Artif Intell Rev (2016) 46:543–576
DOI 10.1007/s10462-016-9491-9

Learning from crowdsourced labeled data: a survey

Jing Zhang[1] · **Xindong Wu**[2] · **Victor S. Sheng**[3,4]

Published online: 2 July 2016

Abstract With the rapid growing of crowdsourcing systems, quite a few applications based on a supervised learning paradigm can easily obtain massive labeled data at a relatively low cost. However, due to the variable uncertainty of crowdsourced labelers, learning procedures face great challenges. Thus, improving the qualities of labels and learning models plays a key role in learning from the crowdsourced labeled data. In this survey, we first introduce the basic concepts of the qualities of labels and learning models. Then, by reviewing recently proposed models and algorithms on ground truth inference and learning models, we analyze connections and distinctions among these techniques as well as clarify the level of the progress of related researches. In order to facilitate the studies in this field, we also introduce open accessible real-world data sets collected from crowdsourcing systems and open source libraries and tools. Finally, some potential issues for future studies are discussed.

This research has been supported by the China Postdoctoral Science Foundation under grant 2016M590457, the Program for Changjiang Scholars and Innovative Research Team in University (PCSIRT) of the Ministry of Education, China, under grant IRT13059, the National 973 Program of China under grant 2013CB329604, and the US National Science Foundation under grant IIS-1115417.

✉ Jing Zhang
jzhang@njust.edu.cn

Xindong Wu
xwu@hfut.edu.cn

Victor S. Sheng
ssheng@uca.edu

[1] School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, People's Republic of China

[2] School of Computer Science and Information Engineering, Hefei University of Technology, Hefei 230009, People's Republic of China

[3] Department of Computer Science, University of Central Arkansas, Conway, AR 72035, USA

[4] Jiangsu Engineering Center of Network Monitoring, Nanjing University of Information Science and Technology, Nanjing, People's Republic of China

Keywords Crowdsourcing · Learning from crowds · Multiple noisy labeling · Label quality · Learning model quality · Ground truth inference

1 Introduction

With the emergency of crowdsourcing systems, more and more tasks could be distributed to and completed by ordinary users (as workers) on the Internet via a paradigm of micro outsourcing. The writer Howe published an article in the *Wired* magazine (Howe 2006), where he deeply analyzed the impact of a rising micro outsourcing via Internet on current business environments, and the term *crowdsourcing* was first introduced. The crowdsourcing is defined by Merriam-Webster as the process of obtaining needed services, ideas, or content by soliciting contributions from a large group of people, and especially from an online community, rather than from traditional employees or suppliers. A crowdsourcing process involves several steps. First, a large task is partitioned into numerous small pieces of subtasks which do not require special expertise to complete. All these small subtasks are distributed on a platform and picked up by online workers. After finishing these subtasks, each worker may obtain a small amount of reward. Accomplished subtasks are eventually assembled into a solution for the original task. The amateurism and the independency of workers are fundamental differences between crowdsourcing and traditional employment based labor mode. The success of crowdsourcing is due to the crowd capacity which is a series of organizational processes through the contribution of the group's intelligence via a public IT infrastructure (Prpic and Shukla 2013, 2014). Crowdsourcing not only has a commercial value but also has a social significance which lies in its reconstruction of social intelligence and resources with high efficiency to solve intelligent problems.

Crowdsourcing has already attracted a wide attention in the field of artificial intelligence. Researchers wish that human intelligence could be involved in the computational process and collaborate to solve the problem that cannot be conquered solely by machines, which is so-called human computation (Von Ahn 2009). Researchers have been attempting to utilize crowdsourcing to solve different kinds of problems (Bernstein et al. 2010; Carvalho et al. 2011; Grady and Lease 2010; Urbano et al. 2010; Corney et al. 2010; Muhammadi et al. 2015). As one of the most active branches in artificial intelligence, machine learning and its related fields (including data mining, information retrieval, pattern recognition, computer vision and image, etc.) are the first to realize that the development of crowdsourcing may bring great opportunities to themselves (Lease 2011). A large amount of labeled data can be collected quickly and cheaply via crowdsourcing, which are the cornerstone of a wide range of supervised learning methods. On the one hand, the importance of model training is higher than its complexity (Halevy et al. 2009). The diversity and the specificity of the data are greatly increased. On the other hand, fast and easily generating verification and test data sets on demand continuously optimize the iteration of learning models (Little et al. 2009), with which a learning system will evolve to a hybrid sustained learning system (Yan et al. 2010a). With the growing of crowdsourcing platforms, such as Mechanical Turk Amazon (MTurk) and CrowdFlower, more and more machine learning tasks are posted on these platforms. The tasks include the collection of ranking scores (Su et al. 2007), image and video annotation (Nowak and Rüger 2010; Sorokin and Forsyth 2008; Vondrick et al. 2013; Xu et al. 2012) and other online applications (Doan et al. 2011). Despite their diversity, the core of these applications is utilizing ordinary users to assign labels to objects, and these labeled data are used by intelligent algorithms to solve problems.

Traditionally, labeling tasks are usually processed by domain experts. This way provides accurate labels, but it is not efficient, and involves a high cost. In contrast, crowdsourcing is superior to expert labeling in the cost and efficiency. However, the label qualities in crowdsourcing are varied. This is due to significant differences among labelers in their knowledge levels, dedication and evaluation criteria. Therefore, how to improve the quality of the data and how to use these noisy data to build a learning model have become hot topics in recent machine learning studies. In the context of crowdsourcing, traditional machine learning techniques related to labeled data, such as classification, regression, active learning, transfer learning, and so on, are full of new challenges. Although novel research fruits have been coming out in recent years, research in this field is still in a very young stage, and a lot of issues are worthy of being further studied.

In this paper, we start with the basic concepts of the qualities of labels and learning models, review the mainstream research outcomes in the field of learning from crowdsourced labeled data and discuss some future research directions. This paper is organized as follows. In Sect. 2, the definitions of crowdsourced labeling, label quality and learning model quality are presented. In Sect. 3, the researches on the ground truth inference in crowdsourcing are reviewed. In Sect. 4, the researches on the building of learning models using crowdsourced data including active learning and emerging transfer learning are summarized. In Sect. 5, we summarize some open accessible real-world data sets and open source libraries and tools for crowdsourcing studies. Sect. 6 concludes the paper and discusses some future research directions.

2 Crowdsourced labeling

In this section, we first retrospect a brief history of labeling done by the crowd. Then, we focus on the label quality and the learning model quality.

2.1 Overview

The idea and practice of using ordinary Internet users to label data is proposed in 2004 by Luis von Ahn. He designed a game ESP to label images (Von Ahn and Dabbish 2004). In the ESP game, if two users provide the same labels for an image, both of them are awarded points. Luis von Ahn then developed the famous system reCaptcha (Von Ahn et al. 2008). reCaptcha uses two different Optical Character Recognition (OCR) systems to scan the same document.The respective outputs of these OCRs are aligned and compared with each other. Any word that is deciphered differently by both OCR programs will be marked as *suspicious*. The images of those suspicious words are recognized by human through the way when users sign in a system.

The success of reCaptcha shows the advantages of ordinary Internet users in solving the problem that the machine intelligence hitherto cannot solve. Although crowdsourcing systems had not been proposed during its development, the idea forms the basis of taking advantage of crowdsourcing. The first time that a commercial crowdsourcing system is used to study the label quality started in 2008, Snow et al. (2008) selected a 100-headline sample, posted them on MTurk, and collected 1000 affect labels for each of seven emotion types, where each example was labeled by 10 unique labelers. This work confirms that the quality of labeling by non-expert labelers is almost the same as that of labeling by experts, if a proper integration method is adopted. At this point, a large number of annotation tasks in machine

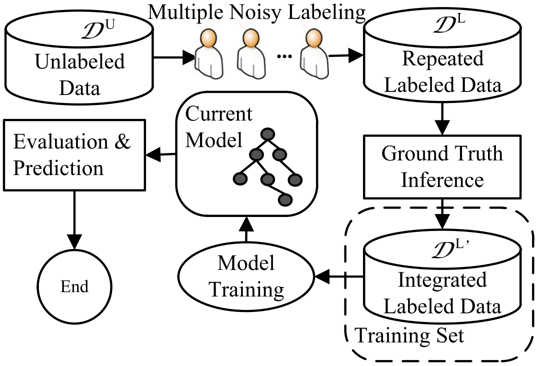


Fig. 1 A general framework for learning from crowdsourced labeled data

learning are completed with the help of crowdsourcing. Figure 1 shows a general framework for learning from crowdsourced labeled data.

2.2 Qualities of labels and learning models

The essence of labeling is to map the underlying features of an object to an upper level concept using the human intelligence (Michalski et al. 2013). Although crowdsourced labeling provides a promising blueprint, it suffers from low qualities of non-expert labelers. At present, since it is not allowed to modify features of objects, data quality and label quality are synonyms in this paper.

Definition 1 (*Label Quality*) The accuracy that the labels of objects derived from the crowdsourced labeling match their true values.

In a crowdsourcing study, a widely acknowledged assumption is that the true labels come from experts who have a perfect labeling quality known as the oracles. These true labels are used as a gold criterion for model and algorithm evaluation. Compared with the experts, labelers from a crowdsourcing platform have uneven expertise and dedication which result in a low quality of the data. The label of an object, which is inconsistent with its true value, is known as a noisy label. Reducing the noisy labels in a data set is a direct means to improve the quality of data.

Definition 2 (*Learning Model Quality*) The prediction performance of a learning model which is trained by a crowdsourced data set through a supervised learning algorithm.

Intuitively, the learning model quality is closely related to the label quality. Because of the noise presented in labels, the quality of the learning model trained by crowdsourced data is usually lower than that trained by expert labeled data. Thus, a direct way to improve the learning model quality is improving the label quality first. However, the learning model quality and the label quality are different things after all. On one hand, from the perspective of applications, some applications end up with the true labels inferred. For example, if we create a large-scale image database for user authentication in a login process, the requirement is that all images are correctly labeled for the subsequent comparison with user inputs. On the other hand, from the perspective of model learning, training a good learning model does not necessarily utilize all examples in a training set. For example, if a support vector machine (SVM) (Steinwart and Christmann 2008) is used to build a classifier, it should be efficient as long as support vectors are correctly labeled (Gu et al. 2014, 2015), even though the overall labeling quality is not so high. That is, in order to efficiently improve the quality of learning models under a constraint of budget, we must improve the labeling quality on those key data points, which can be achieved through active learning (Zhong et al. 2015).

2.3 Improving label quality

Generally, there are two ways to improve the label quality of the crowdsourced labeled data.

One is called *Quality Control on Task Designing* (Allahbakhsh et al. 2013). In this way, some quality control mechanism is introduced to regulate the behaviors of labelers, which tries to guide the labelers to provide high quality labels. For example, Dow et al. (2012) introduced a shepherd mechanism to provide real-time quality assurance. Kittur et al. (2011), Kulkarni et al. (2012) designed a proper workflow (including a quality supervision mechanism) to support complex tasks. Zhang et al. (2013) proposed a strategy for the quality control during label collection, which introduces periodical quality checkpoints to filter out low quality labelers and labels. The common defect of these methods is that the complicated quality control mechanism in data collection phase heavily relies on the background and historical information. Furthermore, the complex task allocation model damages the fairness and the efficiency, which results in a difficulty to design a reasonable reward mechanism. Currently, most commercial crowdsourcing systems, such as MTurk and CrowdFlower, do not provide such complicated functions. The design of the quality control mechanisms belongs to the disciplines of human–computer interaction, collaboration, gaming or operations research. In recent years, some researchers developed effective payment mechanisms for crowdsourcing. Shah and Zhou (2015) proposed a simple payment mechanism to incentivize workers to answer only the questions that they are sure of and skip the rest. This mechanism is the an incentive-compatible payment mechanism under a mild and natural no-free-lunch requirement, which can minimize possible payment to spammers. Furthermore, coupling with a (strictly proper) incentive-compatible compensation mechanism, Shah et al. (2015) introduced approval voting to utilize the expertise of workers who have partial knowledge of the true answer to solve the problem that the incentives of the workers are not aligned with those of the requesters.

