A Brief Introduction to Weakly Supervised Learning

Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology
Nanjing University, Nanjing 210023, China
zhouzh@nju.edu.cn

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 data labeling process. Thus, it is desired 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 are given with labels; inexact supervision where the training data are given with only coarse-grained labels; inaccurate supervision where the given labels are not always ground-truth.

1 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 which contains a large amount of training examples, each corresponding to an event/object. A training example consists of two parts: a feature vector (or called *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 [37], require ground-truth labels 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 data labeling process. Thus, it is desired 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 are given with labels whereas the other data remain unlabeled. Such 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 desired to have every object in the images be 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 situation occurs, e.g., when the image annotator is careless or weary, or some images are difficult to be categorized.

Weakly supervised learning is an umbrella covering a variety of studies which attempt to construct predictive models by learning with weak supervision. In this article, we will introduce some progress about 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 the 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 assume that $(\boldsymbol{x}_i, y_i)$'s are generated according to an unknown identical and independent distribution $\mathcal{D}$; in other words, $(\boldsymbol{x}_i, y_i)$'s are *i.i.d.* samples. Figure 1 provides an illustration of the three types of weak supervision we will discuss in this article.

2 Incomplete Supervision

Incomplete supervision concerns about the situation where 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

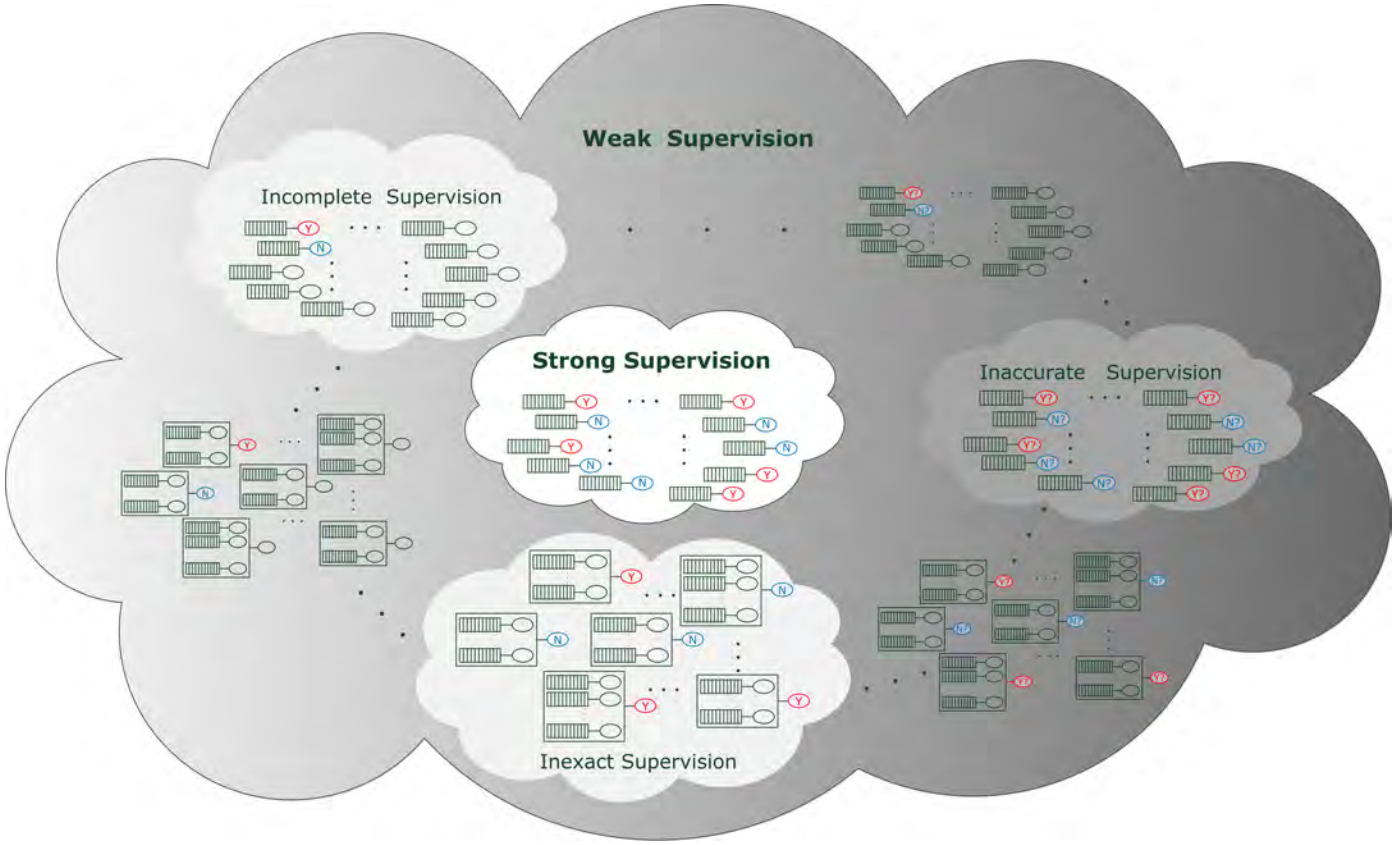


Figure 1. Illustration of three typical types of weak supervision. Bars denote feature vectors; red/blue marks labels; "?" implies label may be inaccurate. Intermediate subgraphs depict some situations with mixed types of weak supervision.

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 that of supervised learning with strong supervision, as defined in the end of Section 1. For the convenience of discussion, we also call the l labeled examples as "labeled data" whereas the u unlabeled instances as "unlabeled data".

There are two major techniques for this purpose, i.e., *active learning* [65] and *semi-supervised learning* [16, 97, 102].

Active learning assumes that there is an "oracle", such as a human expert, 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* whose main difference with (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 "close-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 to be test data. Figure 2 intuitively shows the difference between active learning, (pure) semi-supervised learning and tranductive learning.

2.1 With Human Intervention

Active learning [65] 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

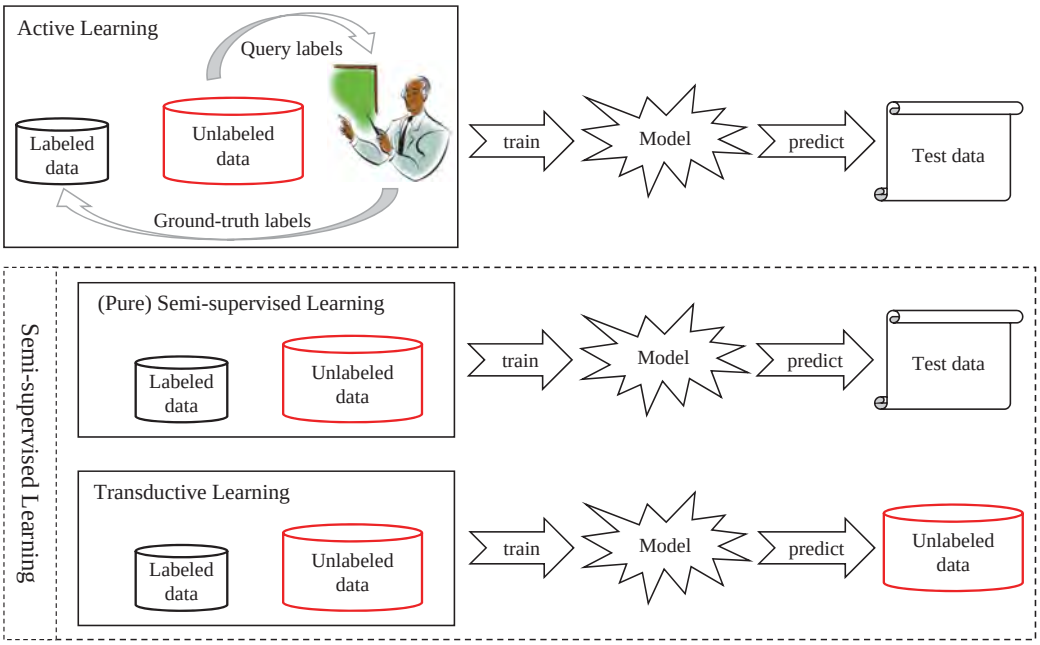


Figure 2. Active learning, (pure) semi-supervised learning, and transductive learning

cost for training a good model can be minimized.

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* [43]. 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 is with the least confidence [49]. The latter generates multiple learners and then queries the unlabeled instance on which the learners disagree to the most [1, 67]. Approaches based on representativeness generally aim to exploit the cluster structure of unlabeled data, usually by a clustering method [23, 57].

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 depend 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 [43, 82].

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 [22, 24]. 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 assumption about noise model, the lower bound of active learning matches the upper bound of passive learning [46]; in other words, active learning does not offer much help. By assuming Tsybakov noise model, it has been proven that exponential improvement can be obtained for bounded noise [7, 38]; if some special data characteristics, such as multi-view structure, can be exploited, exponential improvement can even be achieved for unbounded noise [79]. In other words, even for difficult cases, active learning still can be helpful with delicate designs.

2.2 Without Human Intervention

Semi-supervised learning [16, 97, 102] 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 [55], assume that data come from a Gaussian mixture model with

n mixture 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 posterior* criterion, we have the model

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

where

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

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 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 are not explicitly with label information, they implicitly convey some information about data distribution which can be helpful for predictive modelling.

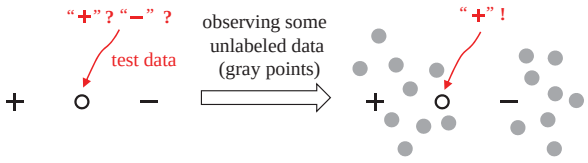


Figure 3. Illustration of the usefulness of unlabeled data

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 [55, 58] 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 [27]. These methods differ by fitting data using different generative models. To get good performance, one usually needs domain knowledge to determine adequate generative model. There are also attempts to combine advantages of generative and discriminative approaches [33].

Graph-based methods [8, 92, 103] construct a graph, where the nodes correspond to training instances and edges correspond to relation (usually some kind of similarity or distance) between instances, and then propagate label information on the graph according to some criteria; for example, labels can be propagated inside different subgraphs separated by minimum cut [8]. Apparently, the performance will heavily depends on how the graph is constructed [14, 39, 77]. 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) [17,44,51]. Figure 4 demonstrates the difference between conventional supervised SVM and S3VM. It is evident that S3VMs try to identify a classification boundary which 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 [11, 96, 97] generate multiple learners and let them collaborate to exploit unlabeled data, where the disagreement among the learners is crucial to ensure the learning process to continue. The most famous representative, co-training [11], works by training two learners from two different feature sets (or called two *views*). In each iteration, each learner picks its most

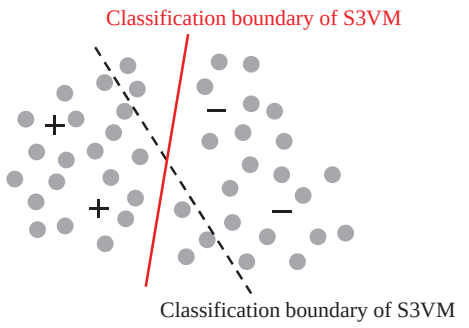


Figure 4. Illustration of the usefulness of unlabeled data

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 [94, 95]. 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 contradictive, 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 [21]; however, only recently, some solid progress are reported [52]. 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 [95].

There are abundant theoretical studies about semi-supervised learning [102], some even earlier than the name of semi-supervised learning being coined [15]. In particular, a thorough study about disagreement-based methods is presented recently [81].

3 Inexact Supervision

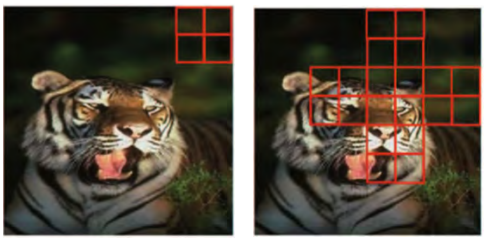
Inexact supervision concerns about the situation where some supervision information is given, but not as exact as desired. A typical scenario is when only coarse-grained label information available. For example, in the problem of drug activity prediction [28], the goal is to build a model to predict 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 the known molecules, however, human experts only know whether the molecules are qualified or not, instead of knowing what special shapes are decisive.

Formally, the task is to learn $f : \mathcal{X} \mapsto \mathcal{Y}$ from a training data set $D = \{(X_1, y_1), \ldots, (X_m, y_m)\}$, where $X_i = \{\boldsymbol{x}_{i1}, \ldots, \boldsymbol{x}_{i,m_i}\} \subseteq \mathcal{X}$ is called a *bag*, $\boldsymbol{x}_{ij} \in \mathcal{X}$ ($j \in \{1, \ldots, m_i\}$) 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 $\boldsymbol{x}_{ip}$ which is positive, while $p \in \{1, \ldots, m_i\}$ is unknown. The goal is to predict labels for unseen bags. This is called *multi-instance learning* [28, 31].

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 [93]; some other algorithms attempt to adapt the multi-instance representation to single-instance algorithms through representation transformation [83, 100]. There is also a categorization [2] which groups the algorithms into instance-space paradigm where the instance-level responses are aggregated, bag-space paradigm where the bags are treated as a whole, and 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, [99] indicates that the instances in multi-instance learning should not be assumed as independent although the bags can be treated as *i.i.d.* samples, and based on this insight, some effective algorithms have been developed [98].

Multi-instance learning has been successfully applied to various tasks, such as image categorization/retrieval/annotation [20, 73, 90], text categorization [3, 66], spam detection [45], medical diagnosis [34], face/object detection [30, 76], object class discovery [101], object tracking [6], 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 formation of instances in a bag (i.e., shapes of a molecule), the instances need to be generated for each bag. 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 significant influence on learning performance, only recently, an extensive study about image bag generators is reported [84], which discloses that some simple dense-sampling bag generators

perform better than complicated ones. Figure 5 shows two simple yet effective image bag generators.



(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. SB will generate 16 instances for the image, by regarding each patch consisted of the four blobs as one instance, and sliding without overlap. SBN will generate 9 instances for the image, by regarding the patch consisted of the twenty blobs as one instance, and sliding with overlap.

The original goal of multi-instance learning is to predict labels for unseen bags; however, there are studies trying to identify the *key instance* which enables a positive bag to be positive [51, 53]. 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 [28] assumes that each positive bag must contain a key instance, whereas there are studies which assume that there is no key instance and every instance contributes to the bag label [19, 87], or even assume that there are multiple concepts and a bag is positive only when the bag contains instances satisfying every concept [85]. More variants can be found in [31].

Early theoretical results [5, 9, 54] show that multi-instance learning is hard for *heterogeneous* case where each instance in the bag is classified by a different rule, while it is learnable for *homogeneous* case where 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 homogeneous class there are at least some cases learnable for arbitrary distribution over bags [63]. 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.

4 Inaccurate Supervision

Inaccurate supervision concerns about the situation where 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 has been shown in the end of Section 1, except that the y_i's in the training data set may be incorrect.

A typical scenario is learning with label noise [32]. There are many theoretical studies [4, 10, 35], among which most assumes random classification noise, i.e., labels are subject to random noise. In practice, a basic idea is to identify the potentially mislabeled examples [13], and then try to make some correction. For example, a *data editing* approach [56] 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 suspectable if it is associated with many cut edges. The suspected instances can be either removed or relabeled, as illustrated in Figure 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 neighborhood is usually less reliable when data are sparse.

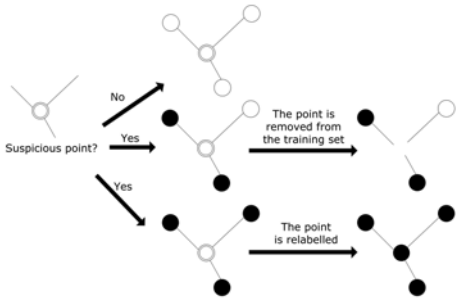


Figure 6. Identify and remove/relabel suspicious points

An interesting recent scenario of inaccurate supervision occurs with *crowdsourcing* [12], 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 on 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 non-trivial 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 [95], is widely used in practice with good performance [69,70], and thus often used as a baseline. It is expected that if worker quality and task difficulty can be modelled, 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 [62, 86]. Minimax entropy principle has also been used [95]. Spammer elimination can be accommodated in probabilistic models [61]. 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 concerned. There are many studies about learning from weak teachers or crowd labels [26, 75], which is closely related to learning with label noise (introduced in the beginning of this section); a distinction lies in the fact that for crowdsourcing setting, one can conveniently draw crowd labels for an instance repeatedly. 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 [78]. 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 off-line [47, 74], and adaptive mechanisms which assign tasks online [18, 41], have both been studied with theoretical supports. Note that most studies adopt the Dawid-Skene model [25] which assumes that the potential cost for different tasks is the same, whereas more complicated cost settings are rarely explored.

