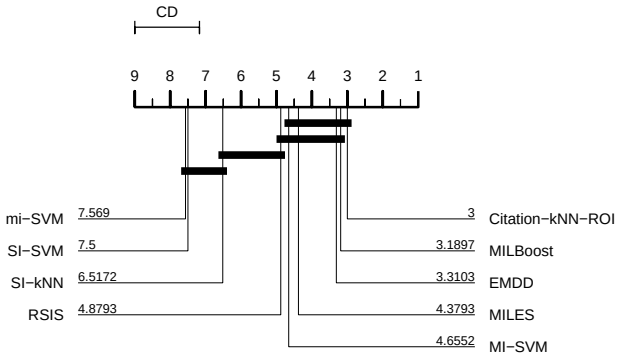

Figure 7: Critical difference diagram for UAR on instance classification ($\alpha = 0.01$). Higher numbers are better.

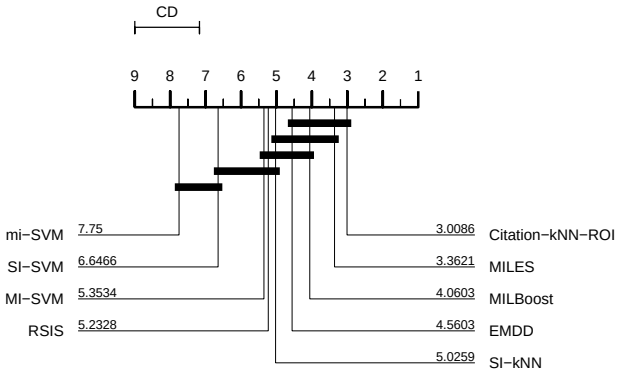

Figure 8: Critical difference diagram for the F_1-score on instance classification ($\alpha = 0.01$). Higher numbers are better.

Instance classification problems often exhibit class imbalance, especially when the WR is small. In these cases, comparing algorithm is 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, Fig. 7 and 8 present the results in the form of critical difference diagrams [150] 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 [53]. These two methods are closely related because SI-SVM corresponds to the first iteration of mi-SVM. SI-kNN also yield 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 in each bag. This strategy seems successful to some degree, but is prone to ignore more ambiguous

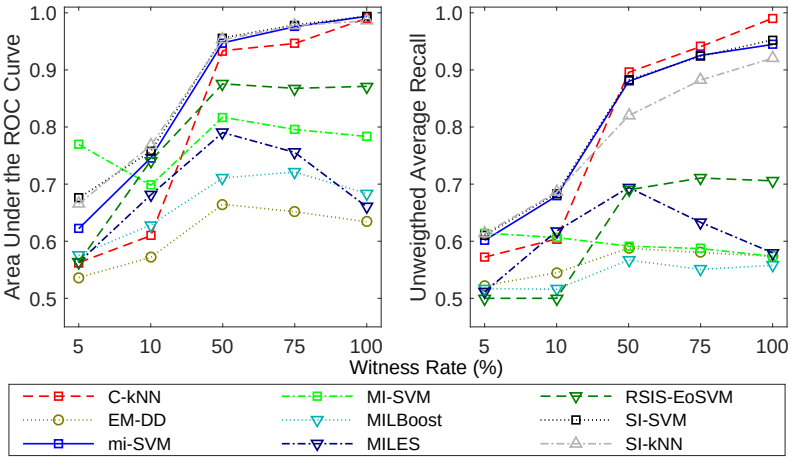

Figure 9: Average performance of the MIL algorithms for instance classification on the Letters data set as the witness rate increases.

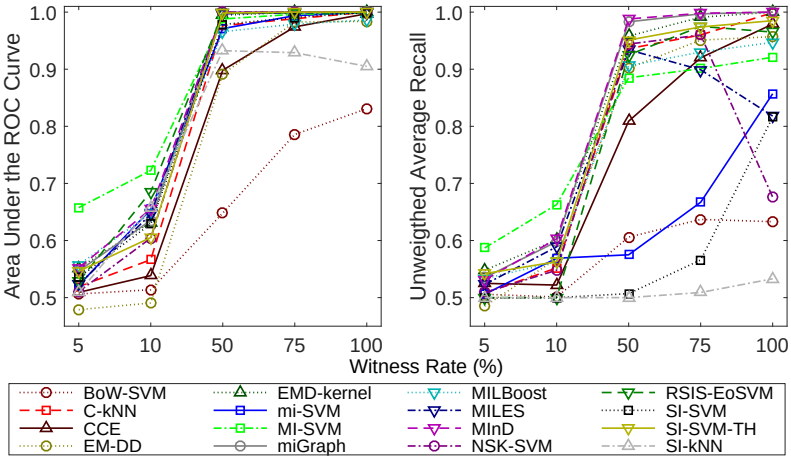

Figure 10: Average performance of the MIL algorithms for bag classification on the Letters data set as the witness rate increases.

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.

6.4 Bag Composition: Witness Rate

These experiments study the effects of the WR on MIL algorithm performances. 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 Fig. 9, 10, 11 and 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.

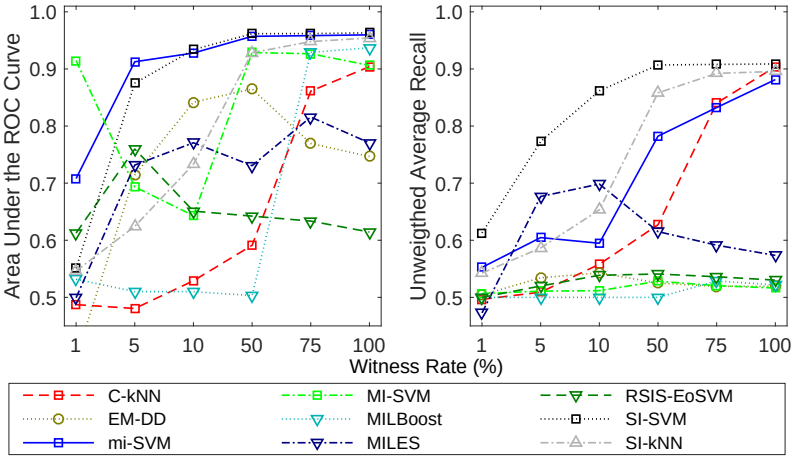

Figure 11: Average performance of the MIL algorithms for instance classification on the HEPMASS data set as the witness rate increases.

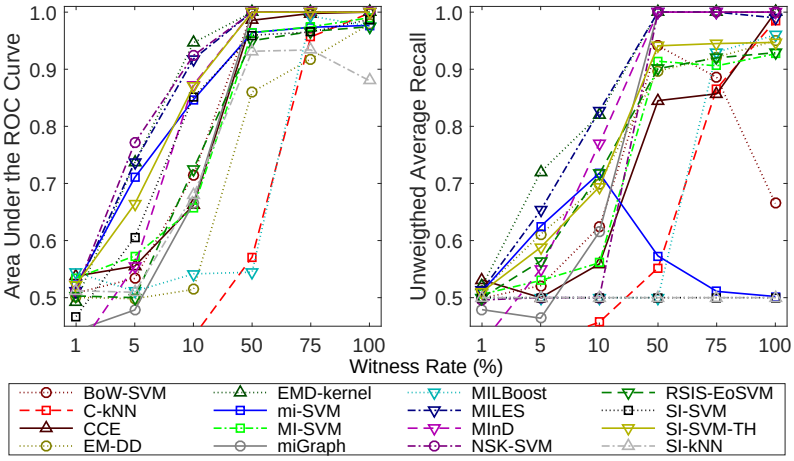

Figure 12: Average performance of the MIL algorithms for bag classification on the HEPMASS data set as the witness rate increases.

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 6.3, some methods 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 Fig. 9 and 11 with Fig. 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 Fig. 9 and 11 suggest that **supervised classifiers are as effective for instance classification as the best MIL classifiers when the WR is over 50%**. In this case, the mislabeled negative instance are just noise in the training set, which easily is 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.

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

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

The results for bag classification are reported in Fig. 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-level methods outperform their instance-based counter-parts at higher WR** ($\geq 50\%$). At lower WR ($5 \sim 10\%$), 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.

6.5 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. Two data sets are used for these experiments: the Letters data set and the synthetic Gaussian toy data set created specially for this experiment. 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 distribution. 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 letter A, B and C are used as negative instances. Gradually, the instance from A, B and C are replaced by instance on the letter D, E and F.

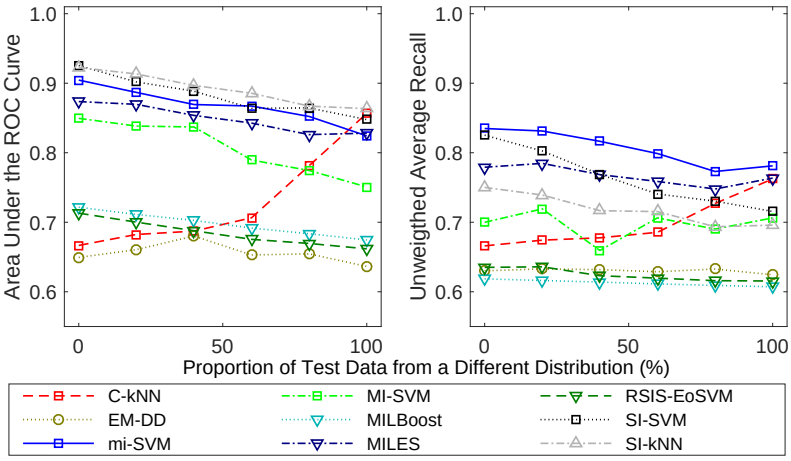

Figure 13: Average performance of the MIL algorithms for instance classification on the Letters data as the test negative instance distribution increasingly differs from the training distribution.

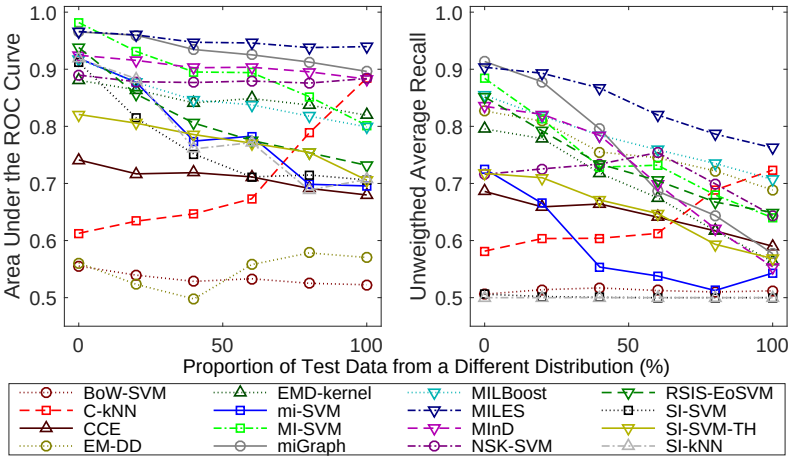

Figure 14: Average performance of the MIL algorithms for bag classification on the Letters data as the test negative instance distribution increasingly differs from the training distribution.

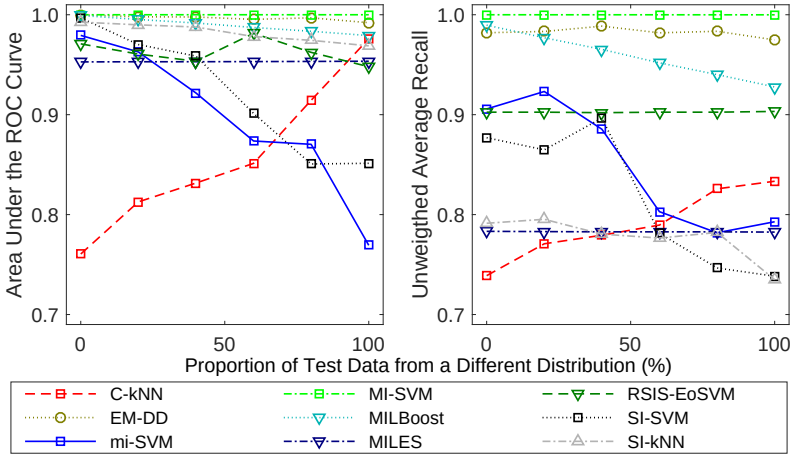

Figure 15: Average performance of the MIL algorithms for instance classification on Gaussian toy data as the test negative instance distribution increasingly differs from the training distribution.

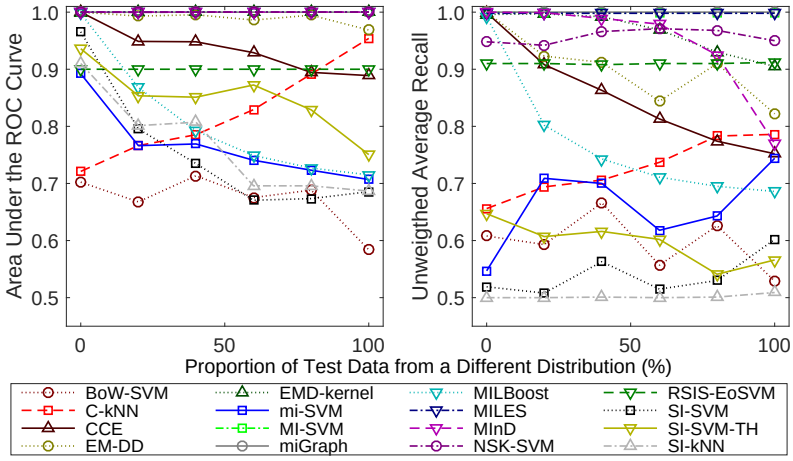

Figure 16: Average performance of the MIL algorithms for bag classification on Gaussian toy data as the test negative instance distribution increasingly differs from the training distribution.

The results of the experiments, illustrated in Fig. 13, 14, 15 and 16, show that **most algorithms have decreasing performance when the test negative instances distribution differs from the training distribution**. However, C-kNN exhibits a contrasting behavior. More the test instances differ from test to training, the better are performances. This is because C-kNN uses the minimal Hausdorff distance as a similarity 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 instance 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-level methods are better for dealing with new negative distributions**. This may contribute to their success in computer vision applications. 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 which would adjust the decision threshold as new data arrives.

7 Discussion

The problem characteristics identified in this paper allow for a discussion on validation procedures of MIL algorithms. These suggestions are also based on the observations from the experiments in the previous section. Then, we identify interesting research avenues for MIL.

7.1 Benchmarks Data Sets

Several characteristics inherent to MIL problems were discussed in this paper. Experiments confirmed what has been observed by many researchers before: algorithms perform differently depending on the type of MIL problem, and several characteristics define a MIL problem. 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 WR is unknown. However, it has been estimated in some papers [28, 65, 151] which reported 82 to 100% for Musk1, 23 to 90% for Musk2 and 38 to 100% for TEF. Moreover, in the Musk data sets, there are 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 [28, 65, 151]. Since the state-of-the-art accuracy on this class is around 70%, it is plausible that a large proportion of the animals in the negative class live in deserts or under the sea. 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 [152] or ImageNet [153]) as benchmarks. These types of data set provide bounding box 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

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-occurence	Structure in bags	Multimodal positive distribution	Non-representative negative distribution	Label noise	Semi-Artificial
Musk [3]			✓			✓	✓		
Tiger, Fox, Elephant [6]			✓	✓		✓	✓		
SIVAL [127]	✓					✓			
Birds [12]	✓	✓		✓					
Newsgroups [127]	✓	✓				✓			✓
Corel [85]			✓	✓		✓	✓		
Messidor Diabetic Retinopathy [53]		✓				✓			
UCSB Breast [154]		✓				✓			
Biocreative [18]			✓	✓		✓			

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 supply 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 task.

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 methods 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 on 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 6.3. Figure 14 and Figure 16 in Section 6.5 clearly illustrate this. As the negative distribution changes, the AUC remains stable for many algorithm, 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 Figure 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 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, accuracy should always be reported alongside AUC.

7.3 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. 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 personalities. 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 observation 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 only 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 [32] 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 6.3 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.

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. It could be beneficial 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 looked upon in other types of MIL methods, except for some methods using graphs [10, 26, 67, 77]. Some of these methods represent similarities between instances or represent whole bag as graph. 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 witness would be an ordered arrangement 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 [155]. There are external methods like SMOTE [156] and RUSBoost [157] 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 [158, 159] 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 [46] as an important topic for future research.

There are other MIL challenges that were not studied in this paper due to space constraints. For one, the computational complexity of the algorithms is important since MIL is often used to leverage large quantities of data. This is generally not a concern for bag-level methods. However, instance-level methods rapidly become difficult to use with large data sets. The elaboration of methods focused on computational efficiency would facilitate the use of MIL in large-scale applications.

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 parts of feature space. 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 instance with weak labels.

Except for a few papers, MIL methods always focus on classification/regression, and features are considered as immutable parameters of the problem. Recently, methods for representation learning [160] 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 [161], or, in deep learning, a supervised training phase is often used to fine tune the features learned in an unsupervised manner [162]. 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 [163–165]. 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 [106, 118]. 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 [166]. Alternatively, events in videos can be represented, among others, by frame, texture and motion descriptors [167, 168]. In both cases, a video sequence is represented by a collection of feature vectors, which corresponds to a bag in MIL. The difference with existing MIL problems is that these instances belong to a different feature spaces. This is an interesting problem that has yet to be addressed by the MIL community. This will be useful in rising research areas, such as multimedia analysis or problems related to the Internet-of-things, which necessitate the fusion of diverse sources of information. By their nature these applications imply large quantity of data, and thus MIL would be a perfect tool to leverage all this information and reduce the burden of annotation. Several fusion strategies should be explored. Instance could be mapped to the same semantic space to be compared directly, graph model could be used to aggregate several heterogeneous descriptors or instances could be combined in pairs to create new spaces for comparison similarly to [169].

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 experiments 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 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 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-level 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.
- 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 amount of data, multiple modalities and class imbalance require further investigation.

References

[1] J. Hoffman, D. Pathak, T. Darrell, and K. Saenko, "Detector Discovery in the Wild: Joint Multiple Instance and Representation Learning," in *CVPR*, 2015.

[2] J. Wu, Y. Yu, C. Huang, and K. Yu, "Deep multiple instance learning for image classification and auto-annotation," in *CVPR*, 2015.

[3] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez, "Solving the Multiple Instance Problem with Axis-parallel Rectangles," *Artif. Intell.*, vol. 89, no. 1-2, pp. 31–71, Jan. 1997.

[4] Y. Chen, J. Bi, and J. Z. Wang, "MILES: Multiple-Instance Learning via Embedded Instance Selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 12, pp. 1931–1947, Dec. 2006.

[5] R. Rahmani and S. A. Goldman, "MISSL: Multiple-instance Semi-supervised Learning," in *ICML*, 2006.

[6] S. Andrews, I. Tsochantaridis, and T. Hofmann, "Support Vector Machines for Multiple-Instance Learning," in *NIPS*, 2002.

[7] Q. Zhang, S. A. Goldman, W. Yu, and J. Fritts, "Content-Based Image Retrieval Using Multiple-Instance Learning," in *ICML*, 2002.

[8] S. Phan, D.-D. Le, and S. Satoh, "Multimedia Event Detection Using Event-Driven Multiple Instance Learning," in *ACMMM*, 2015.

[9] R. G. Cinbis, J. Verbeek, and C. Schmid, "Weakly Supervised Object Localization with Multi-fold Multiple Instance Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

[10] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li, "Multi-Instance Learning by Treating Instances As non-I.I.D. Samples," in *ICML*, 2009.

[11] R. Bunescu and R. Mooney, "Learning to Extract Relations from the Web using Minimal Supervision," in *ACL*, 2007.

[12] F. Briggs, X. Z. Fern, and R. Raich, "Rank-Loss Support Instance Machines for MIML Instance Annotation," in *KDD*, 2012.

[13] Z.-h. Zhou, "Multi-Instance Learning : A Survey," AI Lab, Department of Computer Science and Technology, Nanjing University, Tech. Rep., 2004.

[14] B. Babenko, "Multiple Instance Learning : Algorithms and Applications," University of California, San Diego, USA, Tech. Rep., 2008.

[15] J. Amores, "Multiple Instance Classification: Review, Taxonomy and Comparative Study," *Artif. Intell.*, vol. 201, pp. 81–105, Aug. 2013.

[16] G. Doran and S. Ray, "A Theoretical and Empirical Analysis of Support Vector Machine Methods for Multiple-Instance Classification," *Machine Learning*, vol. 97, no. 1-2, pp. 79–102, 2014.

[17] J. Foulds and E. Frank, "A Review of Multi-Instance Learning Assumptions," *Knowl. Eng. Review*, vol. 25, no. 1, pp. 1–25, Mar. 2010.

[18] S. Ray and M. Craven, "Supervised Versus Multiple Instance Learning: An Empirical Comparison," in *ICML*, 2005.

[19] V. Cheplygina, D. M. Tax, and M. Loog, "On classification with bags, groups and sets," *Pattern Recognit. Lett.*, vol. 59, pp. 11–17, Jul 2015.

[20] G. Vanwinckelen, V. do O, D. Fierens, and H. Blockeel, "Instance-level accuracy versus bag-level accuracy in multi-instance learning," *Data Mini. and Knowl. Discovery*, pp. 1–29, 2015.

[21] E. Alpaydın, V. Cheplygina, M. Loog, and D. M. Tax, "Single- vs. multiple-instance classification," *Pattern Recognit.*, vol. 48, no. 9, pp. 2831–2838, Sep 2015.

[22] V. Cheplygina, L. Sørensen, D. M. J. Tax, M. Bruijne, and M. Loog, "Label Stability in Multiple Instance Learning," in *MICCAI*, 2015.

[23] V. Cheplygina and D. M. J. Tax, "Characterizing multiple instance datasets," in *SIMBAD*, 2015.

[24] F. Li and C. Sminchisescu, "Convex Multiple-Instance Learning by Estimating Likelihood Ratio," *NIPS*, 2010.

[25] Y. Han, Q. Tao, and J. Wang, "Avoiding False Positive in Multi-Instance Learning," in *NIPS*, 2010.

[26] S. Yan, X. Zhu, G. Liu, and J. Wu, "Sparse multiple instance learning as document classification," *Multimedia Tools and Appl.*, pp. 1–18, 2016.

[27] R. C. Bunescu and R. J. Mooney, "Multiple Instance Learning for Sparse Positive Bags," in *ICML*, 2007.

[28] Y. Li, D. M. Tax, R. P. Duin, and M. Loog, "Multiple-Instance Learning as a Classifier Combining Problem," *Pattern Recognit.*, vol. 46, no. 3, pp. 865–874, Mar. 2013.

[29] O. Maron and T. Lozano-Pérez, "A Framework for Multiple-Instance Learning," in *NIPS*, 1998.

[30] M.-A. Carbonneau, E. Granger, A. J. Raymond, and G. Gagnon, "Robust multiple-instance learning ensembles using random subspace instance selection," *Pattern Recognit.*, vol. 58, pp. 83–99, 2016.

[31] Y. Xiao, B. Liu, and Z. Hao, "A Sphere-Description-Based Approach For Multiple-Instance Learning." *IEEE Trans. Pattern Anal. Mach. Intell.*, 2016.

[32] M.-A. Carbonneau, E. Granger, and G. Gagnon, "Decision Threshold Adjustment Strategies for Increased Accuracy in Multiple Instance Learning," in *IPTA*, 2016.

[33] Q. Zhang and S. A. Goldman, "EM-DD : An Improved Multiple-Instance Learning Technique," in *NIPS*, 2001.

[34] Z.-H. Zhou and M.-L. Zhang, "Solving Multi-instance Problems with Classifier Ensemble Based on Constructive Clustering," *Knowl. Inf. Syst.*, vol. 11, no. 2, pp. 155–170, 2007.

[35] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, and S. Vluymans, *Multiple Instance Multiple Label Learning*. Springer, 2016, pp. 209–230.

[36] D. R. Dooly, Q. Zhang, S. A. Goldman, and R. A. Amar, "Multiple Instance Learning of Real Valued Data," *J. Mach. Learn. Res.*, vol. 3, pp. 651–678, Mar 2003.

[37] S. Ray and D. Page, "Multiple Instance Regression," in *ICML*, 2001.

[38] Z. Wang, V. Radosavljevic, B. Han, Z. Obradovic, and S. Vucetic, "Aerosol Optical Depth Prediction from Satellite Observations by Multiple Instance Regression," in *SDM*, 2008.

[39] K. L. Wagstaff and T. Lane, "Salience Assignment for Multiple-Instance Regression," in *ICML*, 2007.

[40] N. Pappas and A. Popescu-Belis, "Explaining the Stars: Weighted Multiple-Instance Learning for Aspect-Based Sentiment Analysis," in *EMNLP*, 2014.

[41] Y. EL-Manzalawy, D. Dobbs, and V. Honavar, "Predicting MHC-II Binding Affinity Using Multiple Instance Regression," *IEEE/ACM Trans. Comput. Biol. Bioinf*, vol. 8, no. 4, pp. 1067–1079, Jul 2011.

[42] C. Bergeron, G. Moore, J. Zaretzki, C. M. Breneman, and K. P. Bennett, "Fast bundle algorithm for multiple-instance learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 34, no. 6, pp. 1068–1079, Jun 2012.

[43] Y. Hu, M. Li, and N. Yu, "Multiple-instance ranking: Learning to rank images for image retrieval," in *CVPR*, 2008.

[44] M.-L. Zhang and Z.-H. Zhou, "Multi-instance clustering with applications to multi-instance prediction," *Applied Intelligence*, vol. 31, no. 1, pp. 47–68, 2009.

[45] D. Zhang, F. Wang, L. Si, and T. Li, "Maximum Margin Multiple Instance Clustering With Applications to Image and Text Clustering," *IEEE Trans. on Neural Networks*, vol. 22, no. 5, pp. 739–751, May 2011.

[46] F. Herrera, S. Ventura, R. Bello, C. Cornelis, A. Zafra, D. Sánchez-Tarragó, and S. Vluymans, *Multiple Instance Learning: Foundation and Algorithms*. Springer, 2016.

[47] S. Sabato and N. Tishby, "Multi-instance Learning with Any Hypothesis Class," *J. Mach. Learn. Res.*, vol. 13, no. 1, pp. 2999–3039, Oct 2012.

[48] M.-A. Carbonneau, E. Granger, and G. Gagnon, "Witness Identification in Multiple Instance Learning Using Random Subspaces," in *ICPR*, 2016.

[49] X.-S. Wei and Z.-H. Zhou, "An empirical study on image bag generators for multi-instance learning," *Machine Learning*, pp. 1–44, 2016.

[50] E. Nowak, F. Jurie, and B. Triggs, "Sampling Strategies for Bag-of-Features Image Classification," in *ECCV*, 2006.

[51] H. Wang, M. M. Ullah, A. Klaser, I. Laptev, and C. Schmid, "Evaluation of Local Spatio-Temporal Features for Action Recognition," in *BMVC*, 2009.

[52] R. Venkatesan, P. Chandakkar, and B. Li, "Simpler Non-Parametric Methods Provide as Good or Better Results to Multiple-Instance Learning," in *ICCV*, 2015.

[53] M. Kandemir and F. A. Hamprecht, "Computer-aided diagnosis from weak supervision: a benchmarking study." *Computerized Medical Imaging and Graphics*, vol. 42, pp. 44–50, Jun 2015.

[54] B. Babenko, P. Dollár, Z. Tu, and S. Belongie, "Simultaneous Learning and Alignment: Multi-Instance and Multi-Pose Learning," in *ECCV*, 2008.

[55] W. J. Li and D. Y. Yeung, "MILD: Multiple-Instance Learning via Disambiguation," *IEEE Trans. on Knowledge and Data Engineering*, vol. 22, no. 1, pp. 76–89, Jan 2010.

[56] Y. Jia and C. Zhang, "Instance-level Semisupervised Multiple Instance Learning," in *AAAI*, 2008.

[57] C. Yang, M. Dong, and J. Hua, "Region-based Image Annotation using Asymmetrical Support Vector Machine-based Multiple-Instance Learning," in *CVPR*, 2006.

[58] Z.-H. Zhou and J.-M. Xu, "On the Relation Between Multi-instance Learning and Semi-supervised Learning," in *ICML*, 2007.

[59] A. Blum and A. Kalai, "A Note on Learning from Multiple-Instance Examples," *Mach. Learn.*, vol. 30, no. 1, pp. 23–29, 1998.

[60] J. Amores, "Vocabulary-Based Approaches for Multiple-Instance Data: A Comparative Study," in *ICPR*, 2010.

[61] G. Doran and S. Ray, "Learning Instance Concepts from Multiple-instance Data with Bags As Distributions," in *AAAI*, 2014.

[62] X. S. Wei, J. Wu, and Z. H. Zhou, "Scalable Multi-instance Learning," in *ICD*, 2014.

[63] T. Gärtner, P. A. Flach, A. Kowalczyk, and A. J. Smola, "Multi-Instance Kernels," in *ICML*, 2002.

[64] X. Xu and E. Frank, "Logistic Regression and Boosting for Labeled Bags of Instances," in *PAKDD*, 2004.

[65] P. Gehler and O. Chapelle, "Deterministic Annealing for Multiple-Instance Learning," in *AISTATS*, 2007.

[66] K. Ali and K. Saenko, "Confidence-Rated Multiple Instance Boosting for Object Detection," in *CVPR*, 2014.

[67] D. Zhang, Y. Liu, L. Si, J. Zhang, and R. D. Lawrence, "Multiple Instance Learning on Structured Data," in *NIPS*, 2011.

[68] F. Kang, R. Jin, and R. Sukthankar, "Correlated label propagation with application to multi-label learning," in *CVPR*, 2006.

[69] V. Cheplygina, D. M. Tax, and M. Loog, "Multiple instance learning with bag dissimilarities," *Pattern Recognit.*, vol. 48, no. 1, pp. 264–275, Jan. 2015.

[70] G. Csurka, C. R. Dance, L. Fan, J. Willamowski, and C. Bray, "Visual categorization with bags of keypoints," in *ECCV*, 2004.

[71] W. Ping, Y. Xu, J. Wang, and X.-S. Hua, "FAMER: Making Multi-Instance Learning Better and Faster," in *SDM*, 2011.

[72] H.-Y. Wang, Q. Yang, and H. Zha, "Adaptive P-posterior Mixture-model Kernels for Multiple Instance Learning," in *ICML*, 2008.

[73] G. J. Qi, X. S. Hua, Y. Rui, T. Mei, J. Tang, and H. J. Zhang, "Concurrent multiple instance learning for image categorization," in *CVPR*, 2007.

[74] Z.-J. Zha, X.-S. Hua, T. Mei, J. Wang, G.-J. Qi, and Z. Wang, "Joint multi-label multi-instance learning for image classification," in *CVPR*, 2008.

[75] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld, "Learning realistic human actions from movies," in *CVPR*, 2008.

[76] M. S. Ryoo and J. K. Aggarwal, "Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities," in *ICCV*, 2009.

[77] A. Mcgovern and D. Jensen, "Identifying Predictive Structures in Relational Data Using Multiple Instance Learning," in *ICML*, 2003.

[78] K. Grauman and T. Darrell, "The pyramid match kernel: discriminative classification with sets of image features," in *ICCV*, 2005.

[79] S. Lazebnik, C. Schmid, and J. Ponce, "Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories," in *CVPR*, 2006.

[80] D. M. J. Tax, E. Hendriks, M. F. Valstar, and M. Pantic, "The Detection of Concept Frames Using Clustering Multi-instance Learning," in *ICPR*, 2010.

[81] X. Guan, R. Raich, and W.-K. Wong, "Efficient Multi-Instance Learning for Activity Recognition from Time Series Data Using an Auto-Regressive Hidden Markov Model," in *ICML*, 2016.

[82] J. Warrell and P. H. S. Torr, "Multiple-Instance Learning with Structured Bag Models," in *EMMCVPR*, 2011.

[83] Z. Li, G.-H. Geng, J. Feng, J.-y. Peng, C. Wen, and J.-l. Liang, "Multiple instance learning based on positive instance selection and bag structure construction," *Pattern Recognit. Lett.*, vol. 40, pp. 19–26, 2014.

[84] J. Wang and J.-D. Zucker, "Solving the Multiple-Instance Problem: A Lazy Learning Approach," in *ICML*, 2000.

[85] Y. Chen and J. Z. Wang, "Image Categorization by Learning and Reasoning with Regions," *J. Mach. Learn. Res.*, vol. 5, pp. 913–939, Dec. 2004.

[86] Q. Wang, L. Si, and D. Zhang, "A Discriminative Data-Dependent Mixture-Model Approach for Multiple Instance Learning in Image Classification," in *ECCV*, 2012.

[87] G. Doran, "Multiple Instance Learning from Distributions," Ph.D. dissertation, Case Western Reserve University, 2015.

[88] D. M. Tax and R. P. Duin, "Learning Curves for the Analysis of Multiple Instance Classifiers," in *IAPR*, 2008.

[89] C. Zhang, X. Chen, M. Chen, S.-C. Chen, and M.-L. Shyu, "A Multiple Instance Learning Approach for Content Based Image Retrieval Using One-Class Support Vector Machine," in *ICME*, 2005.

[90] R.-S. Wu and W.-H. Chung, "Ensemble one-class support vector machines for content-based image retrieval," *Expert Syst. Appl.*, vol. 36, no. 3, pp. 4451–4459, 2009.

[91] Z. Wang, Z. Zhao, and C. Zhang, "Learning with only multiple instance positive bags," in *IJCNN*, 2016.

[92] W. Li and N. Vasconcelos, "Multiple instance learning for soft bags via top instances," in *CVPR*, Jun 2015.

[93] Y. Rubner, C. Tomasi, and L. J. Guibas, "The Earth Mover's Distance As a Metric for Image Retrieval," *Int. J. Comput. Vision*, vol. 40, no. 2, pp. 99–121, Nov 2000.

[94] A. Erdem and E. Erdem, "Multiple-Instance Learning with Instance Selection via Dominant Sets," in *SIMBAD*, 2011.

[95] Z. Fu, A. Robles-Kelly, and J. Zhou, "MILIS: Multiple Instance Learning with Instance Selection," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 5, pp. 958–977, May 2011.

[96] S. Bandyopadhyay, D. Ghosh, R. Mitra, and Z. Zhao, "MBSTAR: multiple instance learning for predicting specific functional binding sites in microRNA targets," *Sci. Rep.*, vol. 5, p. 8004, 2015.

[97] D. Palachanis, "Using the Multiple Instance Learning framework to address differential regulation," Master, Delft University of Technology, 2014.

[98] R. Eksi, H.-D. Li, R. Menon, Y. Wen, G. S. Omenn, M. Kretzler, and Y. Guan, "Systematically differentiating functions for alternatively spliced isoforms through integrating RNA-seq data." *PLoS Comput. Biol.*, vol. 9, no. 11, Jan 2013.

[99] S. Vijayanarasimhan and K. Grauman, "Keywords to Visual Categories: Multiple-Instance Learning For Weakly Supervised Object Categorization," in *CVPR*, Jun. 2008.

[100] O. Maron and A. L. Ratan, "Multiple-Instance Learning for Natural Scene Classification," in *ICML*, 1998.

[101] C. Leistner, A. Saffari, and H. Bischof, "MIForests: Multiple-instance Learning with Randomized Trees," in *ECCV*, 2010.

[102] X. Song, L. Jiao, S. Yang, X. Zhang, and F. Shang, "Sparse Coding and Classifier Ensemble Based Multi-Instance Learning for Image Categorization," *Signal Process.*, vol. 93, no. 1, pp. 1–11, Jan. 2013.

[103] H. Xu, S. Venugopalan, V. Ramanishka, M. Rohrbach, and K. Saenko, "A Multi-scale Multiple Instance Video Description Network," *CoRR*, vol. abs/1505.0, 2016.

[104] A. Karpathy and L. Fei-Fei, "Deep Visual-Semantic Alignments for Generating Image Descriptions," in *CVPR*, 2015.

[105] 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, and G. Zweig, "From Captions to Visual Concepts and Back," in *CVPR*, 2015.

[106] J. Y. Zhu, J. Wu, Y. Xu, E. Chang, and Z. Tu, "Unsupervised Object Class Discovery via Saliency-Guided Multiple Class Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 37, no. 4, pp. 862–875, Apr 2015.

[107] H. O. Song, R. Girshick, S. Jegelka, J. Mairal, Z. Harchaoui, and T. Darrell, "On learning to localize objects with minimal supervision," in *ICML*, 2014.

[108] B. Babenko, M.-H. Yang, and S. Belongie, "Robust Object Tracking with Online Multiple Instance Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[109] M. Sapienza, F. Cuzzolin, and P. H. S. Torr, "Learning Discriminative Space–Time Action Parts from Weakly Labelled Videos," *Int. J. Comput. Vis.*, vol. 110, no. 1, pp. 30–47, 2014.

[110] A. Müller and S. Behnke, "Multi-instance Methods for Partially Supervised Image Segmentation," in *IAPR*, 2012, pp. 110–119.

[111] B. Hariharan, P. Arbeláez, R. Girshick, and J. Malik, "Simultaneous Detection and Segmentation," in *ECCV*, 2014.

[112] A. Vezhnevets and J. M. Buhmann, "Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning," in *CVPR*, 2010.

[113] K. T. Lai, F. X. Yu, M. S. Chen, and S. F. Chang, "Video Event Detection by Inferring Temporal Instance Labels," in *CVPR*, 2014.

[114] J. Wang, B. Li, W. Hu, and O. Wu, "Horror video scene recognition via Multiple-Instance learning," in *ICASSP*, 2011.

[115] B. Babenko, M.-H. Yang, and S. Belongie, "Robust Object Tracking with Online Multiple Instance Learning," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 8, pp. 1619–1632, Aug. 2011.

[116] K. Zhang and H. Song, "Real-time visual tracking via online weighted multiple instance learning," *Pattern Recognit.*, vol. 46, no. 1, pp. 397–411, 2013.

[117] H. Lu, Q. Zhou, D. Wang, and R. Xiang, "A co-training framework for visual tracking with multiple instance learning," in *FG'11*, 2011.

[118] J. Zhu, B. Wang, X. Yang, W. Zhang, and Z. Tu, "Action Recognition with Actons," in *ICCV*, 2013.

[119] Y. Xu *et al.*, "Weakly supervised histopathology cancer image segmentation and classification," *MedIA*, vol. 18, no. 3, pp. 591–604, 2014.

[120] G. Quellec *et al.*, "A multiple-instance learning framework for diabetic retinopathy screening," *MedIA*, vol. 16, no. 6, pp. 1228–1240, 2012.

[121] 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," *Medical image analysis*, vol. 18, no. 5, pp. 808–818, 2014.

[122] J. Melendez *et al.*, "A novel multiple-instance learning-based approach to computer-aided detection of tuberculosis on chest x-rays," *TMI*, vol. 31, no. 1, pp. 179–192, 2014.

[123] V. Cheplygina, L. Sørensen, D. M. J. Tax, J. H. Pedersen, M. Loog, and M. de Bruijne, "Classification of COPD with multiple instance learning," in *ICPR*, 2014.

[124] Z. S. Harris, "Distributional structure." *Word*, vol. 10, pp. 146–162, 1954.

[125] Y. Zhang, A. C. Surendran, J. C. Platt, and M. Narasimhan, "Learning from Multi-topic Web Documents for Contextual Advertisement," in *KDD*, 2008.

[126] D. Zhang, J. He, and R. Lawrence, "Mi2ls: Multi-instance learning from multiple informationsources," in *KDD*, 2013.

[127] B. Settles, M. Craven, and S. Ray, "Multiple-Instance Active Learning," in *NIPS*, 2008.

[128] Z. Jorgensen, Y. Zhou, and M. Inge, "A Multiple Instance Learning Strategy for Combating Good Word Attacks on Spam Filters," *J. Mach. Learn. Res.*, vol. 9, pp. 1115–1146, Jun. 2008.

[129] D. Kotzias, M. Denil, P. Blunsom, and N. de Freitas, "Deep Multi-Instance Transfer Learning," *CoRR*, vol. abs/1411.3, 2014.

[130] D. Kotzias, M. Denil, N. de Freitas, and P. Smyth, "From Group to Individual Labels Using Deep Features," in *KDD*, 2015.

[131] Z.-H. Zhou, K. Jiang, and M. Li, "Multi-Instance Learning Based Web Mining," *Appl. Intell.*, vol. 22, no. 2, pp. 135–147, Mar 2005.

[132] A. Zafra, S. Ventura, E. Herrera-Viedma, and C. Romero, "Multiple Instance Learning with Genetic Programming for Web Mining," *Computational and Ambient Intell.*, vol. 4507, pp. 919–927, 2007.

[133] M. I. Mandel and D. P. W. Ellis, "Multiple-instance learning for music information retrieval," 2008.

[134] R. F. Lyon, "Machine Hearing: An Emerging Field [Exploratory DSP]," *Signal Process. Mag. IEEE*, vol. 27, no. 5, pp. 131–139, Sep 2010.

[135] J. F. Ruiz-Muñoz, M. Orozco-Alzate, and G. Castellanos-Dominguez, "Multiple Instance Learning-based Birdsong Classification Using Unsupervised Recording Segmentation," in *IJCAI*, 2015.

[136] M.-A. Carbonneau, E. Granger, Y. Attabi, and G. Gagnon, "Feature learning from spectrograms for assessment of personality traits," *arXiv preprint arXiv:1610.01223*, 2016.

[137] A. Kumar and B. Raj, "Weakly Supervised Scalable Audio Content Analysis," *CoRR*, vol. abs/1606.0, 2016.

[138] M. Stikic, D. Larlus, S. Ebert, and B. Schiele, "Weakly Supervised Recognition of Daily Life Activities with Wearable Sensors," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 33, no. 12, pp. 2521–2537, Dec 2011.

[139] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado, "Machine Learning Methods for Predicting Failures in Hard Drives: A Multiple-Instance Application," *J. Mach. Learn. Res.*, vol. 6, pp. 783–816, Dec 2005.

[140] A. Manandhar, K. D. Morton, L. M. Collins, and P. A. Torrione, "Multiple instance learning for landmine detection using ground penetrating radar," in *SPIE*, 2012.

[141] A. Karem and H. Frigui, "A multiple instance learning approach for landmine detection using Ground Penetrating Radar," in *IGARSS*, 2011.

[142] P. Viola, J. C. Platt, and C. Zhang, "Multiple Instance Boosting for Object Detection," in *NIPS*, 2006.

[143] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *The Annals of Statistics*, vol. 29, no. 5, pp. 1189–1232, 2001.

[144] Z.-H. Zhou, X.-B. Xue, and Y. Jiang, "Locating Regions of Interest in CBIR with Multi-instance Learning Techniques," in *AUS-AI*, 2005.

[145] R. Rahmani, S. A. Goldman, H. Zhang, J. Krettek, and J. E. Fritts, "Localized Content Based Image Retrieval," in *SIGMM*, 2005.

[146] K. Lang, "Newsweeder: Learning to filter netnews," in *ICML*, 1995.

[147] P. Baldi, K. Cranmer, T. Faucett, P. Sadowski, and D. Whiteson, "Parameterized machine learning for high-energy physics," *arXiv preprint arXiv:1601.07913*, 2016.

[148] P. W. Frey and D. J. Slate, "Letter recognition using holland-style adaptive classifiers," *Mach. Learn.*, vol. 6, no. 2, pp. 161–182, Mar. 1991.

[149] M. Stone, "Cross-Validatory Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society. Series B (Methodological)*, vol. 36, no. 2, pp. 111–147, 1974.

[150] J. Demsar, "Statistical Comparisons of Classifiers over Multiple Data Sets," *J. Mach. Learn. Res.*, vol. 7, pp. 1–30, 2006.

[151] F. Li and C. Sminchisescu, "Convex Multiple-Instance Learning by Estimating Likelihood Ratio," *NIPS*, 2010.

[152] M. Everingham, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman, "The PASCAL visual object classes (VOC) challenge," *International journal of computer vision*, vol. 88, no. 2, pp. 303–338, 2010.

[153] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei, "Imagenet large scale visual recognition challenge," *Int. J. Comput. Vision*, vol. 115, no. 3, pp. 211–252, Dec. 2015.

[154] M. Kandemir, C. Zhang, and F. A. Hamprecht, "Empowering Multiple Instance Histopathology Cancer Diagnosis by Cell Graphs," *MICCAI*, 2014.

[155] P. Branco, L. Torgo, and R. P. Ribeiro, "A Survey of Predictive Modeling on Imbalanced Domains," *ACM Comput. Surv.*, vol. 49, no. 2, pp. 31:1—-31:50, Aug 2016.

[156] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *J. Artif. Int. Res.*, vol. 16, no. 1, pp. 321–357, Jun 2002.

[157] C. Seiffert, T. M. Khoshgoftaar, J. Van Hulse, and A. Napolitano, "RUSBoost: A Hybrid Approach to Alleviating Class Imbalance," *Syst. Man Cybern. Part A Syst. Humans, IEEE Trans.*, vol. 40, no. 1, pp. 185–197, Jan 2010.

[158] T. Imam, K. M. Ting, and J. Kamruzzaman, "z-SVM: An SVM for Improved Classification of Imbalanced Data," in *AJCAI*, 2006.

[159] K. Veropoulos, C. Campbell, and N. Cristianini, "Controlling the Sensitivity of Support Vector Machines," in *IJCAI*, 1999.

[160] Y. Bengio, A. Courville, and P. Vincent, "Representation learning: A review and new perspectives," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 35, no. 8, pp. 1798–1828, 2013.

[161] J. Mairal, F. Bach, J. Ponce, G. Sapiro, and A. Zisserman, "Discriminative learned dictionaries for local image analysis," in *CVPR*, 2008.

[162] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin, "Exploring Strategies for Training Deep Neural Networks," *J. Mach. Learn. Res.*, vol. 10, pp. 1–40, Jun 2009.

[163] A. Hauptmann, R. Yan, W. H. Lin, M. Christel, and H. Wactlar, "Can High-Level Concepts Fill the Semantic Gap in Video Retrieval? A Case Study With Broadcast News," *IEEE Trans. Multimed.*, vol. 9, no. 5, pp. 958–966, aug 2007.

[164] L.-j. Li, H. Su, L. Fei-fei, and E. P. Xing, "Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification," in *NIPS*, 2010.

[165] S. Sadanand and J. J. Corso, "Action bank: A high-level representation of activity in video," in *CVPR*, 2012.

[166] F. Ringeval, A. Sonderegger, J. Sauer, and D. Lalanne, "Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions," *FG'13*, 2013.

[167] M. Merler, B. Huang, L. Xie, G. Hua, and A. Natsev, "Semantic Model Vectors for Complex Video Event Recognition," *IEEE Trans. Multimed.*, vol. 14, no. 1, pp. 88–101, Feb 2012.

[168] K. Tang, B. Yao, L. Fei-Fei, and D. Koller, "Combining the Right Features for Complex Event Recognition," in *ICCV*, 2013.

[169] H. Daumé III, "Frustratingly easy domain adaptation," *arXiv preprint arXiv:0907.1815*, 2009.