The other is called *Quality Improvement after Data Collection*. Current crowdsourcing systems only provide simple trust models to filter out spammers, by which the label quality still hardly be guaranteed. Additional procedures are used to further improve the label quality. A common idea is repeated labeling. That is, an example is labeled by different labelers. Therefore, each example could obtain a set of multiple noisy labels. Then, we can design algorithms to induce an integrated label from its multiple label set. These algorithms are called ground truth inference algorithms. We hope that the integrated label of each example is its true label. Using repeated labeling to improve the label quality can be traced back to 1994. Smyth et al. used a maximum likelihood estimation algorithm and repeated labeling to deal with the uncertainties in Venus image annotations (Smyth et al. 1994, 1995). The ground truth inference algorithms will be discussed in Sect. 3. The integrated labels greatly improve the data quality, which makes the collected data to meet the needs of training models.

2.4 Challenges of learning from crowdsourcing

Both label integration and learning model training under the crowdsourcing environment face with many challenges.

1. The expertise and dedication of non-expert labelers are different from one another. The statistical characteristics of these variables are not very clear, which bring difficulties in

modeling. All these make a ground truth inference algorithm perform inconsistently on different data sets. Moreover, there are still no effective ways to conduct model selection.

2. Crowdsourcing systems have an open nature, i.e., not all historical information of participants can be obtained. It is impossible to make decisions based on the information other than collected multiple noisy labels. That is, ground truth inference algorithms must be agnostic, which do not rely on any other additional historical or supervised information.
3. The quality control in crowdsourcing is difficult, especially some systems provide somewhat privacy protection (Downs et al. 2010; Ross et al. 2010), which makes it even impossible to improve label quality only during the data collection. That is why improving quality after collection is so important.
4. The phenomenon of *biased labeling* is widely spread (Wauthier and Jordan 2011; Faltings et al. 2014). Bias is different from an individual error, which is a common tendency of a large number of labels. It is caused not only by the differences between non-expert labelers and experts in their expertise and their individual preference, but also by the different scales of measures when making decisions. The detection and the modeling of bias are very difficult, and bias has a negative effect on inference algorithms and training models.
5. Both spam and adversarial labelers may exist (Ipeirotis et al. 2010; Wang et al. 2014). Spammers usually provide fixed labels to most objects, while adversarial labelers intend to provide error labels even though they know right ones. These two kinds of labelers are very harmful to the quality of collected data. Some studies (Wang et al. 2014; Li et al. 2014; Tong et al. 2014; Raykar and Yu 2012) are committed to detect and clean out the results provided by the ”malicious” labelers. But under the agnostic environment, detections are often inefficient and conservative. Even though, a large number of non-malicious results could be washed out.

3 Ground truth inference

Since crowdsourcing is a open labor force market place, the quality of work must be a great challenge. Some restrictions can be affiliated to a task which ensures that labelers must comply with certain characteristics such as total working time, the proportion of approvals of their work, etc. These stipulations are setup with the hope that annotators would complete their assigned tasks seriously. However, due to some other factors such as lack of expertise, misunderstanding directions, preferences and so on, human error, bias, even sabotage cannot be completely avoided. We cannot rely on a single labeler’s decision, but infer the true labels of instances through multiple noisy labelers, which is one of effective ways to improve the quality of labels, requiring no additional infrastructure improvement to prevalent crowdsourcing platforms. Lin et al. (2014) pointed out relabeling provides the highest benefit on the domains with a large number of features (e.g., image, video).

Ground truth inference is defined as a process of estimating the true label of each example from its multiple label set. If we only focus on the label itself, it is also called label integration. But the meanings of the former are more wide, because a lot of inference algorithms not only estimate the label of examples, but also estimate other parameters, such as the levels of expertise knowledge of labelers, the difficulty of examples, and so on. In this section, we first describe the problem of the ground truth inference. Then, we focus on expectation maximization based ground truth inference algorithms. Finally, some other algorithms are briefly discussed.

3.1 Problem definition

For a crowdsourcing system, the sample set is defined as $E = \{e_i\}_{i=1}^{I}$, where each example is $e_i = <\mathbf{x}_i, y_i>$, $\mathbf{x}_i$ is the feature vector, and y_i is the true label. The labeler set is defined as $U = \{u_j\}_{j=1}^{J}$. Each label is an element of the class set $C = \{c_k\}_{k=1}^{K}$. For simplicity, labelers, examples and classes can be denoted by their indexes, saying that example i is labeled k by labeler j. For a binary labeling, we respectively map c_1 (i.e., $k = 1$) and c_2 (i.e., $k = 2$) to the negative and the positive classes. For example i, it associates a multiple noisy label set $\mathbf{l}_i = \{l_{ij}\}_{j=1}^{J}$, where element l_{ij} comes from labeler j, and $l_{ij} \in \{0, c_1, c_2, \ldots, c_k\}$, where 0 means that the corresponding labeler does not provide any label. All multiple noisy label sets in the data set form a matrix $L = \{\mathbf{l}_i\}_{i=i}^{I}$. For each labeler j, there is a matrix $T^{(j)} = \{t_{ik}^{(j)}\}$, where $1 \leq i \leq I$ and $1 \leq k \leq K$. Each element $t_{ik}^{(j)}$ denotes the count number that labeler j provides the label k to example i. In practice, for the sake of cost and consistency, it is not usual to allow multiple labeling by the same labeler to an example, i.e., $t_{ik}^{(j)} \in \{0, 1\}$. In addition, for a binary labeling, we define the prior probabilities of the negative and positive classes of the data set as p^- and p^+.

The ground truth inference is to obtain an integrated label $\hat{y}_i$ to each sample i as its estimated true label, and to minimize the empirical risk

$$R_{emp} = \frac{1}{I} \sum_{i=1}^{I} \mathbb{I}\left(\hat{y}_i \neq y_i\right) \tag{1}$$

given the whole noisy label matrix L, where $\mathbb{I}$ is an indicator function whose output will be 1 if the test condition satisfies. Otherwise, its output will be 0.

3.2 General ground truth inference algorithm: an overview

A general ground truth inference algorithm satisfies following two constraints. It at least infers integrated labels for examples; and it does not depend on any additional information (such as the historical labeling qualities, true labels of some examples, the features of examples, etc.) other than observed multiple noisy labels. Majority vote (MV) is a simple but an effective method. For a binary case, as long as more than a half labelers provide negative labels, the integrated label will be negative, and vice versa. If the numbers of both kinds of labels are the same, the integrated label will be randomly determined. Both Sheng et al. (2008) and Ipeirotis et al. (2014) carefully studied the algorithm MV and proposed a simple probabilistic model to describe the label quality of a single example sample. This model assumes that each labeler has the same labeling quality p. If each example obtains $2N+1$ labels, the probability of the integrated label to be true obeys a binomial distribution

$$q = \sum_{i=0}^{N} \binom{2N+1}{i} p^{2N+1-i} (1-p)^{i}. \tag{2}$$

Since labelers obey the same probabilistic model, when the number of noisy labels for each example increases, the accuracies of integrated labels are quickly improved. Thus, this model provides an upper boundary of the performance of inference algorithms. For a multi-class labeling case, this algorithm is call plurality vote (Parhami 1994). In this paper, we do not distinguish the differences between MV and plurality vote, and use MV for simplicity. Jung and Lease (2011) proposed a method to improve the accuracy of MV using z-score and weights for examples.

Besides MV, quite a few novel general ground truth inference algorithms have been proposed in recent years. Table 1 summarizes these algorithms and makes comparisons from different aspects. These algorithms can be classified into two categories according to their mathematical methodologies: machine learning-based and linear algebra-based methods. Machine learning based methods are the main stream of current researches, where inference models are usually based on the probabilistic graphical models (Koller and Friedman 2009). One important mainstream general method that conducts inference in probabilistic graphical models is expectation-maximization (EM) algorithm (Watanabe and Yamaguchi 2003) which is at present widely used in the estimation of latent variables (Li et al. 2015; Liu et al. 2015). Since the majority of ground truth inference algorithms are based on EM, we focus on this kind of algorithms in the next section. From the label type viewpoint, most inference algorithms can be applied to binary labeling, and only a few are suitable for multi-class labeling. There are two studies related to ordinal labels. As Fig. 1 shows, in most cases inference and model learning are loosely coupled. However, there exist a small amount of methods, where inference and model learning are closely coupled. That is, during the process of label inference, the learning model is simultaneously trained. Such methods usually need to utilize the features of the examples. If the features cannot be obtained, the algorithms degrade into pure inference algorithm. For all general ground truth inference algorithms, although they model different aspects of a crowdsourcing system, a learning model can be trained after inference as long as the features of examples can be acquired.

Crowdsourcing is based on two basic common assumptions: labelers have different reliabilities and independently make decisions. Some inference algorithms only follow these assumptions, such as MV, DS (Dawid and Skene 1979), and ZenCrowd (Demartini et al. 2012). However, other ones, such as RY (Raykar et al. 2010), IEThresh (Donmez et al. 2010), PLAT (Zhang et al. 2015c), and LC-ME (Tian and Zhu 2015), are designed based on some other additional assumptions. Assumptions of an inference algorithm determine on which conditions it reaches its good performance. For example, RY (Raykar et al. 2010) assumes that labelers have biases toward negative and positive examples. If real-world situations follow this assumption, RY performs well. Otherwise, it may perform not so well. We list the special assumptions of reviewed ground truth inference inference algorithms in Table 2.

3.3 Inference based on EM

Early in 1979, Dawid and Skene proposed a ground truth inference algorithm DS based on maximum likelihood estimation (Dawid and Skene 1979). In addition to infer the integrated label for each example, DS also estimate a confusion matrix for each labeler. The element $\pi_{kl}^{(j)}$ in the confusion matrix of labeler j denotes the probability that the labeler provides the label l to the example with the true label k. In E-step, DS estimates the probability that example i belongs to the class k as

$$P(\hat{y}_i = c_k|L) = \frac{\prod_{j=1}^{J}\prod_{l=1}^{K}\left(\pi_{kl}^{(j)}\right)^{t_{il}^{j}} P(c_k)}{\sum_{q=1}^{K}\prod_{j=1}^{J}\prod_{l=1}^{K}\left(\pi_{ql}^{(j)}\right)^{t_{il}^{j}} P(c_q)}. \tag{3}$$

In M-step, DS updates the confusion matrix of each labeler and the prior probability of each class as follows.

Table 1 Comparisons among general ground truth inference algorithms

Algorithm	Main function	Label type	Methodology	Reliability of labeler	Difficulty of example	Other comments
MV Sheng et al. (2008)	Inference	Binary	Statistics	√		Majority voting
DS David and Skene (1979)	Inference	Multiclass	EM	√		Model confusion matrices of labelers
GLAD Whitehill et al. (2009)	Inference	Binary	EM	√	√	
IEThresh Donmez et al. (2009)	Inference	Binary	Statistics	√	√	Interval estimation learning
Welinder et al. (2010)	Inference	Binary	EM	√	√	Model competence and bias of labelers
Welinder and Perona (2010)	Inference	Multiclass/ordinal	EM	√	√	Optimize the number of labels
RY Raykar et al. (2009, 2010)	Inference + Learning	Binary	EM	√		Model sensitivity & Specificity of labelers
Weighted voting Jung and Lease (2011)	Inference	Multiclass	Statistics	√		Filter unreliable labelers
KOS Karger et al. (2011, 2014)	Inference	Binary	Algebra			Minimum the cost; belief propagation

Table 2 Special assumptions of reviewed ground truth inference algorithms

Algorithm	Assumptions
GLAD Whitehill et al. (2009)	Different levels of expertise of labelers
	Different levels of difficulty of examples
IEThresh Donmez et al. (2009)	A learner should acquire labeler and label knowledge through repeated trials, balancing the exploration versus exploitation tradeoff by first favoring the former and moving gradually to increase exploitation
Welinder et al. (2010)	Each image has different characteristics that are represented in an abstract Euclidean space
Welinder and Perona (2010)	For a label provided by an expert annotator, we can probably rely on it
	For labels provided by unreliable annotators, we should probably ask for more labels until we find an expert or until we have enough labels from non-experts to let the majority decide the label
RY Raykar et al. (2009, 2010)	Labelers have biases towards negative and/or positive examples
KOS Karger et al. (2011, 2014)	Using a simple model to capture the presence of spammers, which is called the spammer-hammer model
Yan et al. (2010c, 2011)	The probability of a labeler providing true labels obeys a Bernoulli or Gaussian distributions
Ghosh et al. (2011)	If a labeler has the probability of providing correct answers greater than 1/2, the labeler can be identified as a trustworthy labeler
Liu et al. (2012)	Different levels of ability of labelers
	The same level of difficulty of examples
Kajino and Kashima (2012)	Introducing a personal classifier for each of workers, and estimating the base classifier by relating it to the personal models
Zhou et al. (2012)	When many items are simultaneously labeled, the performance of a worker is consistent across different items
PLAT Zhang et al. (2015a)	Labelers have different correction rates on negative and positive examples
Kurve et al. (2015)	A stochastic model for answer generation that plausibly captures interplay between worker skills, intentions, and task difficulties
LC-ME Tian and Zhu (2015)	Each item belongs to one latent class, and labelers have a consistent view on items of the same class but inconsistent views on items of different classes
	Several different latent classes consist in one label category
GTIC Zhang et al. (2015a, b, c, d)	Labelers' biases from one class towards another show some consistency in a multi-class categorization

$$\hat{\pi}_{kl}^{(j)} = \sum_{i=1}^{I} \mathbb{I}(\hat{y}_i = c_k) t_{il}^{(j)} \Bigg/ \sum_{l=1}^{K} \sum_{i=1}^{I} \mathbb{I}(\hat{y}_i = c_k) t_{il}^{(j)} \tag{4}$$

$$\hat{P}(c_k) = \frac{1}{I} \sum_{i=1}^{I} \mathbb{I}(\hat{y}_i = c_k) \tag{5}$$

Although DS has been widely used in (Snow et al. 2008; Smyth et al. 1994, 1995; Ipeirotis et al. 2010; Rodrigues et al. 2013; Eagle 2009), it has a defect: if the number of classes is large and the number of labels is relatively small, the estimated confusion matrix is very sparse, which results in inaccurate estimation. Recently, DS has been considered to be a good theoretical model to analyze the error rate bounds of label aggregation. Li and Yu (2014) derived error rate bounds of a general type of aggregation rules with any finite number of workers and items under the DS model, which can be used for designing optimal weighted majority voting. Khetan and Oh (2016) also analyzed the reliability of crowdsourcing under the generalized DS model.

Focusing on binary labeling, Raykar et al. (2009, 2010) proposed a Bayesian estimation based algorithm RY, which reduces the number of parameters, compared with DS. It only concerns two parameters *sensitivity* and *specificity*. In their model, sensitivity parameter α_j denotes the labeler's bias towards the positive and the specificity parameter β_j denotes the bias towards the negative. The two parameters and the prior probability of the positive class obey a Beta distribution (Gelman et al. 2014).

$$P(\alpha_j|a_j^+, a_j^-) = Beta(\alpha_j|a_j^+, a_j^-) \tag{6}$$

$$P(\beta_j|b_j^+, b_j^-) = Beta(\beta_j|b_j^+, b_j^-) \tag{7}$$

$$P(p^+|n^+, n^-) = Beta(p^+|n^+, n^-) \tag{8}$$

Parameters a_j^+ (b_j^+) and a_j^- (b_j^-) are the numbers of the positive and the negative labels respectively, provided by labeler j to those examples currently estimated as positive (negative). Parameters n^+ and n^- are the total numbers of the positive and the negative labels respectively, provided by all labelers. RY estimates the probability that example i belongs to the positive in its E-step as $P(y_i = +|x_i, L, \alpha, \beta, p^+)$. In M-step, RY updates the parameters (sensitivity, specificity) and the prior probability of the positive class. RY shows some advantages over DS for a binary labeling. However, for a multi-class labeling, no empirical investigation has been reported.

Both algorithms above do not take the difficulties of labeling examples into account. Some studies (Welinder et al. 2010; Brew et al. 2010) have shown that considering the difficulties of examples is useful. The algorithm GLAD (Whitehill et al. 2009) models both the levels of users' expertise and the difficulties of examples. GLAD uses the parameter $1/\beta_i \in [0, +\infty)$ to model the difficulty of example i and the parameter $\alpha_j \in (-\infty, +\infty)$ to model the level of the expertise of labeler j. The larger the $1/\beta_i$, the harder an example can be labeled; and the lager the α_j, the more professional a labeler would be. $\alpha_j < 0$ suggests that labeler j is adversarial. GLAD adopts a logistic regression model as follows.

$$P(l_{ij} = y_i|\alpha_j, \beta_i) = \sigma(\alpha_j\beta_i) = \frac{1}{1 + e^{-\alpha_j\beta_i}} \tag{9}$$

In E-step, based on the two parameters ($\boldsymbol{\alpha}, \boldsymbol{\beta}$) and all observed labels, GLAD calculates the posterior probabilities of being positive and being negative for all examples as follows.

$$P(y_i|L, \alpha, \beta) \propto P(y_i)\prod_{j=1}^{J} p(l_{ij}|y_i, \alpha_j, \beta_i) \tag{10}$$

In M-step, GLAD uses a gradient descent algorithm to maximize a standard auxiliary function Q and updates the two parameters ($\boldsymbol{\alpha}, \boldsymbol{\beta}$), where function Q is defined as follows.

$$Q(\boldsymbol{\alpha}, \boldsymbol{\beta}) = \sum_i E[\ln P(y_i)] + \sum_{ij} E[\ln P(l_{ij}|y_i, \alpha_j, \beta_i)] \tag{11}$$

Welinder et al. (2010) proposed a more complex multidimensional model. In their model, parameter $z_i \in \{0, 1\}$ represents a two-value judgment to an example. When labeler j judges example i, the decision is not directly made by evaluating $\mathbf{x}_i$, but is made based on the evaluation of $\mathbf{r}_{ij} = \mathbf{x}_i + \mathbf{n}_{ij}$, where $\mathbf{n}_{ij}$ is a labeler-specific and instance-specific noise. These noises vary with different labelers, each of which is parameterized as $\boldsymbol{\sigma}_j$. Each labeler is parameterized as a vector $\hat{\mathbf{w}}_j$, which encodes the expertise of a labeler to a high dimensional space. The scalar project of $<\mathbf{r}_{ij}, \hat{\mathbf{w}}_j>$ is compared with a threshold $\hat{\tau}_j$. If the former is larger, then $l_{ij} = 1$. Otherwise, $l_{ij} = 0$. The inference of the method is based on a complicated model as follows.

$$p(L, \mathbf{z}, \mathbf{x}, \mathbf{r}, \boldsymbol{\sigma}, \hat{\mathbf{w}}, \hat{\boldsymbol{\tau}}) = \prod_{j=1}^{J} p(\boldsymbol{\sigma}_j) p(\hat{\tau}_j) p(\hat{\mathbf{w}}_j)$$
$$\prod_{i=1}^{I} (p(z_i) p(\mathbf{x}_i | z_i) \prod_{j \in \mathbf{l}_i} p(\mathbf{r}_{ij} | \mathbf{x}_i, \boldsymbol{\sigma}_j) p(l_{ij} | \hat{\mathbf{w}}_j, \hat{\tau}_j, \mathbf{r}_{ij})) \tag{12}$$

Demartini et al. (2012) found that RY, GLAD, and other algorithms (Welinder et al. 2010) attempted to model the system from different aspects. However, due to the sparsity of the sample, the accuracy of the inference faces with some risks. The simplified model may perform better than a complex one under the sparsity of data. Thus, they proposed algorithm ZenCrowd, which uses only a two-element *{good, bad}* parameter to model the reliability of a labeler. The advantage of using one parameter is that it avoids the problem of the large estimation deviation of the variables when applying inference on the data set with sparsity. ZenCrowd is slightly more complex than MV, but simpler than other EM based algorithms. If no prior information is provided, it starts with $P(u_j = reliable) = 0.5$ based on maximum entropy principle. In E-step, ZenCrowd calculates the probability of each example belonging to a particular class as follows.

$$P(y_i = c_k) = \frac{\prod_{j=1}^{J} [P(u_j = reliable)]^{\mathbb{I}(\hat{y}_i = c_k)}}{\sum_{k=1}^{K} \prod_{j=1}^{J} [P(u_j = reliable)]^{\mathbb{I}(\hat{y}_i = c_k)}} \tag{13}$$

In M-step, since the label of each example is estimated as the class with the maximum probability, ZenCrowd uses these estimated labels to update the reliability of each labeler.

$$P(u_j = reliable) = \frac{\sum_{i=1}^{I} \mathbb{I}(l_{ij} = \hat{y}_i)}{\sum_{k=1}^{K} \sum_{i=1}^{I} t_{ik}^{(j)}} \tag{14}$$

Although the above algorithms respectively model the ability, the bias of labelers, and the difficulties of examples, they cannot tell whether a labeler is a spammer or adversarial. For example, some labelers do have good abilities, but for some reason, they provide opposite labels. Kurve et al. (2015) added the *intention* of labelers into the model for inference, where labelers are classified into two categories: *honesty* and *adversarial*. The parameters of their model are as follows: $\tilde{d}_i \in (-\infty, +\infty)$ represents the difficulty of example i; labeler's intention is denoted by $v_j \in [0, 1]$, where 0 means *adversarial* and 1 means *honesty*; $d_j \in (-\infty, +\infty)$ represents the expertise of labeler j; and the additional α_j denotes the tendency of labeler j providing correct labels given the difference $d_j - \tilde{d}_i$. The parameter set of their model is denoted by $\Lambda = \{\{(v_j, d_j, \alpha_j)\forall j\}, \tilde{d}_i \forall i\}$. Supposed the iteration count of the algorithm is t, it calculates the labels of examples and the parameters in E-step as follows.

$$P_i(y_i = k|E, \Lambda^t) = \frac{\prod_{j=1}^{J} \beta(l_{ij}|\Lambda_{ij}^t, y_i = k)}{\sum_{c=1}^{K} \prod_{j=1}^{J} \beta(l_{ij}|\Lambda_{ij}^t, y_i = c)} \tag{15}$$

In M-step, it optimizes the expectation of the log likelihood of the whole data set with respect to the parameters. The optimal parameters will be used in the next E-step.

$$\Lambda^{t+1} = \arg\max_{\Lambda} \mathbb{E}[\log L_c(\Lambda)|E, \Lambda] \tag{16}$$

The minimax entropy principle can be used for estimating the true labels from the judgements of a crowd of nonexperts. Zhou et al. (2012) combines the ideas of MV and confusion matrix. Their work is based on an assumption that labels are generated by a probability distribution over workers, items, and labels. By maximizing the entropy of this distribution, this method naturally infers item confusability and worker expertise. By minimizing this entropy, the method infers the integrated labels. Let p_{ijd} be the probability of labeler j providing a correct label to example i whose true label is d. According to MV, the true label of example i should most commonly appear in its multiple noisy label set. Thus, we have the first constraint

$$\sum_j p_{ijd} = \sum_j \mathbb{I}(l_{ij} = d), \forall (i, d). \tag{17}$$

According to confusion matrix, the same labeler should have a consistent performance on different examples belonging to the same class. Thus, we have the second constraint

$$\sum_{\substack{j \\ s.t. y_i = d}} p_{ijd} = \sum_{\substack{i \\ s.t. y_i = d}} \mathbb{I}(l_{ij} = d), \quad \forall (j, d)\,. \tag{18}$$

Finally, by applying the minimax entropy principle, we use the set of class C to minimize the entropy under above two constraints (i.e., Eqs. 19, 20), and use the set of labeler probabilities P to maximum the entropy.

$$\min_C \max_P - \sum_{i,j,d} p_{ijd} \log p_{ijd} \tag{19}$$

Tian and Zhu (2015) proposed a nonparametric model LC-ME which extends the minimax entropy estimator to learn latent structures. The LC-ME estimator is based on two assumptions: each example only belongs to one latent class and the behaviors of labelers are different when the examples that they label belong to different latent classes. Experimental results on real-world data sets have shown that LC-ME is superior to the original minimax entropy estimator.

