REVIEW

National Science Review
5: 44–53, 2018
doi: 10.1093/nsr/nwx106
Advance access publication 25 August 2017

INFORMATION SCIENCE

Special Topic: Machine Learning

A brief introduction to weakly supervised learning

Zhi-Hua Zhou

National Key Laboratory for Novel Software Technology, Nanjing University, Nanjing 210023, China

E-mail: zhouzh@nju.edu.cn

Received 13 June 2017; **Revised** 24 July 2017; **Accepted** 10 August 2017

ABSTRACT

Supervised learning techniques construct predictive models by learning from a large number of training examples, where each training example has a *label* indicating its ground-truth output. Though current techniques have achieved great success, it is noteworthy that in many tasks it is difficult to get strong supervision information like fully ground-truth labels due to the high cost of the data-labeling process. Thus, it is desirable for machine-learning techniques to work with weak supervision. This article reviews some research progress of *weakly supervised learning*, focusing on three typical types of weak supervision: incomplete supervision, where only a subset of training data is given with labels; inexact supervision, where the training data are given with only coarse-grained labels; and inaccurate supervision, where the given labels are not always ground-truth.

Keywords: machine learning, weakly supervised learning, supervised learning

INTRODUCTION

Machine learning has achieved great success in various tasks, particularly in *supervised learning* tasks such as classification and regression. Typically, predictive models are learned from a training data set that contains a large amount of training examples, each corresponding to an event/object. A training example consists of two parts: a feature vector (or *instance*) describing the event/object, and a *label* indicating the ground-truth output. In classification, the label indicates the class to which the training example belongs; in regression, the label is a real-value response corresponding to the example. Most successful techniques, such as deep learning [1], require ground-truth labels to be given for a big training data set; in many tasks, however, it can be difficult to attain strong supervision information due to the high cost of the data-labeling process. Thus, it is desirable for machine-learning techniques to be able to work with weak supervision.

Typically, there are three types of weak supervision. The first is *incomplete* supervision, i.e. only a (usually small) subset of training data is given with labels while the other data remain unlabeled. Such a situation occurs in various tasks. For example, in image categorization the ground-truth labels are given by human annotators; it is easy to get a huge number of images from the Internet, whereas only a small subset of images can be annotated due to the human cost. The second type is *inexact* supervision, i.e. only coarse-grained labels are given. Consider the image categorization task again. It is desirable to have every object in the images annotated; however, usually we only have image-level labels rather than object-level labels. The third type is *inaccurate* supervision, i.e. the given labels are not always ground-truth. Such a situation occurs, e.g. when the image annotator is careless or weary, or some images are difficult to categorize.

Weakly supervised learning is an umbrella term covering a variety of studies that attempt to construct predictive models by learning with weak supervision. In this article, we will discuss some progress in this line of research, focusing on learning with incomplete, inexact and inaccurate supervision. We will treat these types of weak supervision separately, but it is worth mentioning that in real practice they often occur simultaneously. For simplicity, in this article we consider binary classification concerning two exchangeable classes Y and N. Formally, with strong supervision, the supervised learning task is to learn $f : \mathcal{X} \mapsto \mathcal{Y}$ from a training data set $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_m, y_m)\}$, where $\mathcal{X}$ is the feature space, $\mathcal{Y} = \{Y, N\}$, $\boldsymbol{x}_i \in \mathcal{X}$, and $y_i \in \mathcal{Y}$. We

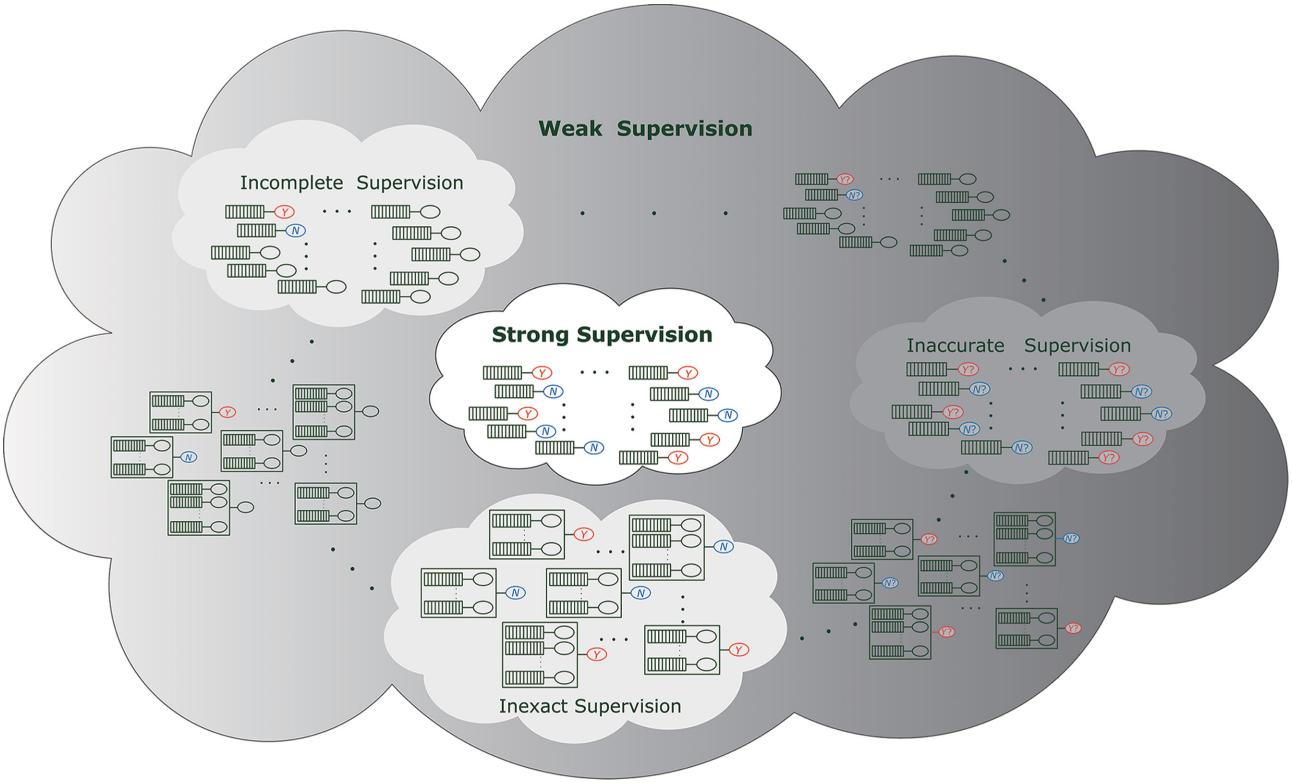


Figure 1. Illustration of three typical types of weak supervision. Bars denote feature vectors; red/blue marks labels; '?' implies that the label may be inaccurate. Intermediate subgraphs depict some situations with mixed types of weak supervision.

assume that $(\boldsymbol{x}_i, y_i)$ are generated according to an unknown identical and independent distribution $\mathcal{D}$; in other words, $(\boldsymbol{x}_i, y_i)$ are i.i.d. samples. Figure 1 provides an illustration of the three types of weak supervision that we will discuss in this article.

INCOMPLETE SUPERVISION

Incomplete supervision concerns the situation in which we are given a small amount of labeled data, which is insufficient to train a good learner, while abundant unlabeled data are available. Formally, the task is to learn $f : \mathcal{X} \mapsto \mathcal{Y}$ from a training data set $D = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_l, y_l), \boldsymbol{x}_{l+1}, \ldots, \boldsymbol{x}_m\}$, where there are l number of labeled training examples (i.e. those given with y_i) and $u = m - l$ number of unlabeled instances; the other conditions are the same as in supervised learning with strong supervision, as defined at the end of the introduction. For the convenience of discussion, we also call the l labeled examples 'labeled data' and the u unlabeled instances 'unlabeled data'.

There are two major techniques for this purpose, i.e. *active learning* [2] and *semi-supervised learning* [3–5].

Active learning assumes that there is an 'oracle', such as a human expert, that can be queried to get ground-truth labels for selected unlabeled instances. In contrast, semi-supervised learning attempts to automatically exploit unlabeled data in addition to labeled data to improve learning performance, where no human intervention is assumed. There is a special kind of semi-supervised learning called *transductive learning*; the main difference between this and (pure) semi-supervised learning lies in their different assumptions about test data, i.e. data to be predicted by the trained model. Transductive learning holds a 'closed-world' assumption, i.e. the test data are given in advance and the goal is to optimize performance on the test data; in other words, the unlabeled data are exactly test data. Pure semi-supervised learning holds an 'open-world' assumption, i.e. the test data are unknown and the unlabeled data are not necessarily test data. Figure 2 intuitively shows the difference between active learning, (pure) semi-supervised learning and transductive learning.

With human intervention

Active learning [2] assumes that the ground-truth labels of unlabeled instances can be queried from an oracle. For simplicity, assume that the labeling cost depends only on the number of queries. Thus, the goal of active learning is to minimize the number of queries such that the labeling cost for training a good model can be minimized.

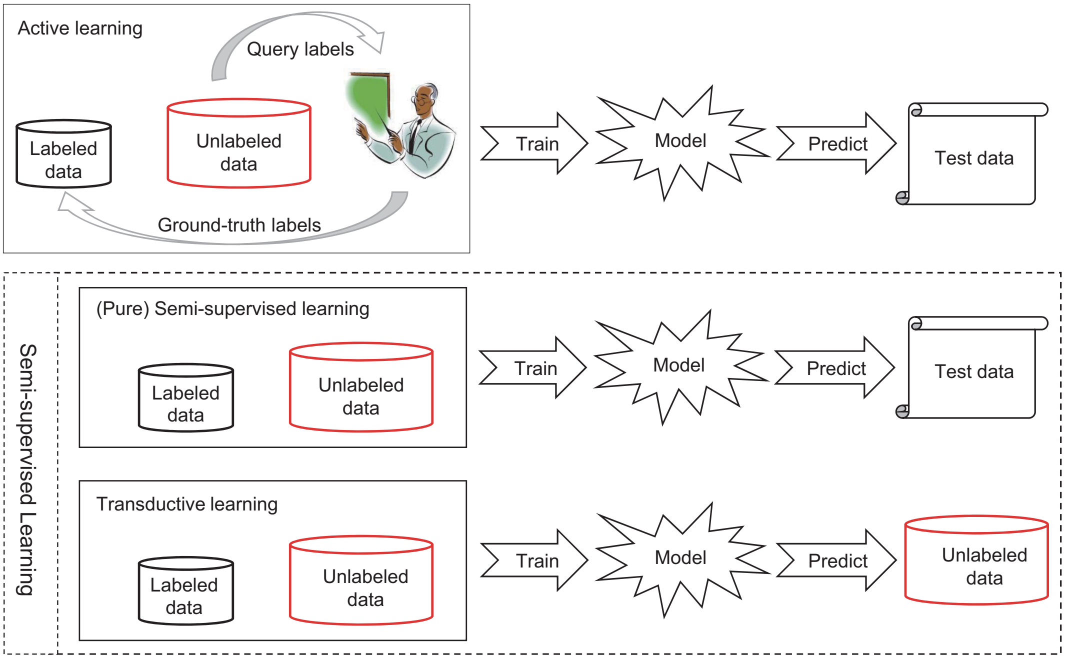


Figure 2. Active learning, (pure) semi-supervised learning and transductive learning.

Given a small set of labeled data and abundant unlabeled data, active learning attempts to select the most valuable unlabeled instance to query. There are two widely used selection criteria, i.e. *informativeness* and *representativeness* [6]. Informativeness measures how well an unlabeled instance helps reduce the uncertainty of a statistical model, whereas representativeness measures how well an instance helps represent the structure of input patterns.

Uncertainty sampling and *query-by-committee* are representative approaches based on informativeness. The former trains a single learner and then queries the unlabeled instance on which the learner has the least confidence [7]. The latter generates multiple learners and then queries the unlabeled instance on which the learners disagree the most [8,9]. Approaches based on representativeness generally aim to exploit the cluster structure of unlabeled data, usually by a clustering method [10,11].

The main weakness of informativeness-based approaches lies in the fact that they rely seriously on labeled data for constructing the initial model to select the query instance, and the performance is often unstable when there are only a few labeled examples available. The main weakness of representativeness-based approaches lies in the fact that the performance heavily depends on the clustering results dominated by unlabeled data, especially when there are only a few labeled examples. Thus, several recent active learning approaches try to leverage informativeness and representativeness [6,12].

There are many theoretical studies about active learning. For example, it has been proven that for *realizable* cases (where there exists a hypothesis perfectly separating the data in the hypothesis class), exponential improvement in sample complexity can be obtained by active learning [13,14]. For *non-realizable* cases (where the data cannot be perfectly separated by any hypothesis in the hypothesis class because of noise) it has been shown that, without assumptions about noise models, the lower bound of active learning matches the upper bound of passive learning [15]; in other words, active learning does not offer much help. By assuming a Tsybakov noise model, it has been proven that exponential improvement can be obtained for bounded noise [16,17]; if some special data characteristics, such as multi-view structure, can be exploited, exponential improvement can even be achieved for unbounded noise [18]. In other words, even for difficult cases, active learning can still be helpful with delicate design.

Without human intervention

Semi-supervised learning [3–5] attempts to exploit unlabeled data without querying human experts. One might be curious about why data without labels can help construct predictive models. For a simple explanation [19], assume that the data come from a Gaussian mixture model with n mixture

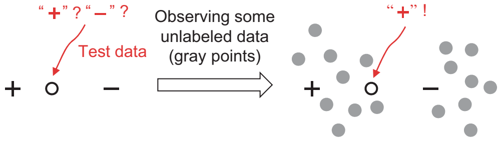


Figure 3. Illustration of the usefulness of unlabeled data.

components, i.e.

$$f(\boldsymbol{x}|\Theta)=\sum_{j=1}^{n}\alpha_j f(\boldsymbol{x}|\theta_j), \tag{1}$$

where α_i is the mixture coefficient, $\sum_{i=1}^{n}\alpha_i=1$, and $\Theta=\{\theta_i\}$ are the model parameters. In this case, label y_i can be considered as a random variable whose distribution $P(y_i|\boldsymbol{x}_i, g_i)$ is determined by the mixture component g_i and the feature vector $\boldsymbol{x}_i$. According to the maximum *a posteriori* criterion, we have the model

$$\begin{aligned} h(\boldsymbol{x}) = \underset{c\in\{Y,N\}}{\arg\max} \sum_{j=1}^{n} P\left(y_i=c|g_i=j, \boldsymbol{x}_i\right) \\ \times P\left(g_i=j|\boldsymbol{x}_i\right), \end{aligned} \tag{2}$$

where

$$P\left(g_i=j|\boldsymbol{x}_i\right)=\frac{\alpha_j f(\boldsymbol{x}_i|\theta_j)}{\sum_{k=1}^{n}\alpha_k f(\boldsymbol{x}_i|\theta_k)}. \tag{3}$$

The objective is accomplished by estimating the terms $P\left(y_i=c|g_i=j, \boldsymbol{x}_i\right)$ and $P\left(g_i=j|\boldsymbol{x}_i\right)$ from the training data. It is evident that only the first term requires label information. Thus, unlabeled data can be used to help improve the estimate of the second term, and hence improve the performance of the learned model.

Figure 3 provides an intuitive explanation. If we have to make a prediction based on the only positive and negative points, what we can do is just a random guess because the test data point lies exactly in the middle between the two labeled data points; if we are allowed to observe some unlabeled data points like the gray ones in the figure, we can predict the test data point as positive with high confidence. Here, although the unlabeled data points do not explicitly have label information, they implicitly convey some information about data distribution that can be helpful for predictive modeling.

Actually, in semi-supervised learning there are two basic assumptions, i.e. the *cluster assumption* and the *manifold assumption*; both are about data distribution. The former assumes that data have inherent cluster structure, and thus, instances falling into the same cluster have the same class label. The latter assumes that data lie on a manifold, and thus, nearby instances have similar predictions. The essence of both assumptions lies in the belief that similar data points should have similar outputs, whereas unlabeled data can be helpful to disclose which data points are similar.

There are four major categories of semi-supervised learning approaches, i.e. generative methods, graph-based methods, low-density separation methods and disagreement-based methods.

Generative methods [19,20] assume that both labeled and unlabeled data are generated from the same inherent model. Thus, labels of unlabeled instances can be treated as missing values of model parameters, and estimated by approaches such as the EM (expectation–maximization) algorithm [21]. These methods differ by fitting data using different generative models. To get good performance, one usually needs domain knowledge to determine an adequate generative model. There are also attempts to combine the advantages of the generative and discriminative approaches [22].

Graph-based methods [23–25] construct a graph, where the nodes correspond to training instances and the edges to the relation (usually some kind of similarity or distance) between instances, and then propagate label information on the graph according to some criteria; e.g. labels can be propagated inside different subgraphs separated by a minimum cut [23]. Apparently, the performance will heavily depend on how the graph is constructed [26–28]. Note that for m data points such approaches generally require about $O(m^2)$ storage and almost $O(m^3)$ computational complexity. Thus, they suffer seriously from scalability; in addition, they are inherently transductive, because it is difficult to accommodate new instances without graph reconstruction.

Low-density separation methods enforce the classification boundary to go across the less-dense regions in input space. The most famous representatives are S3VMs (semi-supervised support vector machines) [29–31]. Figure 4 demonstrates the difference between conventional supervised SVMs and S3VMs. It is evident that S3VMs try to identify a classification boundary that goes across the less-dense region while keeping the labeled data correctly classified. Such a goal can be accomplished by trying different label assignments for unlabeled data points in different ways, leading to complicated optimization problems. Thus, much effort in this line of research is devoted to efficient approaches for the optimization.

Disagreement-based methods [5,32,33] generate multiple learners and let them collaborate to exploit unlabeled data, where the disagreement among the learners is crucial to allow the learning process to continue. The most famous representative, co-training [32], works by training two learners

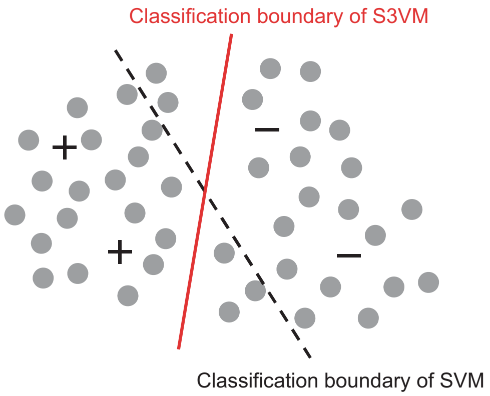

Classification boundary of S3VM
+
+
−
−
Classification boundary of SVM

Figure 4. Illustration of different classification boundaries of SVM which considers only labeled data ("+/−" points), and S3VM which considers labeled and unlabeled data (gray points).

from two different feature sets (or two *views*). In each iteration, each learner picks its most confidently predicted unlabeled instances, and assigns its predictions as pseudo-labels for the training of its peer learner. Such approaches can be further enhanced by combining the learners as an ensemble [34,35]. Note that disagreement-based methods offer a natural way to combine semi-supervised learning with active learning: in addition to letting the learners teach each other, some unlabeled instances, on which the learners are all unconfident or highly confident but contradictory, can be selected to query.

It is worth mentioning that although the learning performance is expected to be improved by exploiting unlabeled data, in some cases the performance may become worse after semi-supervised learning. This issue has been raised and studied for many years [36]; however, only recently has some solid progress been reported [37]. We now understand that the exploitation of unlabeled data naturally leads to more than one model option, and inadequate choice may lead to poor performance. The fundamental strategy to make semi-supervised learning 'safer' is to optimize the worst-case performance among the options, possibly by incorporating ensemble mechanisms [35].

There are abundant theoretical studies about semi-supervised learning [4], some even earlier than the coinage of the term 'semi-supervised learning' [38]. In particular, a thorough study about disagreement-based methods has recently been presented [39].

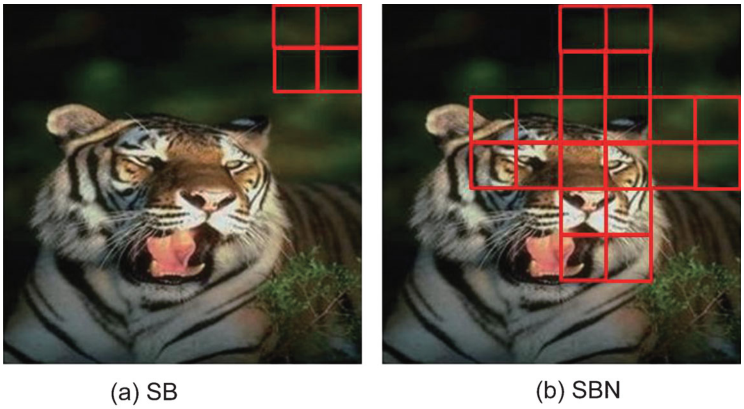
INEXACT SUPERVISION

Inexact supervision concerns the situation in which some supervision information is given, but not as exact as desired. A typical scenario is when only coarse-grained label information is available. For example, in the problem of drug-activity prediction [40], the goal is to build a model to predict whether a new molecule is qualified to make a special drug or not, by learning from a set of known molecules. One molecule can have many low-energy shapes, and whether the molecule can be used to make the drug depends on whether the molecule has some special shapes. Even for known molecules, however, human experts only know whether the molecules are qualified or not, instead of knowing which special shapes are decisive.

Formally, the task is to learn $f : \mathcal{X} \mapsto \mathcal{Y}$ from a training data set $D = \{(X_1, y_1), \dots, (X_m, y_m)\}$, where $X_i = \{\boldsymbol{x}_{i1}, \dots, \boldsymbol{x}_{i,m_i}\} \subseteq \mathcal{X}$ is called a *bag*, $\boldsymbol{x}_{ij} \in \mathcal{X}$ $(j \in \{1, \dots, m_i\})$ is an instance, m_i is the number of instances in X_i, and $y_i \in \mathcal{Y} = \{Y, N\}$. X_i is a *positive* bag, i.e. $y_i = Y$, if there exists $\boldsymbol{x}_{ip}$ that is positive, while $p \in \{1, \dots, m_i\}$ is unknown. The goal is to predict labels for unseen bags. This is called *multi-instance learning* [40,41].

Many effective algorithms have been developed for multi-instance learning. Actually, almost all supervised learning algorithms have their multi-instance peers. Most algorithms attempt to adapt single-instance supervised learning algorithms to the multi-instance representation, mainly by shifting their focus from the discrimination on instances to the discrimination on bags [42]; some other algorithms attempt to adapt the multi-instance representation to single-instance algorithms through representation transformation [43,44]. There is also a categorization [45] that groups the algorithms into an instance-space paradigm, where the instance-level responses are aggregated; a bag-space paradigm, where the bags are treated as a whole; and an embedded-space paradigm, where learning is performed in an embedded feature space. Note that the instances are usually regarded as i.i.d. samples; however, [46] indicates that the instances in multi-instance learning should not be assumed to be independent although the bags can be treated as i.i.d. samples, and based on this insight, some effective algorithms have been developed [47].

Multi-instance learning has been successfully applied to various tasks, such as image categorization/retrieval/annotation [48–50], text categorization [51,52], spam detection [53], medical diagnosis [54], face/object detection [55,56], object class discovery [57], object tracking [58], etc. In these tasks it is natural to regard a real object (such as an image or text document) as a bag; however, in contrast to drug-activity prediction where there are natural formations of instances in a bag (i.e. shapes of



(a) SB

(b) SBN

Figure 5. Image bag generators. Suppose each image is of size 8 × 8 and each blob is of size 2 × 2. Single Blob (SB) will generate 16 instances for the image, by regarding each patch consisting of four blobs as one instance, and sliding without overlap. Single Blob with Neighbors (SBN) will generate nine instances for the image, by regarding the patch consisting of 20 blobs as one instance, and sliding with overlap.

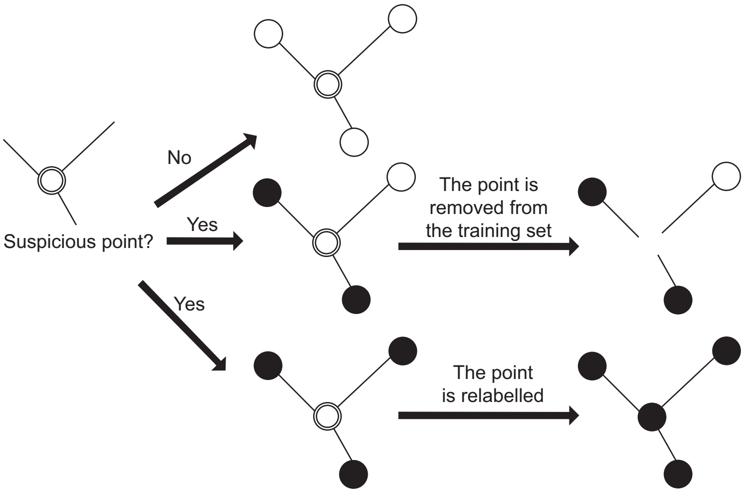


Figure 6. Identifying and removing/relabeling suspicious points.

a molecule), the instances need to be generated for each bag. A bag generator specifies how instances are generated to constitute a bag. Typically, many small patches can be extracted from an image as its instances, whereas sections/paragraphs or even sentences can be used as instances for text documents. Although bag generators have a significant influence on learning performance, only recently has an extensive study about image bag generators been reported [59]; this study discloses that some simple dense-sampling bag generators perform better than complicated ones. Figure 5 shows two simple yet effective image bag generators.

The original goal of multi-instance learning is to predict labels for unseen bags; however, there are studies trying to identify the *key instance* that enables a positive bag to be positive [31,60]. This is quite helpful in tasks such as locating regions of interest in images without fine-grained labeled training data. It is noteworthy that standard multi-instance learning [40] assumes that each positive bag must contain a key instance, whereas there are studies that assume that there is no key instance and every instance contributes to the bag label [61,62], or even assume that there are multiple concepts and a bag is positive only when the bag contains instances satisfying every concept [63]. More variants can be found in [41].

Early theoretical results [64–66] show that multi-instance learning is hard for *heterogeneous* cases in which each instance in the bag is classified by a different rule, while it is learnable for *homogeneous* cases in which all instances are classified by the same rule. Fortunately, almost all practical multi-instance tasks belong to the homogeneous class. These analyses assume that instances in the bags are independent. Analysis without assuming instance independence is more challenging and appears much later, disclosing that in the homogeneous class there are at least some cases learnable for arbitrary distribution over bags [67]. Nevertheless, in contrast to the flourishing studies in algorithms and applications, theoretical results on multi-instance learning are very rare because the analysis is quite hard.

INACCURATE SUPERVISION

Inaccurate supervision concerns the situation in which the supervision information is not always ground-truth; in other words, some label information may suffer from errors. The formulation is almost the same as what was shown at the end of the introduction, except that the y_i in the training data set may be incorrect.

A typical scenario is learning with label noise [68]. There are many theoretical studies [69–71], among which most assume random classification noise, i.e. labels are subject to random noise. In practice, a basic idea is to identify the potentially mislabeled examples [72], and then try to make some correction. For example, a *data-editing* approach [73] constructs a relative neighborhood graph where each node corresponds to a training example, and an edge connecting two nodes with different labels is called a *cut edge*. Then, a cut-edge weight statistic is measured, with the intuition that an instance is suspicious if it is associated with many cut edges. The suspicious instances can be either removed or relabeled, as illustrated in Fig. 6. It is worth mentioning that such approaches generally rely on consulting neighborhood information, and thus, they are less reliable in high-dimensional feature space because the identification of neighborhoods is usually less reliable when data are sparse.

An interesting recent scenario of inaccurate supervision occurs with *crowdsourcing* [74], a popular paradigm to outsource work to individuals. For machine learning, crowdsourcing is commonly used as a cost-saving way to collect labels for training data. Specifically, unlabeled instances are outsourced to a large group of workers to label. A famous crowdsourcing system, Amazon Mechanical Turk (AMT), is a market where the user can submit a task, such as annotating images of trees versus non-trees, to be completed by workers in exchange for small monetary payments. The workers usually come from a large society and each of them is presented with multiple tasks. They are usually independent and relatively inexpensive, and will provide labels based on their own judgments. Among the workers, some may be more reliable than others; however, the user usually does not know this in advance because the identities of workers are protected. There may exist 'spammers' who assign almost random labels to the tasks (e.g. robots pretend to be a human for the monetary payment), or 'adversaries' who give incorrect answers deliberately. Moreover, some tasks may be too difficult for many workers. Thus, it is nontrivial to maintain learning performance using the inaccurate supervision information returned by the crowd.

Many studies attempt to infer ground-truth labels from the crowd. The majority voting strategy, with theoretical support in ensemble methods [35], is widely used in practice with good performance [75,76], and thus often used as a baseline. It is expected that if worker quality and task difficulty can be modeled, better performance can be achieved, typically by weighting different workers for different tasks. For this purpose, some approaches try to construct probabilistic models and then adopt the EM algorithm for the estimation [77,78]. The minimax entropy principle has also been used [35]. Spammer elimination can be accommodated in probabilistic models [79]. General theoretical conditions about eliminating low-quality workers have been given recently [80].

For machine learning the crowdsourcing step is generally used to collect labels, whereas the performance of the model learned with these data, rather than the quality of labels themselves, is more important. There are many studies about learning from weak teachers or crowd labels [81,82], which is closely related to learning with label noise (introduced at the beginning of this section); a distinction lies in the fact that, for a crowdsourcing setting, one can conveniently draw crowd labels repeatedly for an instance. Thus, in crowdsourcing learning it is crucial to consider the cost-saving effect, and an upper bound for the minimally sufficient number of crowd labels, i.e. the minimal cost required for effective crowdsourcing learning, is given [83]. Many studies work on task assignment and budget allocation, trying to balance between accuracy and label cost. For this purpose, non-adaptive task assignment mechanisms, which assign tasks offline [84,85], and adaptive mechanisms, which assign tasks online [86,87], have both been studied with theoretical support. Note that most studies adopt the Dawid–Skene model [88], which assumes that the potential cost for different tasks is the same, whereas more complicated cost settings are rarely explored.

Designing an effective crowdsourcing protocol is also important. In [89], an *unsure* option is provided, such that workers are not forced to give a label when they have low confidence; this option helps improve the labeling reliability with theoretical support [90]. In [91], a 'double or nothing' incentive compatible mechanism is proposed to ensure workers behave honestly based on their self-confidence; this protocol is provable to avoid spammers from the crowd, under the assumption that every worker wants to maximize their expected payment.

CONCLUSION

Supervised learning techniques have achieved great success when there is strong supervision information like a large amount of training examples with ground-truth labels. In real tasks, however, collecting supervision information requires costs, and thus, it is usually desirable to be able to do weakly supervised learning.

This article focuses on three typical types of weak supervision: incomplete, inexact and inaccurate supervision. Though they are discussed separately, in practice they often occur simultaneously, as illustrated in Fig. 1, and there are some relevant studies on such 'mixed' cases [52,92,93]. In addition, there are some other types of weak supervision. For example, time-delayed supervision, which is mainly tackled by reinforcement learning [94], can also be regarded as weak supervision. Note that due to the page limit, this article actually serves more as a literature index rather than a comprehensive review. Readers interested in some details are encouraged to read the corresponding references. Note that more and more researchers have recently been attracted to weakly supervised learning; e.g. *partially supervised learning* focuses mostly on learning with incomplete supervision [95], and there have been some other discussions about weak supervision [96,97].

To simplify the discussion, this article focuses on binary classification, although most discussions can be extended to *multi-class* or *regression* learning with slight modifications. Note that more complicated situations may occur with multi-class tasks [98]. It will become even more complicated if *multi-label learning* [99] is considered, where each example can be associated with multiple labels simultaneously. Take incomplete supervision as an example: in addition to labeled/unlabeled instances, multi-label tasks may encounter partially labeled instances, i.e. a training instance is given with ground-truth for a subset of its labels [100]. Even if only labeled/unlabeled data are considered, there are more design options than the single-label setting; e.g. for active learning, given a selected unlabeled instance, in multi-label tasks it is possible to query all labels of the instance [101], a specific label of the instance [102], or relevance ordering of a pair of labels for the instance [103]. Nevertheless, no matter what kinds of data and tasks are concerned, weakly supervised learning is becoming more and more important.

FUNDING

This work was supported by the National Natural Science Foundation of China (61333014), the National Key Basic Research Program of China (2014CB340501), and the Collaborative Innovation Center of Novel Software Technology and Industrialization.

REFERENCES

1. Goodfellow I, Bengio Y and Courville A. *Deep Learning*. Cambridge: MIT Press, 2016.
2. Settles B. Active learning literature survey. *Technical Report 1648*. Department of Computer Sciences, University of Wisconsin at Madison, Wisconsin, WI, 2010 [http://pages.cs.wisc.edu/~bsettles/pub/settles.activelearning.pdf].
3. Chapelle O, Schölkopf B and Zien A (eds). *Semi-Supervised Learning*. Cambridge: MIT Press, 2006.
4. Zhu X. Semi-supervised learning literature survey. Technical Report 1530. Department of Computer Sciences, University of Wisconsin at Madison, Madison, WI, 2008 [http://www.cs.wisc.edu/~jerryzhu/pub/ssl˙survey.pdf].
5. Zhou Z-H and Li M. Semi-supervised learning by disagreement. *Knowl Inform Syst* 2010; **24**: 415–39.
6. Huang SJ, Jin R and Zhou ZH. Active learning by querying informative and representative examples. *IEEE Trans Pattern Anal Mach Intell* 2014; **36**: 1936–49.
7. Lewis D and Gale W. A sequential algorithm for training text classifiers. In *17th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland*, 1994; 3–12.
8. Seung H, Opper M and Sompolinsky H. Query by committee. In *5th ACM Workshop on Computational Learning Theory, Pittsburgh, PA*, 1992; 287–94.
9. Abe N and Mamitsuka H. Query learning strategies using boosting and bagging. In *15th International Conference on Machine Learning, Madison, WI*, 1998; 1–9.
10. Nguyen HT and Smeulders AWM. Active learning using pre-clustering. In *21st International Conference on Machine Learning, Banff, Canada*, 2004; 623–30.
11. Dasgupta S and Hsu D. Hierarchical sampling for active learning. In *25th International Conference on Machine Learning, Helsinki, Finland*, 2008; 208–15.
12. Wang Z and Ye J. Querying discriminative and representative samples for batch mode active learning. In *19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Chicago, IL*, 2013; 158–66.
13. Dasgupta S, Kalai AT and Monteleoni C. Analysis of perceptron-based active learning. In *28th Conference on Learning Theory, Paris, France*, 2005; 249–63.
14. Dasgupta S. Analysis of a greedy active learning strategy. In *Advances in Neural Information Processing Systems 17, Cambridge, MA: MIT Press*, 2005; 337–44.
15. Kääriäinen M. Active learning in the non-realizable case. In *21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics, Sydney, Australia*, 2006; 63–77.
16. Balcan MF, Broder AZ and Zhang T. Margin based active learning. In *20th Annual Conference on Learning Theory, San Diego, CA*, 2007; 35–50.
17. Hanneke S. Adaptive rates of convergence in active learning. In *22nd Conference on Learning Theory, Montreal, Canada*, 2009.
18. Wang W and Zhou ZH. Multi-view active learning in the non-realizable case. In *Advances in Neural Information Processing Systems 23, Cambridge, MA: MIT Press*, 2010; 2388–96.
19. Miller DJ and Uyar HS. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *Advances in Neural Information Processing Systems 9, Cambridge, MA: MIT Press*, 1997; 571–7.
20. Nigam K, McCallum AK and Thrun S *et al.* Text classification from labeled and unlabeled documents using EM. *Mach Learn* 2000; **39**: 103–34.
21. Dempster AP, Laird NM and Rubin DB. Maximum likelihood from incomplete data via the EM algorithm. *J Roy Stat Soc B Stat Meth* 1977; **39**: 1–38.
22. Fujino A, Ueda N and Saito K. A hybrid generative/discriminative approach to semi-supervised classifier design. In *20th National Conference on Artificial Intelligence, Pittsburgh, PA*, 2005; 764–9.
23. Blum A and Chawla S. Learning from labeled and unlabeled data using graph mincuts. In *ICML*, 2001; 19–26.
24. Zhu X, Ghahramani Z and Lafferty J. Semi-supervised learning using Gaussian fields and harmonic functions. In *20th International Conference on Machine Learning, Washington, DC*, 2003; 912–9.
25. Zhou D, Bousquet O and Lal TN *et al.* Learning with local and global consistency. In *Advances in Neural Information Processing Systems 16, Cambridge, MA: MIT Press*, 2004; 321–8.
26. Carreira-Perpinan MA and Zemel RS. Proximity graphs for clustering and manifold learning. In *Advances in Neural Information Processing Systems 17, Cambridge, MA: MIT Press*, 2005; 225–32.

27. Wang F and Zhang C. Label propagation through linear neighborhoods. In *23rd International Conference on Machine Learning, Pittsburgh, PA*, 2006; 985–92.
28. Hein M and Maier M. Manifold denoising. In *Advances in Neural Information Processing Systems 19, Cambridge, MA: MIT Press*, 2007; pp. 561–8.
29. Joachims T. Transductive inference for text classification using support vector machines. In *16th International Conference on Machine Learning, Bled, Slovenia*, 1999; 200–9.
30. Chapelle O and Zien A. Semi-supervised learning by low density separation. In *10th International Workshop on Artificial Intelligence and Statistics, Barbados*, 2005; 57–64.
31. Li YF, Tsang IW and Kwok JT *et al.* Convex and scalable weakly labeled SVMs. *J Mach Learn Res* 2013; **14**: 2151–88.
32. Blum A and Mitchell T. Combining labeled and unlabeled data with co-training. In *11th Conference on Computational Learning Theory, Madison, WI*, 1998; 92–100.
33. Zhou Z-H and Li M. Tri-training: exploiting unlabeled data using three classifiers. *IEEE Trans Knowl Data Eng* 2005; **17**: 1529–41.
34. Zhou Z-H. When semi-supervised learning meets ensemble learning. In *8th International Workshop on Multiple Classifier Systems, Reykjavik, Iceland*, 2009; 529–38.
35. Zhou Z-H. *Ensemble Methods: Foundations and Algorithms*. Boca Raton: CRC Press, 2012.
36. Cozman FG and Cohen I. Unlabeled data can degrade classification performance of generative classifiers. In *15th International Conference of the Florida Artificial Intelligence Research Society, Pensacola, FL*, 2002; 327–31.
37. Li YF and Zhou ZH. Towards making unlabeled data never hurt. *IEEE Trans Pattern Anal Mach Intell* 2015; **37**: 175–88.
38. Castelli V and Cover TM. On the exponential value of labeled samples. *Pattern Recogn Lett* 1995; **16**: 105–11.
39. Wang W and Zhou ZH. Theoretical foundation of co-training and disagreement-based algorithms. arXiv:1708.04403, 2017.
40. Dietterich TG, Lathrop RH and Lozano-Pérez T. Solving the multiple-instance problem with axis-parallel rectangles. *Artif Intell* 1997; **89**: 31–71.
41. Foulds J and Frank E. A review of multi-instance learning assumptions. *Knowl Eng Rev* 2010; **25**: 1–25.
42. Zhou Z-H. Multi-instance learning from supervised view. *J Comput Sci Technol* 2006; **21**: 800–9.
43. Zhou Z-H and Zhang M-L. Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowl Inform Syst* 2007; **11**: 155–70.
44. Wei X-S, Wu J and Zhou Z-H Scalable algorithms for multi-instance learning. *IEEE Trans Neural Network Learn Syst* 2017; **28**:975–87.
45. Amores J. Multiple instance classification: review, taxonomy and comparative study. *Artif Intell* 2013; **201**: 81–105.
46. Zhou Z-H and Xu J-M. On the relation between multi-instance learning and semi-supervised learning. In *24th International Conference on Machine Learning, Corvallis, OR*, 2007; 1167–74.
47. Zhou Z-H, Sun Y-Y and Li Y-F. Multi-instance learning by treating instances as non-i.i.d. samples. In *26th International Conference on Machine Learning, Montreal, Canada*, 2009; 1249–56.
48. Chen Y and Wang JZ. Image categorization by learning and reasoning with regions. *J Mach Learn Res* 2004; **5**: 913–39.
49. Zhang Q, Yu W and Goldman SA *et al.* Content-based image retrieval using multiple-instance learning. In *19th International Conference on Machine Learning, Sydney, Australia*, 2002; 682–9.
50. Tang JH, Li HJ and Qi GJ *et al.* Image annotation by graph-based inference with integrated multiple/single instance representations. *IEEE Trans Multimed* 2010; **12**: 131–41.
51. Andrews S, Tsochantaridis I and Hofmann T. Support vector machines for multiple-instance learning. In *Advances in Neural Information Processing Systems 15, Cambridge, MA: MIT Press*, 2003; 561–8.
52. Settles B, Craven M and Ray S. Multiple-instance active learning. In *Advances in Neural Information Processing Systems 20, Cambridge, MA: MIT Press*, 2008; 1289–96.
53. Jorgensen Z, Zhou Y and Inge M. A multiple instance learning strategy for combating good word attacks on spam filters. *J Mach Learn Res* 2008; **8**: 993–1019.
54. Fung G, Dundar M and Krishnappuram B *et al.* Multiple instance learning for computer aided diagnosis. In *Advances in Neural Information Processing Systems 19, Cambridge, MA: MIT Press*, 2007; 425–32.
55. Viola P, Platt J and Zhang C. Multiple instance boosting for object detection. In *Advances in Neural Information Processing Systems 18, Cambridge, MA: MIT Press*, 2006; 1419–26.
56. Felzenszwalb PF, Girshick RB and McAllester D *et al.* Object detection with discriminatively trained part-based models. *IEEE Trans Pattern Anal Mach Intell* 2010; **32**: 1627–45.
57. Zhu J-Y, Wu J and Xu Y *et al.* Unsupervised object class discovery via saliency-guided multiple class learning. *IEEE Trans Pattern Anal Mach Intell* 2015; **37**: 862–75.
58. Babenko B, Yang MH and Belongie S. Robust object tracking with online multiple instance learning. *IEEE Trans Pattern Anal Mach Intell* 2011; **33**: 1619–32.
59. Wei X-S and Zhou Z-H. An empirical study on image bag generators for multi-instance learning. *Mach Learn* 2016; **105**: 155–98.
60. Liu G, Wu J and Zhou ZH. Key instance detection in multi-instance learning. In *4th Asian Conference on Machine Learning, Singapore*, 2012; 253–68.
61. Xu X and Frank E. Logistic regression and boosting for labeled bags of instances. In *8th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Sydney, Australia*, 2004; 272–81.
62. Chen Y, Bi J and Wang JZ. MILES: multiple-instance learning via embedded instance selection. *IEEE Trans Pattern Anal Mach Intell* 2006; **28**: 1931–47.
63. Weidmann N, Frank E and Pfahringer B. A two-level learning method for generalized multi-instance problem. In *14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia*, 2003; 468–79.
64. Long PM and Tan L. PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. *Mach Learn* 1998; **30**: 7–21.
65. Auer P, Long PM and Srinivasan A. Approximating hyper-rectangles: learning and pseudo-random sets. *J Comput Syst Sci* 1998; **57**: 376–88.
66. Blum A and Kalai A. A note on learning from multiple-instance examples. *Mach Learn* 1998; **30**: 23–9.
67. Sabato S and Tishby N. Homogenous multi-instance learning with arbitrary dependence. In *22nd Conference on Learning Theory, Montreal, Canada*, 2009.
68. Frénay B and Verleysen M. Classification in the presence of label noise: a survey. *IEEE Trans Neural Network Learn Syst* 2014; **25**: 845–69.
69. Angluin D and Laird P. Learning from noisy examples. *Mach Learn* 1988; **2**: 343–70.
70. Blum A, Kalai A and Wasserman H. Noise-tolerant learning, the parity problem, and the statistical query model. *J ACM* 2003; **50**: 506–19.
71. Gao W, Wang L and Li YF *et al.* Risk minimization in the presence of label noise. In *30th AAAI Conference on Artificial Intelligence, Phoenix, AZ*, 2016; 1575–81.
72. Brodley CE and Friedl MA. Identifying mislabeled training data. *J Artif Intell Res* 1999; **11**: 131–67.
73. Muhlenbach F, Lallich S and Zighed DA. Identifying and handling mislabelled instances. *J Intell Inform Syst* 2004; **22**: 89–109.

74. Brabham DC. Crowdsourcing as a model for problem solving: an introduction and cases. *Convergence* 2008; **14**: 75–90.

75. Sheng VS, Provost FJ and Ipeirotis PG. Get another label? Improving data quality and data mining using multiple, noisy labelers. In *14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, NV*, 2008; 614–22.

76. Snow R, O'Connor B and Jurafsky D *et al.* Cheap and fast - but is it good? Evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, Honolulu, HI*, 2008; 254–63.

77. Raykar VC, Yu S and Zhao LH *et al.* Learning from crowds. *J Mach Learn Res* 2010; **11**: 1297–322.

78. Whitehill J, Ruvolo P and Wu T *et al.* Whose vote should count more: optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems 22, Cambridge, MA: MIT Press*, 2009; 2035–43.

79. Raykar VC and Yu S. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *J Mach Learn Res* 2012; **13**: 491–518.

80. Wang W and Zhou ZH. Crowdsourcing label quality: a theoretical analysis. *Sci China Inform Sci* 2015; **58**: 1–12.

81. Dekel O and Shamir O. Good learners for evil teachers. In *26th International Conference on Machine Learning, Montreal, Canada*, 2009; 233–40.

82. Urner R, Ben-David S and Shamir O. Learning from weak teachers. In *15th International Conference on Artificial Intelligence and Statistics, La Palma, Canary Islands*, 2012; 1252–60.

83. Wang L and Zhou ZH. Cost-saving effect of crowdsourcing learning. In *25th International Joint Conference on Artificial Intelligence, New York, NY*, 2016; 2111–7.

84. Karger DR, Sewoong O and Devavrat S. Iterative learning for reliable crowdsourcing systems. In *Advances in Neural Information Processing Systems 24, Cambridge, MA: MIT Press*, 2011; 1953–61.

85. Tran-Thanh L, Venanzi M and Rogers A *et al.* Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks. In *12th International conference on Autonomous Agents and Multi-Agent Systems, Saint Paul, MN*, 2013; 901–8.

86. Ho CJ, Jabbari S and Vaughan JW. Adaptive task assignment for crowdsourced classification. In *30th International Conference on Machine Learning, Atlanta, GA*, 2013; 534–42.

87. Chen X, Lin Q and Zhou D. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *30th International Conference on Machine Learning, Atlanta, GA*, 2013; 64–72.

88. Dawid AP and Skene AM. Maximum likelihood estimation of observer error-rates using the EM algorithm. *J Roy Stat Soc C Appl Stat* 1979; **28**: 20–8.

89. Zhong J, Tang K and Zhou Z-H. Active learning from crowds with unsure option. In *24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina*, 2015; 1061–7.

90. Ding YX and Zhou ZH. *Crowdsourcing with unsure opinion.* arXiv:1609.00292, 2016.

91. Shah NB and Zhou D. Double or nothing: multiplicative incentive mechanisms for crowdsourcing. In *Advances in Neural Information Processing Systems 28, Cambridge, MA: MIT Press*, 2015; 1–9.

92. Rahmani R and Goldman SA. MISSL: multiple-instance semi-supervised learning. In *23rd International Conference on Machine Learning, Pittsburgh, PA*, 2006; 705–12.

93. Yan Y, Rosales R and Fung G *et al.* Active learning from crowds. In *28th International Conference on Machine Learning, Bellevue, WA*, 2011; 1161–8.

94. Sutton RS and Barto AG. *Reinforcement Learning: An Introduction.* Cambridge: MIT Press, 1998.

95. Schwenker F and Trentin E. Partially supervised learning for pattern recognition. *Pattern Recogn Lett* 2014; **37**: 1–3.

96. Garcia-Garcia D and Williamson RC. Degrees of supervision. In *Advances in Neural Information Processing Systems 17, Cambridge, MA: MIT Press Workshops*, 2011.

97. Hernández-González J, Inza I and Lozano JA. Weak supervision and other non-standard classification problems: a taxonomy. *Pattern Recogn Lett* 2016; **69**: 49–55.

98. Kuncheva LI, Rodíguez JJ and Jackson AS. Restricted set classification: who is there? *Pattern Recogn* 2017; **63**:158–70.

99. Zhang M-L and Zhou Z-H. A review on multi-label learning algorithms. *IEEE Trans Knowl Data Eng* 2014; **26**: 1819–37.

100. Sun YY, Zhang Y and Zhou ZH. Multi-label learning with weak label. In *24th AAAI Conference on Artificial Intelligence, Atlanta, GA*, 2010; 593–8.

101. Li X and Guo Y. Active learning with multi-label SVM classification. In *23rd International Joint Conference on Artificial Intelligence, Beijing, China*, 2013; 1479–85.

102. Qi GJ, Hua XS and Rui Y *et al.* Two-dimensional active learning for image classification. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Anchorage, AK*, 2008.

103. Huang SJ, Chen S and Zhou ZH. Multi-label active learning: query type matters. In *24th International Joint Conference on Artificial Intelligence, Buenos Aires, Argentina*, 2015; 946–52.