Designing effective crowdsourcing protocol is also important. In [91], an *unsure* option is provided, such that workers are not forced to give a label when they feel with low confidence; this option helps improve the labeling reliability with theoretical support [29]. In [68], a "double or nothing" incentive compatible mechanism is proposed to ensure workers to 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.

5 Conclusion

Supervised learning techniques have achieved great success when there is strong supervision information like large amount of training examples with ground-truth labels. In real tasks, however, collecting supervision information requires costs, and thus, it is usually desired 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 Figure 1, and there are some relevant studies on such "mixed" cases [60,66,88]. In addition, there are some other types of weak supervision. For example, time-delayed supervision, which is mainly tackled by reinforcement learning [72], can also be regarded as weak supervision. Note that due to page limit, this article actually serves more like a literature index rather than a comprehensive review. Readers interested in some details are encouraged to read the corresponding references. Note that recently more and more researchers are attracted to weakly supervised learning, for example, *partially supervised learning* focuses mostly on learning with incomplete supervision [64], and there are some other discussion about weak supervision [36, 40].

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 [48]. It will become even more complicated if *multi-label learning* [89] is considered, where each example can be associated with multiple labels simultaneously. Taking incomplete supervision for an example: in addition to labeled/unlabeled instances, multi-label tasks may encounter partially labeled instance, i.e., a training instance is given with ground-truth for a subset of its labels [71]. Even if only labeled/unlabeled data are considered, there are more design options than single-label setting; for example, for active learning, given a selected unlabeled instance, in multi-label tasks it is possible to query all labels of the instance [50], a specific label of the instance [59], or relevance ordering of a pair of labels for the instance [42]. Nevertheless, no matter what kind of data and tasks are concerned, weakly supervised learning becomes more and more important.

Funding

This work was supported by the National 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] N. Abe and H. Mamitsuka. Query learning strategies using boosting and bagging. In *ICML*, pages 1–9, 1998.

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

[3] S. Andrews, I. Tsochantaridis, and T. Hofmann. Support vector machines for multiple-instance learning. In *NIPS*, pages 561–568. 2003.

[4] D. Angluin and P. Laird. Learning from noisy examples. *Machine Learning*, 2(4):343–370, 1988.

[5] P. Auer, P. M. Long, and A. Srinivasan. Approximating hyper-rectangles: Learning and pseudo-random sets. *Journal of Computer and System Sciences*, 57(3):376–388, 1998.

[6] B. Babenko, M.-H. Yang, and S. Belongie. Robust object tracking with online multiple instance learning. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 33(8):1619–1632, 2011.

[7] M.-F. Balcan, A. Z. Broder, and T. Zhang. Margin based active learning. In *COLT*, pages 35–50, 2007.

[8] A. Blum and S. Chawla. Learning from labeled and unlabeled data using graph mincuts. In *ICML*, pages 19–26, 2001.

[9] A. Blum and A. Kalai. A note on learning from multiple-instance examples. *Machine Learning*, 30(1):23–29, 1998.

[10] A. Blum, A. Kalai, and H. Wasserman. Noise-tolerant learning, the parity problem, and the statistical query model. *Journal of the ACM*, 50(4):506–519, 2003.

[11] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In *COLT*, pages 92–100, 1998.

[12] D. C. Brabham. Crowdsourcing as a model for problem solving: An introduction and cases. *Convergence*, 14(1):75–90, 2008.

[13] C. E. Brodley and M. A. Friedl. Identifying mislabeled training data. *Journal of Artificial Intelligence Research*, 11:131–167, 1999.

[14] M. A. Carreira-Perpinan and R. S. Zemel. Proximity graphs for clustering and manifold learning. In *NIPS*, pages 225–232. 2005.

[15] V. Castelli and T. M. Cover. On the exponential value of labeled samples. *Pattern Recognition Letters*, 16(1):105–111, 1995.

[16] O. Chapelle, B. Schölkopf, and A. Zien, editors. *Semi-Supervised Learning*. MIT Press, Cambridge, MA, 2006.

[17] O. Chapelle and A. Zien. Semi-supervised learning by low density separation. In *AISTATS*, pages 57–64. 2005.

[18] X. Chen, Q. Lin, and D. Zhou. Optimistic knowledge gradient policy for optimal budget allocation in crowdsourcing. In *ICML*, pages 64–72, 2013.

[19] Y. Chen, J. Bi, and J. Z. Wang. MILES: Multiple-instance learning via embedded instance selection. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 28(12):1931–1947, 2006.

[20] Y. Chen and J. Z. Wang. Image categorization by learning and reasoning with regions. *Journal of Machine Learning Research*, 5:913–939, 2004.

[21] F. G. Cozman and I. Cohen. Unlabeled data can degrade classification performance of generative classifiers. In *FLAIRS*, pages 327–331, 2002.

[22] S. Dasgupta. Analysis of a greedy active learning strategy. In *NIPS*, pages 337–344. 2005.

[23] S. Dasgupta and D. Hsu. Hierarchical sampling for active learning. In *ICML*, pages 208–215, 2008.

[24] S. Dasgupta, A. T. Kalai, and C. Monteleoni. Analysis of perceptron-based active learning. In *COLT*, pages 249–263, 2005.

[25] A. P. Dawid and A. M. Skene. Maximum likelihood estimation of observer error-rates using the EM algorithm. *Applied Statistics*, pages 20–28, 1979.

[26] O. Dekel and O. Shamir. Good learners for evil teachers. In *ICML*, pages 233–240, 2009.

[27] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

[28] T. G. Dietterich, R. H. Lathrop, and T. Lozano-Pérez. Solving the multiple-instance problem with axis-parallel rectangles. *Artificial Intelligence*, 89(1-2):31–71, 1997.

[29] Y.-X. Ding and Z.-H. Zhou. Crowdsourcing with unsure opinion. arXiv:1609.00292, 2016.

[30] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 32(9):1627–1645, 2010.

[31] J. Foulds and E. Frank. A review of multi-instance learning assumptions. *Knowledge Engineering Review*, 25(1):1–25, 2010.

[32] B. Frénay and M. Verleysen. Classification in the presence of label noise: A survey. *IEEE Trans. Neural Networks and Learning Systems*, 25(5):845–869, 2014.

[33] A. Fujino, N. Ueda, and K. Saito. A hybrid generative/discriminative approach to semi-supervised classifier design. In *AAAI*, pages 764–769, 2005.

[34] G. Fung, M. Dundar, B. Krishnappuram, and R. B. Rao. Multiple instance learning for computer aided diagnosis. In *NIPS*, pages 425–432. 2007.

[35] W. Gao, L. Wang, Y.-F. Li, and Z.-H. Zhou. Risk minimization in the presence of label noise. In *AAAI*, pages 1575–1581, 2016.

[36] D. Garcia-Garcia and R. C. Williamson. Degrees of supervision. In *NIPS Workshops*, 2011.

[37] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, MA, 2016.

[38] S. Hanneke. Adaptive rates of convergence in active learning. In *COLT*, 2009.

[39] M. Hein and M. Maier. Manifold denoising. In *NIPS*, pages 561–568. 2007.

[40] J. Hernández-González, I. Inza, and J. A. Lozano. Weak supervision and other non-standard classificaiton problems: A taxonomy. *Pattern Recognition Letters*, 69(1):49–55, 2016.

[41] C.-J. Ho, S. Jabbari, and J. W. Vaughan. Adaptive task assignment for crowdsourced classification. In *ICML*, pages 534–542, 2013.

[42] S.-J. Huang, S. Chen, and Z.-H. Zhou. Multi-label active learning: Query type matters. In *IJCAI*, pages 946–952, 2015.

[43] S.-J. Huang, R. Jin, and Z.-H. Zhou. Active learning by querying informative and representative examples. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 36:1936–1949, 2014.

[44] T. Joachims. Transductive inference for text classification using support vector machines. In *ICML*, pages 200–209, 1999.

[45] Z. Jorgensen, Y. Zhou, and M. Inge. A multiple instance learning strategy for combating good word attacks on spam filters. *Journal of Machine Learning Research*, 8:993–1019, 2008.

[46] M. Kääriäinen. Active learning in the non-realizable case. In *ACL*, pages 63–77, 2006.

[47] D. R. Karger, O. Sewoong, and S. Devavrat. Iterative learning for reliable crowdsourcing systems. In *NIPS*, pages 1953–1961. 2011.

[48] L. I. Kuncheva, J. J. Rodíguez, and A. S. Jackson. Restricted set classification: Who is there? *Pattern Recognition*, 63:158–170, 2017.

[49] D. Lewis and W. Gale. A sequential algorithm for training text classifiers. In *SIGIR*, pages 3–12, 1994.

[50] X. Li and Y. Guo. Active learning with multi-label SVM classification. In *IJCAI*, pages 1479–1485, 2013.

[51] Y.-F. Li, I. W. Tsang, J. T. Kwok, and Z.-H. Zhou. Convex and scalable weakly labeled SVMs. *Journal of Machine Learning Research*, 14:2151–2188, 2013.

[52] Y.-F. Li and Z.-H. Zhou. Towards making unlabeled data never hurt. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 37(1):175–188, 2015.

[53] G. Liu, J. Wu, and Z.-H. Zhou. Key instance detection in multi-instance learning. In *ACML*, pages 253–268, 2012.

[54] P. M. Long and L. Tan. PAC learning axis-aligned rectangles with respect to product distributions from multiple-instance examples. *Machine Learning*, 30(1):7–21, 1998.

[55] D. J. Miller and H. S. Uyar. A mixture of experts classifier with learning based on both labelled and unlabelled data. In *NIPS*, pages 571–577. 1997.

[56] F. Muhlenbach, S. Lallich, and D. A. Zighed. Identifying and handling mislabelled instances. *Journal of Intelligent Information Systems*, 22:89–109, 2004.

[57] H. T. Nguyen and A. W. M. Smeulders. Active learning using pre-clustering. In *ICML*, pages 623–630, 2004.

[58] K. Nigam, A. K. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine Learning*, 39(2-3):103–134, 2000.

[59] G.-J. Qi, X.-S. Hua, Y. Rui, J. Tang, and H.-J. Zhang. Two-dimensional active learning for image classification. In *CVPR*, 2008.

[60] R. Rahmani and S. A. Goldman. MISSL: Multiple-instance semi-supervised learning. In *ICML*, pages 705–712, 2006.

[61] V. C. Raykar and S. Yu. Eliminating spammers and ranking annotators for crowdsourced labeling tasks. *Journal of Machine Learning Research*, 13:491–518, 2012.

[62] V. C. Raykar, S. Yu, L. H. Zhao, G. H. Valadez, C. Florin, L. Bogoni, and L. Moy. Learning from crowds. *Journal of Machine Learning Research*, 11:1297–1322, 2010.

[63] S. Sabato and N. Tishby. Homogenous multi-instance learning with arbitrary dependence. In *COLT*, 2009.

[64] F. Schwenker and E. Trentin. Partially supervised learning for pattern recognition. *Pattern Recognition Letters*, 37:1–3, 2014.

[65] B. Settles. 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.

[66] B. Settles, M. Craven, and S. Ray. Multiple-instance active learning. In *NIPS*, pages 1289–1296. 2008.

[67] H. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *COLT*, pages 287–294, 1992.

[68] N. B. Shah and D. Zhou. Double or nothing: Multiplicative incentive mechanisms for crowdsourcing. In *NIPS*, pages 1–9. 2015.

[69] V. S. Sheng, F. J. Provost, and P. G. Ipeirotis. Get another label? improving data quality and data mining using multiple, noisy labelers. In *KDD*, pages 614–622, 2008.

[70] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *EMNLP*, pages 254–263, 2008.

[71] Y.-Y. Sun, Y. Zhang, and Z.-H. Zhou. Multi-label learning with weak label. In *AAAI*, pages 593–598, 2010.

[72] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[73] J.-H. Tang, H.-J. Li, G.-J. Qi, and T.-S. Chua. Image annotation by graph-based inference with integrated multiple/single instance representations. *IEEE Trans. Multimedia*, 12(2):131–141, 2010.

[74] L. Tran-Thanh, M. Venanzi, A. Rogers, and N. R. Jennings. Efficient budget allocation with accuracy guarantees for crowdsourcing classification tasks. In *AAMAS*, pages 901–908, 2013.

[75] R. Urner, S. Ben-David, and O. Shamir. Learning from weak teachers. In *AISTATS*, pages 1252–1260, 2012.

[76] P. Viola, J. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, pages 1419–1426. 2006.

[77] F. Wang and C. Zhang. Label propagation through linear neighborhoods. In *ICML*, pages 985–992, 2006.

[78] L. Wang and Z.-H. Zhou. Cost-saving effect of crowdsourcing learning. In *IJCAI*, pages 2111–2117, 2016.

[79] W. Wang and Z.-H. Zhou. Multi-view active learning in the non-realizable case. In *NIPS*, pages 2388–2396. 2010.

[80] W. Wang and Z.-H. Zhou. Crowdsourcing label quality: A theoretical analysis. *Science China Information Sciences*, 58(11):1–12, 2015.

[81] W. Wang and Z.-H. Zhou. Theoretical foundation of co-training and disagreement-based algorithms. arXiv:1708.04403, 2017.

[82] Z. Wang and J. Ye. Querying discriminative and representative samples for batch mode active learning. In *KDD*, pages 158–166, 2013.

[83] X.-S. Wei, J. Wu, and Z.-H. Zhou. Scalable algorithms for multi-instance learning. *IEEE Trans. Neural Networks and Learning Systems*, 28(4):975–987, 2017.

[84] X.-S. Wei and Z.-H. Zhou. An empirical study on image bag generators for multi-instance learning. *Machine Learning*, 105(2):155–198, 2016.

[85] N. Weidmann, E. Frank, and B. Pfahringer. A two-level learning method for generalized multi-instance problem. In *ECML*, pages 468–479, 2003.

[86] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. R. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, pages 2035–2043. 2009.

[87] X. Xu and E. Frank. Logistic regression and boosting for labeled bags of instances. In *PAKDD*, pages 272–281, 2004.

[88] Y. Yan, R. Rosales, G. Fung, and J. G. Dy. Active learning from crowds. In *ICML*, pages 1161–1168, 2011.

[89] M.-L. Zhang and Z.-H. Zhou. A review on multi-label learning algorithms. *IEEE Trans. Knowledge and Data Engineering*, 26(8):1819–1837, 2014.

[90] Q. Zhang, W. Yu, S. A. Goldman, and J. E. Fritts. Content-based image retrieval using multiple-instance learning. In *ICML*, pages 682–689, 2002.

[91] J. Zhong, K. Tang, and Z.-H. Zhou. Active learning from crowds with unsure option. In *IJCAI*, pages 1061–1067, 2015.

[92] D. Zhou, O. Bousquet, T. N. Lal, J. Weston, and B. Schölkopf. Learning with local and global consistency. In *NIPS*, pages 321–328. 2004.

[93] Z.-H. Zhou. Multi-instance learning from supervised view. *Journal of Computer Science and Technology*, 21(5):800–809, 2006.

[94] Z.-H. Zhou. When semi-supervised learning meets ensemble learning. In *MCS*, pages 529–538, 2009.

[95] Z.-H. Zhou. *Ensemble Methods: Foundations and Algorithms*. CRC Press, 2012.

[96] Z.-H. Zhou and M. Li. Tri-training: Exploiting unlabeled data using three classifiers. *IEEE Trans. Knowledge and Data Engineering*, 17(11):1529–1541, 2005.

[97] Z.-H. Zhou and M. Li. Semi-supervised learning by disagreement. *Knowledge and Information Systems*, 24(3):415–439, 2010.

[98] Z.-H. Zhou, Y.-Y. Sun, and Y.-F. Li. Multi-instance learning by treating instances as non-i.i.d. samples. In *ICML*, pages 1249–1256, 2009.

[99] Z.-H. Zhou and J.-M. Xu. On the relation between multi-instance learning and semi-supervised learning. In *ICML*, pages 1167–1174, 2007.

[100] Z.-H. Zhou and M.-L. Zhang. Solving multi-instance problems with classifier ensemble based on constructive clustering. *Knowledge and Information Systems*, 11(2):155–170, 2007.

[101] 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 Analysis and Machine Intelligence*, 37(4):862–875, 2015.

[102] X. Zhu. 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.

[103] X. Zhu, Z. Ghahramani, and J. Lafferty. Semi-supervised learning using Gaussian fields and harmonic functions. In *ICML*, pages 912–919, 2003.