Although the inference algorithms based on expectation maximization are widely used in the processing of crowdsourced labeled data, they are not robust on different kinds of data sets because of the inherent defects of EM (Zhang et al. 2014). First, the likelihood function of EM is non convex, which makes the EM algorithm cannot converge to the global optimal. Secondly, there is no guidance to set the initial values of the parameters in EM, and different settings result in different results on the same data set. Finally, the iteration count of EM when it converges depends on the initial settings and the data set. There is no way to guarantee that it can converge fast and efficiently. Some work attempts to overcome the defects of EM. Zhang et al. (2014) utilized a spectral method (Bernardi and Maday 1997) to estimate the initial values in the confusion matrix of each labeler before the algorithm DS starts. In their method, all the labelers are divided into three disjoint groups. The average confusion matrix of the three groups is estimated. Then, the initial state of the confusion matrix of each labeler is estimated using the average confusion matrix. Empirical study has shown that the performance of DS is improved by the spectral base initialization compared with randomly settings.

3.4 Inference based on linear algebra and statistics

Besides EM being widely used for inference, some other methods are based on linear algebra or simple statistics.

Karger et al. (2011, 2014) proposed a label integration algorithm based on the reliabilities of labelers. The algorithm uses a belief propagation-like method to achieve the goal of inference. The following algorithm KOS shows its main steps. One notable feature of this method is that when the noisy label sets L is a (l, r)-regular bipartite graph with $l = r$, it is equivalent to a singular value decomposition (SVD) of a low rank matrix. Ghosh et al. (2011) proposed a similar SVD-based method, which aims to infer the ratings of user generated contents.

Algorithm 1 KOS

```
Input:  $E, L, k_{max}$
Output:  $E$, where example $i$ is assigned $\hat{y}_i$
$\forall (i, j) \in E$ initialize $p^{(0)}_{j \to i}$ with random $z_{ij} \sim N(1, 1)$
for $k = 1, \dots, k_{max}$ do
    $\forall (i, j) \in E, s^{(k)}_{i \to j} \leftarrow \sum_{j' \notin \partial_i \setminus j} l_{ij'} p^{(k-1)}_{j' \to i}$
    $\forall (i, j) \in E, p^{(k)}_{j \to i} \leftarrow \sum_{i' \notin \partial_j \setminus i} l_{i'j} p^{(k)}_{i' \to j}$
end for
$\forall i, s_i \leftarrow \sum_{j \in \partial_i} l_{ij} p^{(k_{\max}-1)}_{j \to i}$
$\forall i, \hat{y}_i \leftarrow sign(s_i)$
return $E$
```

The algorithm KOS is not a standard belief propagation model, though it has good guarantees on (locally tree-like) random assignment graphs, but does not have an obvious interpretation as a standard inference method on a generative probabilistic model (Liu et al. 2012), which makes it difficult to either extend KOS to more complicated models or adapt it to improve its performance on real-world data sets. Liu et al. (2012) proposed a standard belief-propagation-based method based on variational inference (Wainwright and Jordan 2008) on a generative probabilistic model. Both KOS and MV are special cases under this framework. For example, if the reliabilities of labelers obey Haldane prior (Zellner 1996), the method is equivalent to KOS.

SVD based methods require that the label matrix is full, i.e., all examples are labeled by every labeler. However, in reality, this prerequisite cannot always been guaranteed. The method proposed by Dalvi et al. (2013) also take advantage of SVD, but looses the above prerequisite. The model introduce an additional matrix $G \in \{0, 1\}^{I \times J}$ to represent whether labeler j provides a label to example i. The estimated vector of the reliabilities of labelers is denoted by $\hat{\mathbf{w}} \in [-1, 1]^J$ and the estimated vector of the quality of examples is denoted by $\hat{\mathbf{q}} \in [-1, 1]^I$. The algorithm take the advantage of the (scaled) top eigenvector of a matrix defined as $\mathbf{v}_1(M) = \arg\min_{\mathbf{x}} \left\| M - \mathbf{x}\mathbf{x}^T \right\|_2, \mathbf{x}^{(1)} \geq 0$. The steps of inference are shown in the algorithm EigenRatio. The operator $\oslash$ in the algorithm is defined as follows.

$$(M \oslash N)_{ij} = \begin{cases} M_{ij}/N_{ij}, & \text{if } N_{ij} \neq 0 \\ 0, & \text{otherwise} \end{cases} \tag{20}$$

Besides the ground truth inference based on linear algebra, there exist some other methods based on statistics. Donmez et al. (2009) proposed an accuracy model IEThresh based on interval estimation (IE). In statistics, given r_j observations of variable u_j from distribution d, IE method estimates an interval that the next observation belongs to, with the probability $1 - \alpha$. IEThresh selects a labeler with the highest accuracy interval upper bound. A higher upper bound indicates a higher expected accuracy (when the interval length is short) or a higher uncertainty (when the interval length is long). The interval upper bound is estimated as follows.

Algorithm 2 EigenRatio

```
Input:   $L \in \{-1, 0, 1\}^{I \times J}$, $G \in \{0, 1\}^{I \times J}$
Output:  $\hat{\mathbf{w}}, \hat{\mathbf{q}}$
$\hat{\mathbf{w}} = \mathbf{v}_1(L^T L) \oslash \mathbf{v}_1(G^T G)$
$\tilde{w}_j = \text{sgn}(\tilde{w}_j) \max(|\hat{w}_j|, 1)$
$\hat{q}_i = \text{sgn}(\sum_j L_{ij} \tilde{w}_j)$
return $\hat{\mathbf{w}}, \hat{\mathbf{q}}$
```

$$U(u_j) = \mu(u_j) + o_{\alpha/2}^{r_j - 1} \frac{\sigma(u_j)}{\sqrt{r_j}} \tag{21}$$

where $\mu(u_j)$ and $\sigma(u_j)$ are the mean and the standard deviation of correct labels provided by labeler j respectively. $o_{\alpha/2}^{r_j - 1}$ is the value of t-student distribution when the degree of freedom is $r_j - 1$ and the level of confidence is $\alpha/2$. When the observed data is sufficient, even though high quality labels are very rare, IEThresh can lead to very good results.

The labeling bias is always one of research issues in learning from crowds. Many algorithms (Welinder et al. 2010; Raykar et al. 2009, 2010; Kurve et al. 2015) model the labeling bias. On the real-world data sets, these EM-based algorithms probably quickly converge to a local optimal, leading a poor result. Zhang et al. (2015c) analyzed the biased labeling phenomenon in real-world data sets and found that most labelers usually have the consistent tendency when biased labeling occurs. At this point, the class distribution of the labels significantly skew. Based on simple statistics and heuristic search strategies. Zhang et al. (2015c) proposed an algorithm PLAT, which counts the number of positive labels for each example, and then automatically searches a threshold to classify examples into two categories. PLAT has obvious effectiveness on solving the biased labeling problem together with the imbalanced class problem (Prati et al. 2015), and its running speed is higher than that of the EM based algorithms. Besides, it can be adaptively applied to the unbiased data sets. Bias in multi-class labeling is hard to model, which results in a poor performance for most of inference algorithms. Zhang et al. (2016) proposed a novel algorithm GTIC based on Bayesian statistics for multi-class labeling. For a K labeling case, GTIC utilizes the repeated label sets of examples to generate features. Then, it uses a K-Means algorithm to cluster all examples into K different groups, each of which is mapped to a specific class. GTIC captures the tendency of labeling biases with respect to groups of examples and shows a significant improvement on multi-class inference.

3.5 Other ground truth inference algorithms

Besides general ground truth inference algorithms, there are other types of algorithms that are used to describe different characteristics of crowdsourcing system or solve different problems.

Donmez et al. (2010) proposed a time varying algorithm SFilter for user accuracy modeling. The objective of SFilter is to filter out low-quality labelers in active systems. This algorithm is based on Sequential Bayesian Estimation, and assumes that the maximum rate of changes are small and known. Let p_j^t denote the accuracy of labeler j in time t, and $\mathbf{l}_j^t$ denote the label provided by the labeler. The posterior probability of the labeling accuracy in time t is denoted by $P(p_j^t|\mathbf{l}_j^1, \ldots, \mathbf{l}_j^t)$. SFilter uses the following Hidden Markov model for modeling the accuracy changes.

$$p_j^t = f_t\left(p_j^{t-1}, \Delta_{t-1}\right) = p_j^{t-1} + \Delta_{t-1}, \Delta \sim N(0, \sigma^2) \tag{22}$$

The model shows that the accuracy of each labeler is only determined by its accuracy in the last time. SFilter assumes that the accuracy is in a range (0.5, 1], and it calculates a transition probability from time $t-1$ to t using a truncated Gaussian distribution as follows.

$$P\left(p_j^t|p_j^{t-1}, \sigma\right) = \frac{1}{\sigma}\phi\left(\frac{p_j^t - p_j^{t-1}}{\sigma}\right)\Bigg/\left(\Phi(\frac{1-p_j^{t-1}}{\sigma}) - \Phi(\frac{0.5-p_j^{t-1}}{\sigma})\right) \tag{23}$$

where ϕ is Gaussian probability density function and Φ is its cumulative distribution function. Supposed that l_{ij}^t is the label provided by labeler j to example i and $l_{iJ(t)}^t$ are labels provided by other labelers. We have

$$P\left(l_{ij}^t|p_j^t, l_{iJ(t)}^t\right) = \sum_{y\in\{-1,1\}} P\left(l_{ij}^t|p_j^t, y_i = y)P(y_i = y|l_{iJ(t)}^t\right) \tag{24}$$

where $P(y_i|l_{iJ(t)}^t)$ is calculated by the probability of the integrated label inferred from $l_{iJ(t)}^t$ as

$$P\left(y_i|l_{iJ(t)}^t\right) \propto P(y_i)P\left(l_{iJ(t)}^t|y_i\right) \propto P(y_i)\prod_{j\in J(t)} P\left(l_{ij}^t|y_i\right). \tag{25}$$

One of the major drawbacks of SFilter is that it is time consuming. As long as one label updates, the model needs to be re-calculated. To conquer this defect, the authors provided an incremental version of the algorithm. Experimental results show that when the accuracy of labels change in a certain range, SFilter can capture these changes.

Tang and Lease (2011) proposed a semi-supervised ground truth inference algorithm based on DS. In addition to unlabeled data set D^U, the algorithm needs another data set D^L involved in inference, in which the true labels of examples are known. The difference from DS is that a term of likelihood derived from the labeled examples is added at the end of the original likelihood function of DS as follows.

$$\ell\left(p_k, \pi_{km}^{(j)}\right) = \prod_{i\in D^U}\left(\sum_{k=1}^{K} p_k \prod_{j=1}^{J}\prod_{m=1}^{K} (\pi_{km}^{(j)})^{t_{im}^{(j)}}\right) + \prod_{i\in D^L} p_k \prod_{j=1}^{J}\prod_{m=1}^{K}\left(\pi_{km}^{(j)}\right)^{t_{im}^{(j)}} \tag{26}$$

Experimental results show that only a small amount of supervised examples can greatly improve the accuracy of inference. Adding the expert labeled examples violates the agnostic feature of the general inference algorithm. Similar work includes the semi-supervised model proposed by Yan et al. (2010b), where the integrated labels and the expertise of labelers can be simultaneously inferred. Different from Tang and Lease (2011), the model relaxes the prerequisite that all examples must be labeled. That is, there are three kinds of examples: those labeled by experts, by crowdsourced labelers and unlabeled. The applicability of this semi-supervised model is obviously higher.

Another algorithm ELICE proposed by Khattak and Salleb-Aouissi (2011) also focuses on the optimization of the ground truth inference using expert labeled data, where expert labels are injected into the crowdsourced data set. ELICE introduces two parameters $\alpha_j \in [-1, 1]$ and $\beta_i \in [0, 1]$ to represent the reliability of labeler j and the difficulty of example i respectively. By injecting examples labeled by experts and analyzing the noisy labels on these examples, ELICE calculates the parameters α_j and β_i, and then infers the integrated label as follows.

$$IL_i = \frac{1}{J}\sum_{j=1}^{J}\frac{l_{ij}}{1+\exp(-\alpha_j\beta_i)} \tag{27}$$

The symbol IL_i represents the positive and negative categories. Although ELICE shows its effectiveness, some issues need to be addressed. For example, when and how many the expert labeled examples should be injected.

All of the ground truth algorithms mentioned above can be used to deal with either binary or multi-class labeling problems. Even if a few algorithms studied the inference for ordinal label (Welinder and Perona 2010), they still treat it as a special case of multi-class. Zhou et al. (2014) proposed an inference model specially for ordinal labels, which is based on the former minimax entropy principle (Zhou et al. 2012). The model extends the original item confusion vector to a structural item confusion matrix. By introducing structural sequential relationships, the model reduces the number of parameters during the inference, which is beneficial to distinguish adjacent classes.

4 Learning model

Compared with the ground truth inference, studies of building the learning model are still in a young stage. First, the label quality is the basis of a learning model quality. After achieving a relative good quality of integrated labels, the model training can become an independent process. That is why so many researches focus on improving the performance of ground truth inference. Second, the study of the learning model is only one aspect of the studies of crowdsourced labeling, but almost all of follow-up studies, such as information retrieval, must concern the improvement of the inference. Finally, the performance of a learning model is closely related to the features of examples and the selection of classification algorithms, so it is more difficult to propose general methods that work well under most conditions. In this section, we summarize the methods of building the learning model from the crowdsourced labeled data in recent years.

4.1 Ordinary supervised learning

It is straightforward to build a learning model on a data set with integrated labels. For example, RY introduces the logistic regression model

$$P(y=1|\mathbf{x},\mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}} \tag{28}$$

to build a classifier which is used for predicting unlabeled examples. This method provides a general model to conduct binary classification.

A simpler method is proposed to build a learning model after inference with MV. Sheng (2011) proposed an enhanced algorithm *MVBeta*, which considers the distribution of two classes of noisy labels on each examples. The method provides a weight to each example after conducting MV, which is defined as follows.

$$
\begin{aligned}
W &= 1 - \min\{\mathcal{I}_{0.5}(\alpha, \beta), 1 - \mathcal{I}_{0.5}(\alpha, \beta)\} \\
\alpha &= L_p + 1, \beta = L_n + 1
\end{aligned}
\tag{29}
$$

where L_p and L_n are the numbers of positive and negative labels respectively in the multiple noisy label set of an example. The function $\mathcal{I}$ is the cumulative distribution function of Beta distribution, which is defined as

$$
\mathcal{I}_x(\alpha, \beta) = \sum_{j=\alpha}^{\alpha+\beta-1} \frac{(\alpha + \beta - 1)!}{j!(\alpha + \beta - 1 - j)!} x^j (1 - x)^{\alpha+\beta-1-j}. \tag{30}
$$

These weights can be used by different classification algorithms in different forms. For example, a cost sensitive tree classifier (Ting 2002) can directly utilize these weights. Similarly, Naive Bayes classifier also can handle this kind of label with a numeric uncertainty. Experimental results show that assigning weights to examples can significantly improve the accuracy of the classifier in the case of high noise.

In addition to building learning models after inference, there exist some methods that directly build learning models without processing inference first. Sheng (2011) studied a more general method than *MVBeta* to build the learning model without inference, which is named *PairwiseBeta*. Supposed an example has a multiple noisy label set $\{+++----\}$, *PairwiseBeta* does not infer the integrated label for the example, but duplicates this example into two replica $\{(+, WP)\}$ and $\{(-, WN)\}$. Each replica has a corresponding weight. The weights of these two replica are defined as follows.

$$
\begin{aligned}
W_P &= \mathcal{I}_{0.5}(L_n + 1, L_p + 1) \\
W_N &= 1 - \mathcal{I}_{0.5}(L_n + 1, L_p + 1)\}
\end{aligned}
\tag{31}
$$

After example duplication, all examples with their weights are processed by a cost-sensitive classifier to train a learning model. Since this method preserves the information of multiple noisy label sets at a maximum extent without any bias, it is superior to MV and MVBeta in most cases.

As mentioned in Sect. 3, the mainstream methods of ground truth inference are based on the EM algorithm. Due to the inherent defects of EM, inference easily converges into a local optimal, which directly leads to a poor learning model. Kajino and Kashima (2012) proposed a method, which directly builds learning models without the prepositive inference. The objective of their method is to construct a convex function, and then to build a logistic regression model by convex optimization. The method first introduces a base classifier as follows.

$$
P[y = 1|\mathbf{x}, \mathbf{w}] = \sigma\left(\mathbf{w}_0^T \mathbf{x}\right) = \left(1 + \exp(-\mathbf{w}_0^T \mathbf{x})\right)^{-1} \tag{32}
$$

Then, each labeler is treated as an independent classifier with a parameter $\mathbf{w}_j$.

$$
P[y_j = 1|\mathbf{x}, \mathbf{w}_j] = \sigma\left(\mathbf{w}_j^T \mathbf{x}\right) \tag{33}
$$

The parameter of each independent classifier can be viewed as the parameter of the base classifier plus a deviation vector.

$$
\mathbf{w}_j = \mathbf{w}_0 + \mathbf{v}_j \tag{34}
$$

If each labeler is treated as a classifier to handle a batch of tasks, the model is similar with the multi-task learning proposed by Evgeniou and Pontil (2004). Given $W = \{\mathbf{w}_j\}_{j=1}^J$, its target convex functions is

$$F(\mathbf{w}_0, W) = -\sum_{j=1}^{J}\sum_{i\in I_j}\delta\left(l_{ij}, \sigma(\mathbf{w}_j^T\mathbf{x}_i)\right) + \frac{\lambda}{2}\sum_{j=1}^{J}\left\|\mathbf{w}_j - \mathbf{w}_0\right\|^2 + \frac{\eta}{2}\|\mathbf{w}_0\|^2 \quad (35)$$

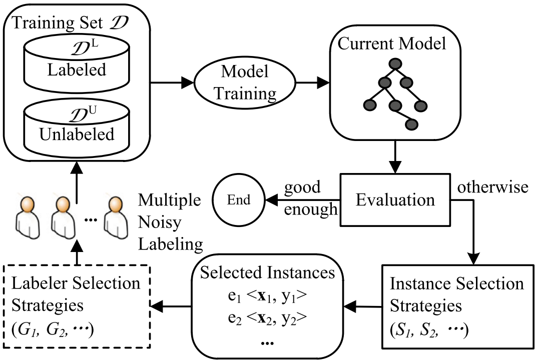
where $\delta(s, t) = s\log t + (1 - s)\log(1 - t)$, I_j represents all examples labeled by labeler j, and both λ and η are positive constants. The model finally optimizes the target function $F(\mathbf{w}_0, W)$ with respect to the parameters $\mathbf{w}_0$ and W. This method theoretically obtains a global optimal solution, and their experimental results shows that it is better than RY.

Learning with crowdsourced labeled data is implicitly related to *learning with noisy labels* (Natarajan et al. 2013). In traditional supervised learning, when training data are polluted by label noises, obvious solutions are to cleanse the training data or to design label noise-tolerant models. Designing label noise-tolerant models, such as (Rätsch et al. 2000; Freund 2001; Sukhbaatar et al. 2014), are not easy. Experimental results in the literature show that the performance of classifiers trained with label noise-tolerant algorithms is still affected by label noises. They seem to be adequate only for simple cases that can be safely managed by overfitting avoidance (Frénay and Verleysen 2014). Label noise cleansing methods, such as model prediction-based filtering (Brodley and Friedl 1999) and label noisy correction (Nicholson et al. 2015), are easy to understand and implement. However, some evidences have shown that precisely distinguishing mislabeled instances from correct ones may be rather difficult (Weiss and Hirsh 1998). Frénay and Verleysen (2014) pointed out that in most cases simply removing mislabeled instances is more efficient than correcting them. However, in the most cases of crowdsourcing tasks, due to the limitation of budgets, removing a large number of potential examples with noisy labels is not acceptable. Some information obtained in an inference stage may help us precisely identify potential noisy labels and make a better correction (Zhang et al. 2015b). Noise correction in crowdsourcing is worthy of being studied in the future.

4.2 Active learning

In many real-world applications, it is necessary to construct a learning model. When there is no enough data available, we need to actively acquire extra data from the oracle. This is active learning (Settles 2010). When the extra data acquisition tasks are outsourced to the crowd, it is called crowdsourcing. From the perspective of model learning, not all instances necessarily need to be labeled and not all labeled instances necessarily have the same number of repeated labels. Active learning allows to dynamically require labels, which is suitable for the open and dynamic environment of crowdsourcing. One of the most important features of active learning is to reduce the total cost of acquiring labels in the premise of keeping the performance of learning model (Settles and Craven 2008). The core of traditional active learning is to design instance selection strategies (Fu et al. 2013). In the crowdsourcing environment, in addition to the instance selection, labeler selection is optionally included. Labeler selection strategies aim at selecting the next labeler to provide labels who is the most beneficial to the improvement of the current model quality. In an active learning paradigm, instance selection is compulsory. Figure 2 shows a general active learning framework for learning with crowdsourced labeled data. In the framework, examples are both labeled and unlabeled. Selected examples are labeled by multiple selected labelers. After inferring the integrated labels of the labeled examples, a learning model is trained and evaluated. If the learning model cannot be further improved or satisfies a preset condition, the iteration of active learning ceases.

Fig. 2 A general framework for active learning from crowdsourced labeled data



Sheng et al. (2008) first investigated the instance selection strategies in crowdsourcing. (1) The simplest strategy in crowdsourcing is to select the example whose multiple noisy label set contains a minimum number of noisy labels. This ensures that each example has the equal opportunity to get labels. Obviously, this simple strategy wastes budget and has poor performance. (2) *Instance selection strategies can be designed based on heterogeneity of labels*. One of these strategies can use information entropy to measure the heterogeneity of examples. Supposed that in the multiple noisy label set of example $\mathbf{x}_i$ the proportion of the majority class is p_i. This strategy can select example l that satisfies the following condition.

$$l = \arg\max_i \{-p_i \log(p_i) - (1-p_i)\log(1-p_i)\} \tag{36}$$

That is, this strategy selects those examples, in whose multiple noisy label sets the proportions of the major classes are close to 0.5. However this strategy could cause that a small portion of examples are always selected again and again, and obtain enormous labels, while other examples are probably neglected, because it does not concern the numbers of labels obtained for each example. (3) *Instance selection strategies can be designed based on label uncertainty*. They proposed such a strategy, which only considers the class distribution of labels in the multiple noisy label set of each example. It selects the example with the maximum label uncertainty to query extra labels. The label uncertainty measure is defined as follows.

$$U_L = \min\{\mathcal{I}_{0.5}(L_p + 1, L_n + 1), 1 - \mathcal{I}_{0.5}(L_p + 1, L_n + 1)\} \tag{37}$$

where L_p and L_n are defined as the same as Eq. (29). (4) *Instance selection strategies can be designed based on model uncertainty*. They proposed such a strategy, which only considers uncertainties of examples to the current learning model, but totally ignores the class distribution of labels. This strategy adopts ensemble learning to build m independent weak classifiers, which work together to determine the probabilities of the classes of examples. The probability of an example being classified as positive is calculated as follows.

$$S_p = \frac{1}{m}\sum_{j=1}^{m} P(+|\mathbf{x}_i, H_j) \tag{38}$$

where $P(+|\mathbf{x}_i, H_j)$ is the probability of example $\mathbf{x}_i$ being classified as positive by model H_j. The learning model can be trained using Random Forest (Breiman 2001). The uncertainty measure of an example is defined as the distance between 0.5 and the estimated probability of being positive or negative.

$$U_M = 0.5 - \left|S_p - 0.5\right| \tag{39}$$

(5) *Instance selection strategies can be designed based on hybrid uncertainty*. They proposed such a strategy, which combines two strategies above and the uncertainty measure of each example is defined as follows.

$$U_{LM} = \sqrt{U_L \cdot U_M} \tag{40}$$

Experimental results show that their hybrid strategy, which combines two mutual complementary uncertainties, has the best performance in most cases. Zhang et al. (2015d) extended these strategies by taking the bias information into account, which makes them to handle biased labeling issues.

Long et al. (2013) proposed a novel instance selection method based on entropy, where the uncertainty measure of an unlabeled example $\mathbf{x}_u$ is defined as follows.

$$H(y_u) = -\sum_{y_u \in \{+,-\}} p\left(y_u|\mathbf{x}_u, D^L\right) \log p\left(y_u|\mathbf{x}_u, D^L\right) \tag{41}$$

where $p(y_u|\mathbf{x}_u, D^L)$ can be solved by an expectation propagation method (Williams and Barber 1998).

Currently, the studies of instance selection in crowdsourcing learning are not sufficient. Novel instance selection strategies are expected. When we design a new instance selection strategy, we must pay attention to the following points: (1) the status of the multiple noisy label set of each example must be taken into account; (2) we should estimate the performance changes of the model (better maximize the model performance) when the selected instances obtain additional labels; (3) we should avoid local optimals.

Yan et al. (2011) proposed an active learning framework based on the logistic regression model proposed in their previous study (Yan et al. 2010c). In this framework, labeler selection is involved as well as instance selection. Each labeler is viewed as logistic regression classifier. The logistic regression function of labeler j is defined as follows.

$$\sigma_j(\mathbf{x}_i) = \left(1 + \exp(-\mathbf{w}_j^T \mathbf{x}_i - \gamma_j)\right)^{-1} \tag{42}$$

The overall final logistic regression classifier is defined as

$$p(y_i = 1|\mathbf{x}_i) = \left(1 + \exp(-\boldsymbol{\alpha}^T \mathbf{x}_i - \boldsymbol{\beta})\right)^{-1}. \tag{43}$$

The instance selection procedure chooses the example with the probability $P(y_i = 1|\mathbf{x}_i)$ mostly close to 0.5. After the example is selected, a labeler selection procedure chooses labeler j with the minimum uncertainty to label the selected example.

$$j* = \arg\min_j \sigma_j(\mathbf{x}_i), \forall j \tag{44}$$

Due to selecting confident labelers to provide labels, the performance of this method is better than that of the methods only involving instance selection.

One defect of the above labeler selection method is that those who are reliable are repeatedly selected, and ordinary labelers are lack of opportunities. In order to select labelers in a wider range and also improve the ability of ordinary labelers, Fang et al. (2012) proposed a self-taught active learning method. The self-taught procedure in the method is essentially using the labels provided by weak labelers to extend the multiple noisy label sets of examples. The main steps are: (1) select an example $\mathbf{x}^*$ to query by some instance selection strategy; (2) select a reliable labeler j^* by some labeler selection strategy; (3) labeler j^* labels example $\mathbf{x}^*$ with label l_{xj}; (4) select the most unreliable labeler w; and (5) Given $\mathbf{I}^w$ is the label set provided by labeler w, let $\mathbf{I}^w \leftarrow \mathbf{I}^w \cup \{\mathbf{x}^*, l_{xj}\}$. The instance and labeler selection strategies used are the same as those proposed by Yan et al. (2011). By adding new reliable labels into the label sets of weak labelers, reliable knowledge is learned by weak labelers. Although it is a very simple method, the experimental results show the obvious improvement of the model quality during active learning. Rodrigues et al. (2014) proposed an active learning framework based on Gaussian process, in which different levels of expertise of labelers are model, instances and labelers are also selected according to their uncertainty and reliability.

For a classification problem, labelers must make choices when providing answers. However, the confidences of labelers are different from one another. Zhong et al. (2015) introduce the labeling confidence into active learning. For a binary labeling, the class set of labels is $\{-1, 0, 1\}$, where 0 indicates that a labeler is unsure about label to provide. Having completed the ith query, a classifier $f_i(\mathbf{x})$ is built. Let function $g_j(\mathbf{x})$ be the reliability of labeler j. $g_j(\mathbf{x}) > 0$ means that the labeler provides a correct label to example $\mathbf{x}$. Define function $\Omega(\mathbf{x}) = \vee_{j \in J} \delta(g_j(\mathbf{x}))$, where δ is a symbolic function. The instance and labeler selection strategies are respectively presented by Eqs. (45) and (46).

$$\mathbf{x}_{i+1} = \arg\max_{\mathbf{x} \in D_i^U} (H(\mathbf{x}|D_i^L, f_i, \theta_i) \cdot \Omega(\mathbf{x})) \tag{45}$$

$$j_{i+1} = \arg\max_{j \in J} g_t(\mathbf{x}_{i+1}) \tag{46}$$

In order to embody the strategies, SVM with Radial Basis Function kernel is applied to build a classifier. Experimental results on real-world data sets show the performance of this method is better than that of the baseline and PMActive proposed by Wu et al. (2013). It is reasonable that the confidence measures of the labelers towards their provided labels are introduced in the learning model. However, how to extend the current binary value confidence measure to a real value measure is worthy of further studies.

4.3 Other learning model

Learning from crowdsourced labeled data is a new research field. Many research issues in traditional machine learning are facing new challenges. For example, researchers have been aware of the potential value of recent widely studied transfer learning (Pan and Yang 2010; Oyen and Lane 2015) to crowdsourcing. Mo et al. (2013) first introduced the techniques of transfer learning in the area of crowdsourcing and proposed the concept of cross-task crowdsourcing which shares the knowledge across different domains and solves the problem of the knowledge sparsity of a particular domain. They gave us a vivid example. After a batch of labelers complete the tasks of labeling razor images, it is easy to infer the gender information of these labelers. The inferred gender information transferred from the source domain will be helpful when we select high quality labelers in active learning to label fragrance brand images. This eventually results in a better performance of learning model. The proposed method is based on a probabilistic graphical model, and its inference procedure utilizes Markov Chain Monte Carlo (MCMC) and gradient descent algorithms.

The transfer learning model can be combined with active learning to provide more plentiful information for instance and labeler selection. Fang et al. (2013) proposed a framework that combines knowledge transfer and active learning. As Fig. 3 shows, in this framework the expertise levels of labelers are modeled from historical labeling information in a source domain, and then used in a target domain to conduct instance and labeler selection. The proposed method jointly considers the probability distributions of different types of labels in both source and target domains. Experimental results on real-world data sets show that the

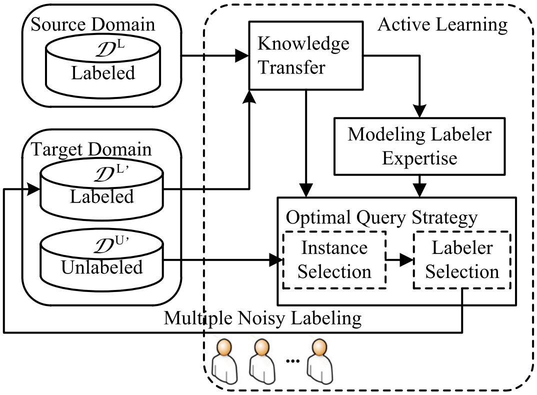


Fig. 3 A framework for transfer learning from crowdsourced labeled data (Fang et al. 2013)

proposed method can significantly improve the performance of active learning if the target domain is similar enough to the source domain.

5 Data sets and tools

Although crowdsourced labeled data is a kind of user generated content, they are really significantly different from social tags from Internet. Crowdsourced data come from the realistic demands of requesters, and its labeling process is accompanied with the expectation of real economic rewards. After labeling, requesters optionally approve answers from labelers to confirm corresponding payments. Because of involving economic interests, the quality of crowdsourced data appears to be higher than ordinary social tags. In fact, it is not the case. On one hand, immature quality control mechanisms cannot guarantee that user will not cheat requesters. On the other hand, once users accept tasks, regardless of the level of their expertise in the application domain, they will provide answers for the sake of rewards. The cost of crowdsourced data collection is higher than that of social tags, but much lower than that of collecting from experts. Currently, crowdsourcing still is a novel research domain less than ten years. Most of the crowdsourced data were collected by researchers themselves, and scattered in their research papers and related academic conferences. To facilitate the future researches in this field, this section lists some information about the open accessible real-world crowdsourced labeled data set and several open source tools.

5.1 Real-world data sets

The earliest contribution of publishing real-world crowdsourced labeled data sets can be traced back to 2008. In order to study whether the crowdsourced label can be really used for natural language processing, Snow et al. (2008) chose data sets containing fine natural language processing tasks, posted them on MTurk and requested users from Internet to label them. The collected data sets are open to the public for research purpose. After that, many researchers chose to publish their data sets collected from crowdsourcing systems in their websites along with their research papers (Ipeirotis et al. 2010; Whitehill et al. 2009, 2010; Rodrigues et al. 2013; Han et al. 2014; Mo et al. 2013). Along with the arrival of the research climax in crowdsourcing, some related international conferences such as TREC 2010 (Buckley et al. 2010) and HCOMP 2013, also organize special tracks and competitions on crowdsourcing topics, which contribute a number of real-world data sets.

Table 3 summarizes commonly used real-world crowdsourced labeled data sets in related researches, and provides the information of their sources, types of labels, URL and brief describes. As the table shows, many crowdsourced labeling tasks were performed on the traditional machine learning related data sets and tasks (Strapparava and Mihalcea 2007; Miller and Charles 1991; Dagan et al. 2006; Pradhan et al. 2007), which facilitates the evaluation of the proposed models and algorithms in their studies.

5.2 Open source tools

Open algorithms and data sets for public acquisition is a tendency in recent data-driven research. Researchers in crowdsourcing also follow this tendency to involve in open source practice (Ipeirotis et al. 2010; Whitehill et al. 2009, 2010; Welinder and Perona 2010; Zhang et al. 2014). Besides these, open source tools for the study of crowdsourced labeling also began to emerge. Nguyen et al. (2013) proposed a tool for label integration research BATC, which implements several ground truth inference algorithms MV, DS, RY, KOS, and GLAD. To facilitate the evaluation of the performance of these algorithms, BATC implements a simulation tool for crowdsourcing labeling with a visual interface. After a user sets parameters such as the number of and the reliability levels of labelers, the simulator in BATC generates noisy labels to form a data set, on which the integration algorithms run. Finally, an analysis report of the running results will be presented visually with charts and tables.

Sheshadri and Lease (2013) proposed another open source tool SQUARE for the ground truth inference study. SQUARE implements and integrates the algorithms MV, DS, RY, GLAD, and ZenCrowd. Different from BATC, SQAURE does not provide visual analysis and simulation tools. Instead, it provides a set of APIs that facilitate users to integrate its function in their own program. SQUARE pays more attention to analyze real-world data sets, so it integrates ten data sets collected from real-world crowdsourcing systems.

The authors of this survey have found that the functions of both BATC and SQUARE have limitation that cannot support different aspects of learning from crowdsourced labeled data. Therefore, we proposed a novel open source tool CEKA (Zhang et al. 2015a) to support the entire research process. CEKA not only contains a large number of ground truth inference algorithms, but also involves the model learning process after inference. CEKA follows the object-oriented design principle, which is fully compatible with a well-known machine learning and data mining tool WEKA (Hall et al. 2009). The sample description file (suffix .arff) and the functions of WAKE can be directly used or called in CEKA. CEKA has a more open architecture, which makes it easy to integrate new algorithms in the future. Figure 4 demonstrates the functional differences between CEKA and the other two tools as well as the high level architecture of CEKA.

5.3 Running examples with CEKA

As running examples, we used real-world binary and multi-class labeling data sets to evaluate the accuracy of several ground truth inference algorithms. These data sets and algorithms have already integrated in CEKA. Table 4 lists the comparison results of the algorithms MV, ZC (Demartini et al. 2012), GLAD (Whitehill et al. 2009), RY Raykar et al. (2010), DS (Dawid and Skene 1979), and PLAT (Zhang et al. 2015c) on nine binary labeling data sets. The worst results are in italic. If there are multiple worst results, the figures are underlined. The best results are in bold. If there are multiple best results, the figures are also underlined. According to the results, we have following observations. (1) Among the four EM-based algorithms, RY and DS have better performance, which suggests DS is a successful model

Table 3 Open accessible real-world crowdsourced labeled data sets

Data set	Source	Label type (#C)	URL	Statistics (#E; #W; #L)	Descriptions
Affect	Snow et al. (2008)	Real [0–10]	https://sites.google.com/site/nlpannotations/	1000; 10; 7000	Judgment of seven emotions based on SemEval Task-14 Strapparava and Mihalcea (2007)
WordSim		Real [0–10]		30; 10; 300	word semantic similarity rating Miller and Charles (1991)
RTE		Binary		800; 164; 8000	Recognize text entailment between two sentences Dagan et al. (2006)
TempOrder		Binary		462; 76; 4620	Judge the order of two events happening (verbs) in a sentence
WSD		Multiclass (3)		177; 10; 1770	Word sense disambiguation Pradhan et al. (2007)
AC2	Ipeirotis et al. (2010)	Multiclass (4)	https://github.com/ipeirotis/get-another-label	333; 269; 3317	Judge whether a webpage contains adult materials, rating (G, PG, R, X)
Spam		Binary		100; 150; 2297	Judge whether an HIT is a spam
BM	Mozafari	Binary	http://people.csail.mit.edu/barzan/datasets/crowdsourcing/	1000; 83; 5000	Judge negative/positive classes of 1000 tweets
WVSCM	Whitehill et al. (2009)	Binary	http://mplab.ucsd.edu/~jake/DuchenneExperiment/DuchenneExperiment.html	159; 17; 1150	Judge whether a smile face in an image is a Duchenne smile
Duck	Welinder et al. (2010)	Binary	https://github.com/welinder/cubam	240; 53; 9600	Judge whether ducks exist in the pictures
Trec2010	TREC 2010 Buckley et al. (2010)	Multiclass (5)	https://www.ischool.utexas.edu/~ml/data/trec-rf10-crowd.tgz	20,232; 766; 98,453	judge the relevance of documents, including five classes: strong relevant, relevant, not relevant, unlabeled and link break
FEJ	HCOMP 2013	Multiclass (3)	http://www.crowdscale.org/shared-task/task-fact-eval	576; 48; 2902	Judge whether statements in Wikipedia are truth, including three classes: yes, no and skip
SAJ		Multiclass (5)	http://www.crowdscale.org/shared-task/sentiment-analysis-judgment-data	300; 461; 1720	judge the sentiment of tweets (five classes)

Table 3 continued

Data set	Source	Label type (#C)	URL	Statistics (#E; #W; #L)	Descriptions
Sentiment	Rodrigues et al. (2013)	Binary	https://eden.dei.uc.pt/~fmpr/malr/	5000; 203; 27,747	Judge the polarity of movie reviews in rottentomatoes.com
MusicGenre		Multiclass ()		700; 44; 2946	Classify an audio clip into one of ten genres
Age	Han et al. (2014)	Real [0–100]	http://www.cse.msu.edu/~hhan/download.htm	1002; 165; 10,020	Estimate the age of a face in a picture
WebSearch	Microsoft	Multiclass (5)	http://research.microsoft.com/en-us/projects/crowd/	2665; 177; 15,567	Judge the relevance of URLs based on five relevance levels
Dog		Multiclass (4)		807; 52; 7354	Classify dogs into one of four breeds
GenderHobby	Mo et al. (2013)	Binary	http://www.cse.ust.hk/~kxmo/materials/GenderHobbyDataSet.rar	200; 40; 4200	Transfer knowledge between two domains (sport and make-up)

#E, **#W**, **#L**, and **#C** represent the number of instances, the number of workers, the number of labels and the number of categories. If the type of a category is real, **#C** represents its range

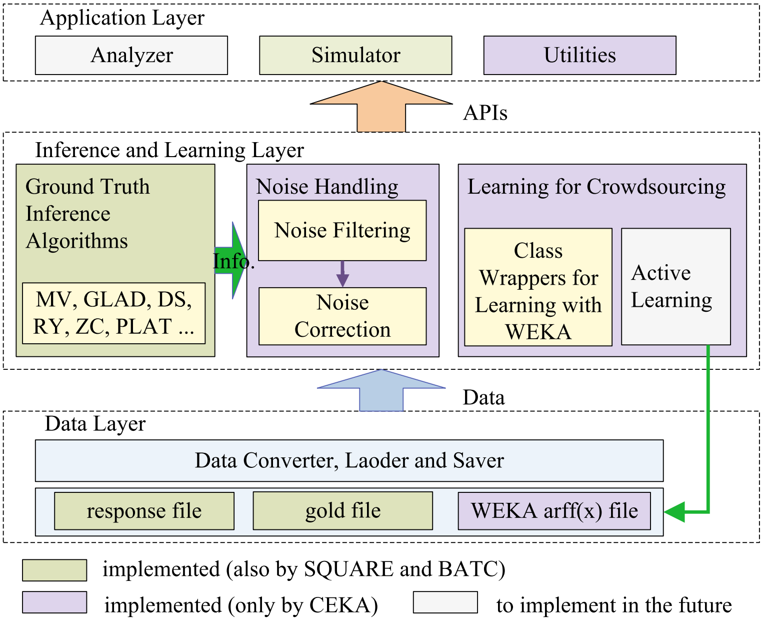

Application Layer
Analyzer
Simulator
Utilities
APIs
Inference and Learning Layer
Ground Truth Inference Algorithms
MV, GLAD, DS, RY, ZC, PLAT ...
Info.
Noise Handling
Noise Filtering
Noise Correction
Learning for Crowdsourcing
Class Wrappers for Learning with WEKA
Active Learning
Data
Data Layer
Data Converter, Laoder and Saver
response file
gold file
WEKA arff(x) file
implemented (also by SQUARE and BATC)
implemented (only by CEKA)
to implement in the future

Fig. 4 The architecture of open source tool CEKA (Zhang et al. 2015a)

Table 4 Comparison results in accuracy (percentage) on nine real-world binary labeling data sets

Data set	MV	ZC	GLAD	RY	DS	PLAT
Fear	*75.0*	*75.0*	*75.0*	79.0	**80.0**	**80.0**
Joy	*66.0*	*66.0*	*66.0*	**77.0**	75.0	75.0
Sadness	76.0	77.0	*74.0*	82.0	**83.0**	81.0
Anger	*70.0*	*70.0*	*70.0*	78.0	79.0	**85.0**
Adult	*84.4*	*84.4*	*84.4*	**88.0**	*84.4*	87.1
WordSim	**90.0**	*86.7*	*86.7*	*86.7*	**90.0**	**90.0**
Trec10	64.2	58.8	*57.8*	67.8	**69.2**	64.6
Duck	68.8	*58.8*	59.6	60.0	60.8	**76.7**
BM	49.7	49.8	*49.3*	50.3	50.7	**51.4**
Average	71.6	69.6	*69.2*	74.3	74.7	**76.7**

for label inference. Although RY is a Bayesian version of DS, it is not always better than DS. (2) In most cases, PLAT has the best performance, because labeling bias is a common phenomenon in crowdsourcing. PLAT is specially proposed for handling the bias, but for unbiased labeling cases it has the same accuracy level as MV does.

For multi-class labeling, we evaluate the algorithms MV, ZC (Demartini et al. 2012), DS (Dawid and Skene 1979), Spectral DS (SDS) (Zhang et al. 2014) and CTIC (Zhang et al. 2016) on nine real-world data sets. From the results shown in Table 5, we have following observations. (1) DS is still a good model for multi-class labeling. It performs much better than ZC, since it has a more complicated model (a confusion matrix for each labeler). (2) Although MV is the simplest method, its performance is not always poor. (3) Spectral DS, an upgraded version of DS, performs worst in average. Spectral DS initializes the parameters of DS using

Table 5 Comparison results in accuracy (percentage) on nine real-world multi-class labeling data sets

Data set	MV	ZC	DS	SDS	GTIC
Fej2013	90.28	90.10	*87.67*	**92.18**	90.76
Trec2010	*44.20*	30.98	**50.27**	47.59	45.48
AC2	75.68	75.98	73.57	*66.06*	**77.78**
Synth4	80.25	*66.75*	68.25	77.33	**81.30**
Valence5	36.00	*32.00*	40.00	45.00	**54.00**
Aircrowd6	80.10	81.45	76.90	*56.70*	**81.96**
Valence7	20.00	9.00	29.00	*19.90*	**32.00**
Leaves9	90.53	90.81	91.92	91.73	**92.20**
Leaves16	60.42	60.27	61.01	*31.10*	**61.46**
Average	64.16	*59.70*	64.28	58.62	**68.55**

a spectral method to prevent trapping into a local optimum. However, it outperforms DS only on three data sets. Especially, as the number of classes increases, Spectral DS performs worse. (4) GTIC has the best performance in average. Maybe it has the ability to cluster examples with similar feature patterns (potentially belonging to one class) together.

6 Conclusions and prospects

We are now in the era of big data. The importance of knowledge discovery from massive data is unprecedented emphasized. In a large number of intelligent systems, the learning models are continuously optimized as the accumulation of the data. Constructing great learning models is one of the most maturely developed branches in machine learning (Wen et al. 2015). Although it has been widely used, it still suffers from the deficiency of the training data. The emergence of crowdsourcing has greatly alleviated the contradiction between the demand of the learning model and the deficiency of the training data. We can obtain plenty of labeled data from crowdsourcing systems. However, the uncertainty of labelers derived from the open nature of the systems is a severe challenge to the quality of collected data and learning models.

In this paper, based on the concepts of the label quality and the learning model quality, we elaborate the relationship between ground truth inference and building learning model. Then, we review the inference and the learning algorithms proposed in the past eight years and discuss the similarities and differences among these methods. Finally, in order to promote the research in this field, we also summarize open accessible real-world crowdsourced data sets as well as open source tools. Since 2008, in a short span of less than ten years, a lot of new algorithms and models have burst out. However, it is still in a young stage. The problems that are worthy of studying in the future at least include the following aspects.

1. *More fine-grained inference methods* At present, the research on the general purpose ground truth algorithms is gradually mature. Various models have been able to describe the system from different aspects. However, empirical studies have shown that due to the huge difference among data, there is no algorithm that are universally superior to others (Sheshadri and Lease 2013). Therefore, in order to further improve the accuracy of the inference, more fine-grained methods need to be further studied. For example, we should consider the historical information of labelers, and the physical features of examples, and assign different weighs to labelers according to application domains.

2. *More learning models* The current researches on ground truth inference and learning model training are mainly focus on binary labeling. Although quit a few models are claimed that they are suitable for multi-class cases, no sufficient empirical study consolidates their arguments. Effective multi-class and multi-label researches in crowdsourcing are limited (Vempaty et al. 2014; Bragg et al. 2013). In multi-class labeling, the sparsity of data and of labels cannot be ignored. The biased labeling issue is more complicated. Similarly, the study of multi-label learning in crowdsourcing is a brand new topic. A series of issues, such as how to reduce the negative impact of the sparsity, need to be studied in the future.
3. *More deeply relationship between the label quality and the learning model quality* Improving the qualities of labels is conducive to improve the quality of learning models. However, they are not equivalent. To find more valuable examples and accurately label them is propitious to make a trade-off between the cost and the quality of a leaning model. At the same time, a large number of practical problems are cost sensitive (Ling and Sheng 2010). It is more useful to make a two-way optimization between labeling cost and mislabel cost in a cost-sensitive context.
4. *Theories of learning from crowdsourced labeled data* Current proposed algorithms on ground truth inference and learning model training are based on existing theories and techniques, such as probabilistic graphical model, EM algorithm, convex optimization and matrix singular value decomposition. Although massive empirical studies have shown the effectiveness of these techniques, there still remain some basic theoretical questions unsolved, for example, the upper and lower bounds of the accuracy of an inference algorithm based on EM, overfitting and the way to avoid it, the impact of the size of training data on the performance, and so on. Uncertainty in crowdsourcing makes theoretical answers to these questions full of challenges.

References

Allahbakhsh M, Benatallah B, Ignjatovic A, Motahari-Nezhad HR, Bertino E, Dustdar S (2013) Quality control in crowdsourcing systems: issues and directions. IEEE Internet Comput 2:76–81

Bernardi C, Maday Y (1997) Handbook of numerical analysis. Spectr Methods 5:209–485

Bernstein MS, Little G, Miller RC, Hartmann B, Ackerman MS, Karger DR, Crowell D, Panovich K (2010) Soylent: a word processor with a crowd inside. In: Proceedings of the 23nd annual ACM symposium on user interface software and technology, ACM, pp 313–322

Bragg J, Weld DS, et al (2013) Crowdsourcing multi-label classification for taxonomy creation. In: First AAAI conference on human computation and crowdsourcing, pp 25–33

Breiman L (2001) Random forests. Mach Learn 45(1):5–32

Brew A, Greene D, Cunningham P (2010) The interaction between supervised learning and crowdsourcing. In: NIPS workshop on computational social science and the wisdom of crowds

Brodley CE, Friedl MA (1999) Identifying mislabeled training data. J Artif Intell Res 11:131–167

Buckley C, Lease M, Smucker MD, Jung HJ, Grady C, Buckley C, Lease M, Smucker MD, Grady C, Lease M, et al (2010) Overview of the trec 2010 relevance feedback track (notebook). In: The nineteenth text retrieval conference (TREC) notebook

Carvalho VR, Lease M, Yilmaz E (2011) Crowdsourcing for search evaluation. ACM Sigir Forum ACM 44:17–22

Corney J, Lynn A, Torres C, Di Maio P, Regli W, Forbes G, Tobin L (2010) Towards crowdsourcing translation tasks in library cataloguing, a pilot study. In: The 4th IEEE international conference on digital ecosystems and technologies(DEST), IEEE, pp 572–577

Dagan I, Glickman O, Magnini B (2006) The pascal recognising textual entailment challenge. In: Quiñonero-Candela J, Dagan I, Magnini B, d'Alché-Buc F (eds) Machine learning challenges. evaluating predictive uncertainty, visual object classification, and recognising tectual entailment, Springer, pp 177–190

Dalvi N, Dasgupta A, Kumar R, Rastogi V (2013) Aggregating crowdsourced binary ratings. In: Proceedings of the 22nd international conference on world wide web, International World Wide Web conferences steering committee, pp 285–294

Dawid AP, Skene AM (1979) Maximum likelihood estimation of observer error-rates using the em algorithm. Appl Stat 28(1):20–28

Demartini G, Difallah DE, Cudré-Mauroux P (2012) Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In: Proceedings of the 21st international conference on world wide web, ACM, pp 469–478

Doan A, Ramakrishnan R, Halevy AY (2011) Crowdsourcing systems on the world-wide web. Commun ACM 54(4):86–96

Donmez P, Carbonell JG, Schneider J (2009) Efficiently learning the accuracy of labeling sources for selective sampling. In: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 259–268

Donmez P, Carbonell JG, Schneider JG (2010) A probabilistic framework to learn from multiple annotators with time-varying accuracy. In: Proceedings of the 10th SIAM international conference on data mining, SIAM, pp 826–837

Dow S, Kulkarni A, Klemmer S, Hartmann B (2012) Shepherding the crowd yields better work. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, ACM, pp 1013–1022

Downs JS, Holbrook MB, Sheng S, Cranor LF (2010) Are your participants gaming the system? Screening mechanical turk workers. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 2399–2402

Eagle N (2009) txteagle: mobile crowdsourcing. In: Aykin N (ed) Internationalization, design and global development, Springer, pp 447–456

Evgeniou T, Pontil M (2004) Regularized multi-task learning. In: Proceedings of the tenth ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 109–117

Faltings B, Jurca R, Pu P, Tran BD (2014) Incentives to counter bias in human computation. In: Second AAAI conference on human computation and crowdsourcing

Fang M, Yin J, Zhu X (2013) Knowledge transfer for multi-labeler active learning. In: Machine learning and knowledge discovery in databases, Springer, pp 273–288

Fang M, Zhu X, Li B, Ding W, Wu X (2012) Self-taught active learning from crowds. In: Data mining (2012 IEEE 12th international conference on ICDM), IEEE, pp 858–863

Frénay B, Verleysen M (2014) Classification in the presence of label noise: a survey. IEEE Trans Neural Netw Learn Syst 25(5):845–869

Freund Y (2001) An adaptive version of the boost by majority algorithm. Mach Learn 43(3):293–318

Fu Y, Zhu X, Li B (2013) A survey on instance selection for active learning. Knowl Inf Syst 35(2):249–283

Gelman A, Carlin JB, Stern HS, Rubin DB (2014) Bayesian data analysis, vol 2. Taylor & Francis, London

Ghosh A, Kale S, McAfee P (2011) Who moderates the moderators? Crowdsourcing abuse detection in user-generated content. In: Proceedings of the 12th ACM conference on electronic commerce, ACM, pp 167–176

Grady C, Lease M (2010) Crowdsourcing document relevance assessment with mechanical turk. In: Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's mechanical turk, association for computational linguistics, pp 172–179

Gu B, Sheng VS, Tay KY, Romano W, Li S (2014) Incremental support vector learning for ordinal regression. IEEE Trans Neural Netw Learn Syst 26(7):1403–1416

Gu B, Sheng VS, Wang Z, Ho D, Osman S, Li S (2015) Incremental learning for ν-support vector regression. Neural Netw 67:140–150

Halevy A, Norvig P, Pereira F (2009) The unreasonable effectiveness of data. IEEE Intell Syst 24(2):8–12

Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH (2009) The weka data mining software: an update. ACM SIGKDD Explor Newslett 11(1):10–18

Han H, Otto C, Liu X, Jain A (2014) Demographic estimation from face images: human vs. machine performance. IEEE Trans Pattern Anal Mach Intell 37(6):1148–1161

Howe J (2006) The rise of crowdsourcing. Wired Mag 14(6):1–4

Ipeirotis PG, Provost F, Wang J (2010) Quality management on amazon mechanical turk. In: Proceedings of the ACM SIGKDD workshop on human computation, ACM, pp 64–67

Ipeirotis PG, Provost F, Sheng VS, Wang J (2014) Repeated labeling using multiple noisy labelers. Data Min Knowl Discov 28(2):402–441

Jung HJ, Lease M (2011) Improving consensus accuracy via z-score and weighted voting. In: Proceedings of the 3rd human computation workshop (HCOMP) at AAAI

Kajino H, Kashima H (2012) Convex formulations of learning from crowds. Trans Jan Soc Artif Intell 27:133–142

Karger DR, Oh S, Shah D (2011) Budget-optimal crowdsourcing using low-rank matrix approximations. In: Communication, control, and computing (Allerton), 2011 49th annual allerton conference on, IEEE, pp 284–291

Karger DR, Oh S, Shah D (2014) Budget-optimal task allocation for reliable crowdsourcing systems. Oper Res 62(1):1–24

Khattak FK, Salleb-Aouissi A (2011) Quality control of crowd labeling through expert evaluation. In: Proceedings of the NIPS 2nd workshop on computational social science and the wisdom of crowds

Khetan A, Oh S (2016) Reliable crowdsourcing under the generalized dawid-skene model. arXiv:1602.03481

Kittur A, Smus B, Khamkar S, Kraut RE (2011) Crowdforge: crowdsourcing complex work. In: Proceedings of the 24th annual ACM symposium on User interface software and technology, ACM, pp 43–52

Koller D, Friedman N (2009) Probabilistic graphical models: principles and techniques. MIT press, Cambridge

Kulkarni A, Can M, Hartmann B (2012) Collaboratively crowdsourcing workflows with turkomatic. In: Proceedings of the ACM 2012 conference on computer supported cooperative work, ACM, pp 1003–1012

Kurve A, Miller DJ, Kesidis G (2015) Multicategory crowdsourcing accounting for variable task difficulty, worker skill, and worker intention. IEEE Trans Knowl Data Eng 27(3):794–809

Lease M (2011) On quality control and machine learning in crowdsourcing. In: Proceedings of the 3rd human computation workshop (HCOMP) at AAAI

Li H, Yu B (2014) Error rate bounds and iterative weighted majority voting for crowdsourcing. arXiv:1411.4086

Li J, Ott M, Cardie C, Hovy E (2014) Towards a general rule for identifying deceptive opinion spam. In: Proceedings of the 52nd annual meeting of the association for computational linguistics, ACL

Li J, Li X, Yang B, Sun X (2015) Segmentation-based image copy-move forgery detection scheme. IEEE Trans Inf Forensics Secur 10(3):507–518

Lin CH, Weld DS, et al (2014) To re (label), or not to re (label). In: Second AAAI conference on human computation and crowdsourcing

Ling CX, Sheng VS (2010) Cost-sensitive learning. In: Sammut C, Webb GI (eds) Encyclopedia of machine learning, Springer, pp 231–235

Little G, Chilton LB, Goldman M, Miller RC (2009) Turkit: tools for iterative tasks on mechanical turk. In: Proceedings of the ACM SIGKDD workshop on human computation, ACM, pp 29–30

Liu K, Cheung WK, Liu J (2015) Detecting multiple stochastic network motifs in network data. Knowl Inf Syst 42(1):49–74

Liu Q, Peng J, Ihler AT (2012) Variational inference for crowdsourcing. In: Advances in neural information processing systems, pp 692–700

Long C, Hua G, Kapoor A (2013) Active visual recognition with expertise estimation in crowdsourcing. In: 2013 IEEE international conference on computer vision (ICCV), IEEE, pp 3000–3007

Michalski RS, Carbonell JG, Mitchell TM (2013) Machine learning: an artificial intelligence approach. Springer, Berlin

Miller GA, Charles WG (1991) Contextual correlates of semantic similarity. Lang Cognit Process 6(1):1–28

Mo K, Zhong E, Yang Q (2013) Cross-task crowdsourcing. In: Proceedings of the 19th ACM SIGKDD international conference on knowledge discovery and data mining, ACM, pp 677–685

Muhammadi J, Rabiee HR, Hosseini A (2015) A unified statistical framework for crowd labeling. Knowl Inf Syst 45(2):271–294

Natarajan N, Dhillon IS, Ravikumar PK, Tewari A (2013) Learning with noisy labels. In: Advances in neural information processing systems, vol 26. pp 1196–1204

Nguyen QVH, Nguyen TT, Lam NT, