On Classification with Bags, Groups and Sets

Veronika Cheplygina, David M.J. Tax, Marco Loog

June 3, 2014

Abstract

Many classification problems may be difficult to formulate in the traditional supervised setting, where both training and test samples are individual feature vectors. It may be the case that samples are better described by sets of feature vectors, that labels are only available for sets rather than individual samples, or, if individual labels are available, that these are not independent. To better deal with such problems, several extensions of supervised learning have been proposed, where either training and/or test objects are sets of feature vectors. However, such extensions are often proposed independently, disregarding important similarities and differences with other existing classification problems. In this work, we provide an overview of such learning scenarios, propose a taxonomy to illustrate the relationships between them, and discuss directions for further research in these areas.

1 Introduction

Some pattern recognition problems are difficult to formulate as regular supervised classification problems where (feature vector, label) pairs are available to train a classifier that, in turn, can predict labels for previously unseen feature vectors. In such cases, the objects are, in fact, not individual feature vectors, but sets or bags of feature vectors or instances. Although terminology differs in different fields, we use the terms bags and instances in the rest of this paper.

The first reason for using a bag of instances is that a single feature vector is often too restrictive to describe an object. For example, in drug activity prediction, an application originally addressed in multiple instance learning [10], we are interested in classifying molecules as having the desired effect (active) or not. However, a molecule is not just a list of its elements: most molecules can fold into different shapes or conformations, which may have an influence on the activity of that molecule. Furthermore, the number of stable shapes is different per molecule. A more logical choice is therefore to represent a molecule as a set of its conformations.

The second reason is that labels on the level of feature vectors are difficult, costly and/or time-consuming to obtain, but labels on a coarser level may be obtained more easily. For computer aided diagnosis applications, it can be very
expensive for a radiologist to label individual pixels or voxels in an image as healthy or diseased, while it is more feasible to label a full image, or large image regions. Such coarsely labeled scans or regions can then be used for predicting labels on bag level by labeling new patient scans or, on a finer grained instance level, by labeling individual pixels or voxels.

Another reason to consider the labeling of bags of instances, instead of single feature vectors, is that there may be structure in the labels of the instances. For example, in face verification, where a video of a person is available, considering all the video frames jointly can provide more confident predictions than labeling each of the frames individually and combining the decisions. Similarly, neighboring objects in images, videos, sounds, time series and so forth are typically very correlated, and thus should not be classified independently.

For all these reasons it may be advantageous to represent the objects by bags of instances. Different applications may ask for different representations in the training and the test phase. All possibilities shown in Fig. 1 occur: both training and test objects may be bags or instances. Traditional supervised learning is in the SI-SI scenario in the top left, where both training and test objects are instances. Predicting molecule activity is in the MI-MI scenario, where both training and test objects are bags. Image classification problems can be found in the MI-MI scenario (training on images, testing on images) as well as the MI-SI scenario (training on images, testing on pixels or patches). The face verification problem is best represented by the SI-MI scenario (training on a single face, testing on a set of faces).

The greater representational power, reduced need for labels and improved classification performance are attractive properties of learning domains where objects can be expressed as sets of feature vectors. This idea has been therefore applied in image recognition [32, 9], face verification [3, 37], tracking [4], computer aided diagnosis [40], molecule activity prediction [10] and document categorization [2], amongst others.

The success of a classifier in one (source) application may motivate other researchers to use the same method in a different (target) application. However, it is not necessarily the case that the assumptions of the source application still hold in the target application, which can lead to poor performances. On the other hand, it may also happen that the same type of problem occurs in two different applications, and that researchers in the respective fields approach the problem in different ways, without benefiting from each other’s findings. We therefore believe that understanding the relationships between such learning scenarios is of importance to researchers in different fields.

With this work, our goal is to provide an overview of learning scenarios in which bags of instances play a role at any of the stages in the learning or classification process. We have gathered the papers in this overview by searching for papers that proposed novel learning scenarios, as well as by combining synonyms of the word “set” with words such as “classification” or “learning”. This work is intended as a survey of learning problems, not of classifiers for a particular scenario, although we refer to existing surveys of this type whenever possible. Furthermore, we mainly focus on a single-label, binary classification scenario, as
Figure 1: Supervised learning (SI-SI) and extensions. In the MI-MI scenario (Section 4), both training and test objects are bags. In the MI-SI scenario (Section 5), the training objects are bags and test objects are instances, while in the MI-SI scenario (Section 6), the training objects are instances and the test objects are bags.
| Term                  | Description                                      |
|----------------------|--------------------------------------------------|
| Instance space       | $\mathcal{X}$, typically $\mathbb{R}^d$          |
| Instance             | $x \in \mathcal{X}$                             |
| Bag                  | $B \in 2^\mathcal{X}$                          |
| Discrete set of class labels | $\mathcal{C}$                          |
| Output space         | $\mathcal{Y}$                                   |
| Instance label       | $y \in \mathcal{Y}$                             |
| Bag label            | $Y \in \mathcal{Y}$                             |
| Instance classifier  | $f : \mathcal{X} \rightarrow \mathcal{Y}$       |
| Bag classifier       | $F : 2^\mathcal{X} \rightarrow \mathcal{Y}$     |

Figure 2: Overview of important notation

many problem formulations can be easily extended to a multi-label [45], multi-class setting.

This paper begins with an overview of notation and learning scenarios in Section 2. We explain the categories of learning scenarios in more detail in Sections 3 to 6. The paper concludes with a discussion in Section 7.

2 Notation and Overview

The basic terminology of bags and instances was already introduced in the previous section. This terminology is borrowed from the field of multiple instance learning (MIL). This is for two reasons: there are more papers on MIL than on other topics covered in this work, and the terms do not have other mathematical definitions that could be confusing.

Mathematically, an instance is represented by a single feature vector $x \in \mathcal{X}$, where $\mathcal{X} = \mathbb{R}^d$ is a $d$-dimensional space, while a bag is represented by a set of $n_i$ feature vectors $B_i = \{x_{ik}; k = 1...n_i\} \in 2^\mathcal{X}$. We denote the set of possible classes $\mathcal{C}$, and the set of possible labels $\mathcal{Y}$. In the case where each object has only one class label (and the focus of this overview), $\mathcal{Y} = \mathcal{C}$, in a multi-label scenario $\mathcal{Y} = 2^\mathcal{C}$.

When a test object is an instance, we are interested in finding an instance classifier $f : \mathcal{X} \rightarrow \mathcal{Y}$. When a test object is a bag, we are interested in finding a bag classifier $F : 2^\mathcal{X} \rightarrow \mathcal{Y}$.

There are several aspects that differ in the learning scenarios covered in this paper. We choose to categorize the learning domains by the following characteristics:

- **Type of training data.** The type of data that is provided to train a classifier: labeled instances, or labeled bags. In the case a bag is provided, usually the labels for the individual instances are not available.

- **Type of test data.** The type of data that is classified by the trained classifier: instances (pixels in an image) or bags (entire images). Often
Table 1: Summary of learning scenarios. The columns show the section where we explain the scenario, the type of training and test data, the assumptions on how the instance and bag labels are related are weak or strong, and the main references where the learning scenario is applied.

| Section | Train | Test | Assumptions | Main references |
|---------|-------|------|-------------|-----------------|
| 3. SI-SI | Instances | Instances | Weak | Supervised learning |
|         | Instances | Instances | Strong | Batch classification [48] |
|         |         |         |         | Collective classification [42, 37, 33] |
| 4. MI-MI | Bags | Bags | Weak | Sets of feature vectors [22, 18, 23] |
|         | Bags | Bags | Strong | Multiple instance learning [10, 31] |
| 5. MI-SI | Bags | Instances | Weak | - |
|         | Bags | Instances | Strong | Multiple instance learning [34, 46] |
|         |         |         |         | Aggregate output learning [36] |
|         |         |         |         | Learning with label proportions [38] |
| 6. SI-MI | Instances | Bags | Weak | - |
|         | Instances | Bags | Strong | Group-based classification [40] |
|         |         |         |         | Set classification [37] |
|         |         |         |         | Full-class set classification [25] |

this also determines how evaluation is done: on instance level or on bag level, but this is not always the case.

• **Assumptions on labels.** Different applications have different assumptions of how the labels of the instances and the labels of the bags are related: for example, an assumption could be that all pixels inside an image region have the same label. These assumptions play an important role in how the learning algorithms are developed.

These characteristics lead us to the categories in the leftmost column of Table 1. In the following sections, which are organized by the first two dimensions (types of training and test data), we will explain each category, the corresponding learning scenarios and assumptions, the equivalence of different terms in literature, or why the category is empty.

3. **SI-SI: Train on instances, test on instances**

The first category of Table 1 contains traditional **supervised learning** where both training and test objects are assumed to be independently generated from some underlying class distributions. We assume that the reader is familiar with supervised learning and keep this section short. For a general introduction, please refer to [17]. With the assumption of independently drawn train and test instances, the best possible approach is to classify each feature vector individually.
However, in some situations data is not independently generated, and we can make more assumptions about the correlations in the test data, and use these assumptions to improve the performance. The classical, rather general way to model dependencies between observations is through Markov random fields \[21\] and the related, currently more popular conditional random fields \[26\]. There are, however, also approaches that act direct on the bag level and do not need an explicit probabilistic model in order to be applied. Examples can be found in batch classification \[48\] and collective classification \[41, 7, 33\]. Batch classification is concerned with classifying parts (segments) of medical images, and the assumption is that neighboring image segments are correlated, and thus often have the same label. Collective classification is often applied to documents, such as identifying requests in emails \[7\] or assigning categories to webpages \[33\]. The underlying assumption is that correlation exists between emails in the same thread, or between webpages that link to each other, therefore, the labels of such neighboring instances are related.

4 MI-MI: Train on bags, test on bags

When both the training objects and test objects are bags, but no additional assumptions about the labels are present, the goal is classification of sets of feature vectors \[22\]. This concept is also used for classifying images as bags of pixels \[18\] or prediction of binding of proteins \[16\] by defining kernels on bags directly. Various kernel functions on computer vision applications are also explored in \[52\], where a kernel on sets of feature vectors is called ensemble similarity. Another application in classifying websites \[23\], where each website is represented as a set of feature vectors, and a set distance together with a nearest neighbor classifier is used.

Another domain where both training and test objects are bags, but stronger assumptions are made is called multiple instance learning (MIL) \[10, 31\]. In MIL, the objects are referred to as bags of instances. Originally, it was assumed that \(\mathcal{Y} = \{-1, +1\}\), and that the bag labels are determined by the (hidden) labels of their instances: a bag is positive if and only if there is at least one positive instance inside the bag; a bag is negative if and only if all of its instances are negative. Such reasoning has been applied to molecule or drug activity prediction \[10, 13\], image classification \[32, 9\], text categorization \[51, 2\], prediction of hard drive failures \[35\] and other settings. For example, in molecule activity prediction, a molecule is considered active if at least one of the conformations demonstrate the activity of interest.

There are two main approaches to achieve the goal of classifying bags. Due to the assumption on the relationship of the bag and instance labels, earlier methods focused on first finding an instance classifier \(f\), and then applying a combining rule \(g\) to the instance outputs. To use the traditional assumption in MIL, \(g\) is defined by the noisy OR function, as follows:
\[ F(B) = \begin{cases} +1 & \text{if } g(\{f(x_k)\}_{k=1}^n) > 1 \\ -1 & \text{otherwise} \end{cases} \]

where \( f(x_k) = p(y_k = 1|x_k) \).

More relaxed formulations of the traditional assumption have also been proposed \([50, 12]\). For instance, for a bag to be positive, it needs to have a specific fraction of positive instances. With such alternative assumptions, it is still possible to find \( f \) first and then apply an appropriate \( g \) to determine the labels of the test bags. By assuming that all instances contribute to the bag label independently, for instance, \( g \) can be replaced by the product of the instance posterior probabilities. Other, generalized rules for combining posterior probabilities for instances can be found in \([29]\).

Several MIL methods have moved away from the assumptions on the relationships of instance and bag labels \([12]\), and learn using assumptions on bags as a whole, therefore taking a detour to the “classification of sets of feature vectors” domain. In other words, such methods aim at finding \( F \) directly rather than through a combination of \( f \) and \( g \). A more general assumption is that bags of the same class are similar to one another, therefore such methods learn by defining distances \([49]\) or kernels \([14]\) between bags. Other approaches include converting bags to a single-instance representation using similarities \([9]\), dissimilarities \([43]\) or histograms in instance space \([47]\), the so-called bag-of-words representation. These methods then borrow techniques from supervised learning to classify bags.

More extensive surveys of MIL assumptions and classifiers can be found in \([53, 12, 4]\).

5 MI-SI: Train on bags, test on instances

This section is concerned with the case where training data is only labeled on bag-level, while instance-level labels are desired in the test phase. Note that this is not possible if no assumptions are made about the label transfer between instances and bags. This is why the “train on bags, test on instances, no assumptions” category in Table 1 is empty (denoted by \(-\)). By making additional assumptions, however, something can be said about the instance-level labels of the test data.

5.1 Learning from weakly labeled data

The standard assumption in multiple instance learning is one of the possibilities we can use to train the classifier using labeled bags, but provide instance-level labels for the test data. Although originally, the goal of MIL was to provide
labels for bags, a side-effect of some algorithms is that instance labels are predicted as well. The fact that only bag labels are required to produce instance labels means that less labels are required than in the usual supervised setting. In several fields, where such weakly labeled data can be (more) easily obtained, the focus has shifted to classifying instances rather than bags.

The goals of classifying instances and classifying bags are not identical, and therefore, in many cases, the optimal bag classifier will not be the optimal instance classifier and vice versa. An important reason in MIL for this is the traditional assumption. If bag classification is done by combining instance predictions, such as in [1], false negative instances are going to have less effect on the bag performance than false positive instances. Consider a positive bag where a positive instance is misclassified as negative: if the bag has any other positive instances, or a negative instance that has been falsely classified as positive, the bag label will still be correct. However, for a negative bag the label changes as soon as a single instance is misclassified. Similar observations have been made in [39] and in [44]. A more general reason why the optimal instance and bag classifiers do not necessarily correspond, is unequal bag sizes. Misclassifying a bag with a few instances will have less effect on the instance performance, than misclassifying a bag with many instances.

There are several examples where MIL is used for the purpose of predicting instance labels, rather than bag labels. For example, in image segmentation or annotation [46, 34] the goal is to label pixels as belonging to the background, or one of the objects portrayed in the image. This goal can be achieved with supervised learning, by providing fully annotated training images, where each pixel is labeled as background or foreground. However, providing such annotated images is costly – it is easier to approximately indicate where the foreground objects are present. MIL is therefore an interesting setting that can still offer image annotation, while only using coarsely labeled images as training input.

Weakly annotated data is also a benefit in tracking [3]. Instead of providing instances (patches) of the tracked object to the learner, bags of patches (with several inexact locations of the tracked object) can be used to improve performance. However, the goal of the tracking algorithm is to again label patches (instances), not bags. Other examples can be found in music information retrieval [30], where the goal is to predict tags (such as “rock”, “pop”) for songs, based on coarser-level tags for the albums or the artists. In [5], the goal is to classify fragments of bird songs, only by learning with bag-level labels for the whole recording.

Although these domains are not directly related to bags of instances, at this point it is important to mention that learning with such weakly annotated data has links to semi-supervised learning [8, 56] and learning with only positive and unlabeled data [11]. Both of these fields deal with weakly annotated data in a sense that some of it is annotated, and some of it is not. In multiple instance learning, all of the data (in the form of bags) is annotated, however, from the perspective of instances, these annotations are weak. More about the links between these fields can be found in [55, 28].
5.2 Learning with other label spaces

Another setting where only training objects are sets of feature vectors is learning about individuals from group statistics [24], aggregate output learning [36], and learning with label proportions [38], independent names for very related ideas. Here the bag labels are not just class labels, but proportions of class labels, \( Y = \{ y_i | i = 1, \ldots, |C|, y_i \in \mathbb{R}, \sum_i y_i = 1 \} \). For instance, a bag can be labeled as “75% positive, 25% negative”.

In [24], the application is image annotation images are provided with labels (“tiger”) along with a fraction of image patches that contain tigers. This is very similar to the image segmentation scenario described in Section 5, but the fraction of positives provides additional information to the classifier. Another possible application addresses privacy issues, such as when it is not desirable to provide the income (label) of a single person, but less problematic to provide the collective income (aggregated label) for a group of people.

The applications addressed in [38] are spam filtering and advertising. In spam filtering, proportions of spam/normal email are easier to estimate for a particular user than the exact labels of each email, however, the goal is to classify individual emails afterwards. In advertising, the proportions are related to customers that would buy a product only on discount, and customers that would buy a product in any circumstances. During an advertising campaign, estimating such proportions can help to predict which groups of customers should receive a discount coupon (and therefore buy the product).

This aggregated output / label proportions setting can be seen as multiple instance learning, where the fraction of positive instances (often called the witness rate) in the bags is already specified. An exact fraction is a stronger assumption than a non-zero fraction, therefore it should be easier to learn when the witness rate is given. For real-life MIL datasets, [24] assumes that a positive bag has exactly \( \frac{1}{n_i} \) positive instances. Other MIL methods take advantage of this by estimating a witness rate first, and then using this estimate to build instance classifiers [27, 15].

6 SI-MI: Train on instances, test on bags

This section is concerned with the scenario where instance-level labels are available for training, but bag-level instances are needed in the test phase. This may seem illogical, because it is already possible to build an instance classifier – why would bags be necessary?

If no assumptions are made about how the bags are generated, there is no added value in considering bags in the test phase, and the reason the category corresponding to SI-MI with few assumptions in Table 1 is empty. However, if additional information is available about the labels inside a bag, it may still be worthwhile to consider sets of feature vectors in the testing phase. Dependencies or constraints between the feature vectors inside a test bag can be exploited to improve the overall classification.
Consider the 1-dimensional binary classification problem in Fig. 3, and assume that given objects from each class, we have found the Bayes optimal instance classifier. The black circles are the test set, and their true labels are $-1$. If we were to classify these instances independently, the error would be equal to $\frac{1}{3}$, because the leftmost object will be misclassified to the positive class. However, with the added constraint that these instances belong to a group of objects from the same class, we could apply a combining rule on the instance outputs, classify the bag as negative, and propagate the label to all the individual instances, reducing the error to 0.

![Figure 3: 1-dimensional binary classification problem. The shaded instances are from the $y = -1$ class. When classifying these instances jointly, the added information that they are all of the same class helps to reduce the classification error.](image)

This situation occurs in group-based classification [40, 6] and set classification [37], independently proposed names for the setting where test objects are sets of feature vectors from the same class. Note that this setting can be easily transferred to the “train on bags, test on bags” category, because if the instances in one bag have the same label, it is straightforward to create bags from instances and vice versa.

In [40], a real-world application involves classifying groups of cells as healthy or anomalous, with the added information that all cells in a group share the same label. The classification of a test bag distance-based and is done by modifying the supervised versions of the nearest neighbor or the nearest mean classifiers. There are two broad approaches called the voting and the pooling scheme. In the voting scheme, each instance is labeled by a classifier $f$, such as the nearest neighbor, and the labels are combined with majority voting as $g$. In the pooling scheme, the distances are aggregated first, and only then converted to a label for the bag. The results show that the pooling scheme (i.e. a nearest neighbor classifier $F$ applied on the bag distances) produces better results.

Another example from computer aided diagnosis is in [19], where classification of cells is applied on two levels: patches (image segments, or instances) and cell slides (full images, or bags). Although some patch-level labels are available and a patch classifier can be built, considering the slide-level labels is still beneficial for performance.

One of the applications in [37] is face classification. When multiple images of the same person are available (such as from different cameras, or from different frames in a video), the fact that the faces share the same label can help identification. The most straightforward approach involves combining predictions
of each instance in a bag during the test phase. The best performing approach actually moves towards the MI-MI scenario, because in both the training and test phase, instance subsets are generated. Kernels are defined on these subsets, and the test bag is classified by combining the predictions of its subsets.

In other literature on face classification, this problem is often referred to as image-set classification [51, 20], although here it is possible that the training objects are bags as well (i.e. there are multiple training faces available for each person). Such problems are therefore often also solved with set distances or kernels.

The added information that all instances in a set share the same label is just one of the examples of a setting where the testing objects are bags. A reversed setting is full-class set classification [25]. It has an additional constraint that each of the instances has a unique label, i.e. it is known beforehand which instance labels will be present in the bag. This is an appropriate setting for registration purposes, where it is known which objects will be present, but not which detected object is which. Here the output of the bag classifier is not a single class label, but a super-label $Y \in \mathcal{I}$, where $\mathcal{I}$ is the set of permutations of the all class labels. Because $|\mathcal{I}| < 2^c$, [25] shows that a classifier $F$ that finds the instance labels jointly is guaranteed to perform better than concatenating the outputs of instance classifiers $f$.

Note that although instance labels are obtained, the labels we are interested in (the super-labels) are bag labels, and the performance is evaluated on bag level: either all instances were labeled correctly, or not. We illustrate this with the diagrams in Fig. 4.

7 Discussion

Many classification problems deal with objects that are represented as sets of feature vectors, or so-called bags of instances. This popularity is not surprising, as there are several motivating reasons for choosing such a representation at one or more stages of the classification process. Firstly, a set of feature vectors provides greater representational power than a single feature vector, and it might not be logical to express multiple entities (such as several face images of one person) as a single entity. Secondly, often labels may be available only on bag level, and too costly to obtain on instance level, therefore using the bag of instances representation as a form of weak supervision. Lastly, it can be advantageous to consider bags as a whole rather than as independent instances, because of relationships of the instances in a single bag.

This popularity is not without dangers: several different learning scenarios may be defined for the same problem, or several different problems may be incorrectly grouped under the same learning scenario. We proposed a taxonomy that illustrates the relationships of scenarios that deal with bags, groups or sets, and could help researchers relate novel problems to existing applications and research directions.

While the proposed taxonomy allows for heterogeneity in training and test
Figure 4: Variants of the SI-MI scenario. The training objects are instances and the test objects are bags, although the bag may be labeled by a set of instance labels (situation on the right). Note that in this case, the instance labels are decided jointly (as a bag super-label) by a bag classifier $F$, not by an instance classifier $f$. 
objects (i.e., where training objects are bags and test objects are instances and vice versa), it is limited because the training or test objects themselves are homogeneous. It would be interesting to investigate what happens in the case where in the training phase both labeled bags and labeled instances are available. As we already discussed in Section 5, the optimal bag classifier does not necessarily correspond with the optimal instance classifier. Therefore, deciding how to best use the available labels should depend on whether bags or instances are to be classified in the test phase. However, what if bags and instances can be expected in both the classification and test phases? A straightforward solution would be to train separate bag and instance classifiers, but when the bag and instance labels are related, an integrated classifier would perhaps be more suitable.

Another interesting observation is that the “hybrid” categories in the taxonomy (Section 6: SI-MI and Section 5: MI-SI) have attracted a lot of attention, and that the learning scenarios proposed here all need to rely on strong assumptions about the relationships of the instance and bag labels. One of the questions this raises is, what are the minimal assumptions needed to learn in such situations? Furthermore, the learning scenarios we reviewed do not exhaustively cover the types of constraints that could be present between the instance and bag labels. Learning scenarios that will be proposed in the future to fill some of these gaps, can now be easily placed in the context of the works described in this overview.

References

[1] J. Amores. Multiple instance classification: Review, taxonomy and comparative study. Artificial Intelligence, 201:81–105, 2013.

[2] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In Advances in Neural Information Processing Systems, volume 15, pages 561–568, 2002.

[3] O. Arandjelovic and R. Cipolla. Face set classification using maximally probable mutual modes. In International Conference on Pattern Recognition, volume 1, pages 511–514. IEEE, 2006.

[4] B. Babenko, M.-H. Yang, and S. Belongie. Visual tracking with online multiple instance learning. In Computer Vision and Pattern Recognition, pages 983–990. IEEE, 2009.

[5] F. Briggs, X. Fern, and R. Raich. Rank-loss support instance machines for MIML instance annotation. In International Conference on Knowledge Discovery and Data Mining, pages 534–542. ACM, 2012.

[6] S. D. Brossi and A. P. Bradley. A comparison of multiple instance and group based learning. In International Conference on Digital Image Computing Techniques and Applications, pages 1–8. IEEE, 2012.
[7] V. R. Carvalho and W. W. Cohen. On the collective classification of email speech acts. In \textit{International ACM SIGIR Conference on Research and Development in Information Retrieval}, pages 345–352. ACM, 2005.

[8] O. Chapelle, B. Schölkopf, A. Zien, et al. \textit{Semi-supervised learning}, volume 2. MIT press Cambridge, 2006.

[9] Y. Chen, J. Bi, and J. Z. Wang. Miles: Multiple-instance learning via embedded instance selection. \textit{IEEE Transactions on Pattern Analysis and Machine Intelligence}, 28(12):1931–1947, 2006.

[10] T. Dietterich, R. Lathrop, and T. Lozano-Pérez. Solving the multiple instance problem with axis-parallel rectangles. \textit{Artificial Intelligence}, 89(1-2):31–71, 1997.

[11] C. Elkan and K. Noto. Learning classifiers from only positive and unlabeled data. In \textit{International Conference on Knowledge Discovery and Data Mining}, pages 213–220. ACM, 2008.

[12] J. Foulds and E. Frank. A review of multi-instance learning assumptions. \textit{Knowledge Engineering Review}, 25(1):1, 2010.

[13] G. Fu, X. Nan, H. Liu, R. Patel, P. Daga, Y. Chen, D. Wilkins, and R. Doerksen. Implementation of multiple-instance learning in drug activity prediction. \textit{BMC Bioinformatics}, 13(Suppl 15):S3, 2012.

[14] T. Gärtner, P. Flach, A. Kowalczyk, and A. Smola. Multi-instance kernels. In \textit{International Conference on Machine Learning}, pages 179–186, 2002.

[15] P. V. Gehler and O. Chapelle. Deterministic annealing for multiple-instance learning. In \textit{International Conference on Artificial Intelligence and Statistics}, pages 123–130, 2007.

[16] B. Hoffmann, M. Zaslavskiy, J.-P. Vert, and V. Stoven. A new protein binding pocket similarity measure based on comparison of clouds of atoms in 3d: application to ligand prediction. \textit{BMC Bioinformatics}, 11(1):99, 2010.

[17] A. K. Jain, R. P. W. Duin, and J. Mao. Statistical pattern recognition: A review. \textit{IEEE Transactions on Pattern Analysis and Machine Intelligence}, 22(1):4–37, 2000.

[18] T. Jebara. Images as bags of pixels. In \textit{International Conference on Computer Vision}, pages 265–272, 2003.

[19] H. Kalkan, M. Nap, R. P. W. Duin, and M. Loog. Automated colorectal cancer diagnosis for whole-slice histopathology. In \textit{Medical Image Computing and Computer-Assisted Intervention}, pages 550–557. Springer, 2012.
[20] T.-K. Kim, J. Kittler, and R. Cipolla. Discriminative learning and recognition of image set classes using canonical correlations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 29(6):1005–1018, 2007.

[21] R. Kindermann, J. L. Snell, et al. *Markov random fields and their applications*, volume 1. American Mathematical Society Providence, RI, 1980.

[22] R. Kondor and T. Jebara. A kernel between sets of vectors. In *International Conference on Machine Learning*, volume 20, page 361, 2003.

[23] H.-P. Kriegel and M. Schubert. Classification of websites as sets of feature vectors. In *Databases and Applications*, pages 127–132, 2004.

[24] H. Kuck and N. de Freitas. Learning about individuals from group statistics. In *Uncertainty in Artificial Intelligence*, pages 332–339, 2005.

[25] L. I. Kuncheva. Full-class set classification using the hungarian algorithm. *International Journal of Machine Learning and Cybernetics*, 1(1):53–61, 2010.

[26] J. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289, 2001.

[27] Y. Li, D. M. J. Tax, R. P. W. Duin, and M. Loog. Multiple-instance learning as a classifier combining problem. *Pattern Recognition*, 2012.

[28] Y. Li, D. M. J. Tax, R. P. W. Duin, and M. Loog. The link between multiple-instance learning and learning from only positive and unlabelled examples. In *Multiple Classifier Systems*, volume 7872, pages 157–166. Springer Berlin Heidelberg, 2013.

[29] M. Loog and B. Van Ginneken. Static posterior probability fusion for signal detection: applications in the detection of interstitial diseases in chest radiographs. In *International Conference on Pattern Recognition*, volume 1, pages 644–647. IEEE, 2004.

[30] M. Mandel and D. Ellis. Multiple-instance learning for music information retrieval. In *International Conference on Music Information Retrieval*, pages 577–582, 2008.

[31] O. Maron and T. Lozano-Pérez. A framework for multiple-instance learning. In *Advances in Neural Information Processing Systems*, pages 570–576. Morgan Kaufmann Publishers, 1998.

[32] O. Maron and A. L. Ratan. Multiple-instance learning for natural scene classification. In *International Conference on Machine Learning*, volume 15, pages 341–349. Citeseer, 1998.
[33] L. K. McDowell, K. M. Gupta, and D. W. Aha. Cautious inference in collective classification. In *National Conference on Artificial Intelligence*, volume 22, page 596, 2007.

[34] A. Müller and S. Behnke. Multi-instance methods for partially supervised image segmentation. In *Partially Supervised Learning*, pages 110–119. Springer, 2012.

[35] J. F. Murray, G. F. Hughes, and K. Kreutz-Delgado. Machine learning methods for predicting failures in hard drives: A multiple-instance application. *Journal of Machine Learning Research*, 6(1):783, 2006.

[36] D. Musicant, J. Christensen, and J. Olson. Supervised learning by training on aggregate outputs. In *International Conference on Data Mining*, pages 252–261. IEEE, 2007.

[37] X. Ning and G. Karypis. The set classification problem and solution methods. In *International Conference on Data Mining Workshops*, pages 720–729. IEEE, 2008.

[38] N. Quadrianto, A. J. Smola, T. S. Caetano, and Q. V. Le. Estimating labels from label proportions. *Journal of Machine Learning Research*, 10:2349–2374, 2009.

[39] S. Ray and M. Craven. Supervised versus multiple instance learning: An empirical comparison. In *International Conference on Machine Learning*, pages 697–704. ACM, 2005.

[40] N. A. Samsudin and A. P. Bradley. Nearest neighbour group-based classification. *Pattern Recognition*, 43(10):3458–3467, 2010.

[41] P. Sen, G. Namata, M. Bilgic, and L. Getoor. Collective classification. In *Encyclopedia of Machine Learning*, pages 189–193. Springer, 2010.

[42] P. Sen, G. Namata, M. Bilgic, L. Getoor, B. Galligher, and T. Eliassi-Rad. Collective classification in network data. *AI magazine*, 29(3):93, 2008.

[43] D. M. J. Tax, M. Loog, R. P. W. Duin, V. Cheplygina, and W. J. Lee. Bag dissimilarities for multiple instance learning. In *Similarity-Based Pattern Recognition*, pages 222–234. Springer, 2011.

[44] V. Tragante do O, D. Fierens, and H. Blockeel. Instance-level accuracy versus bag-level accuracy in multi-instance learning. In *Benelux Conference on Artificial Intelligence*, page 8, 2011.

[45] G. Tsoumakas and I. Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining*, 3(3):1–13, 2007.

[46] A. Vezhnevets and J. M. Buhmann. Towards weakly supervised semantic segmentation by means of multiple instance and multitask learning. In *Computer Vision and Pattern Recognition*, pages 3249–3256. IEEE, 2010.
[47] S. Vijayanarasimhan and K. Grauman. Keywords to visual categories: Multiple-instance learning for weakly supervised object categorization. In Computer Vision and Pattern Recognition, pages 1–8. IEEE, 2008.

[48] V. Vural, G. Fung, B. Krishnapuram, J. Dy, and B. Rao. Batch classification with applications in computer aided diagnosis. In European Conference on Machine Learning, pages 449–460. Springer, 2006.

[49] J. Wang. Solving the multiple-instance problem: A lazy learning approach. In International Conference on Machine Learning, pages 1119–1125, 2000.

[50] N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problems. European Conference on Machine Learning, pages 468–479, 2003.

[51] L. Wolf and A. Shashua. Learning over sets using kernel principal angles. Journal of Machine Learning Research, 4:913–931, 2003.

[52] S. K. Zhou and R. Chellappa. From sample similarity to ensemble similarity: Probabilistic distance measures in reproducing kernel hilbert space. IEEE Transactions on Pattern Analysis and Machine Intelligence, 28(6):917–929, 2006.

[53] Z.-H. Zhou. Multi-instance learning: A survey. Technical report, Department of Computer Science and Technology, Nanjing University, 2004.

[54] Z. H. Zhou, Y. Y. Sun, and Y. F. Li. Multi-instance learning by treating instances as non-iid samples. In International Conference on Machine Learning, pages 1249–1256. ACM, 2009.

[55] Z.-H. Zhou and J.-M. Xu. On the relation between multi-instance learning and semi-supervised learning. In International conference on Machine learning, pages 1167–1174. ACM, 2007.

[56] X. Zhu. Semi-supervised learning literature survey. Technical Report 1530, Computer Sciences, University of Wisconsin-Madison, 2005.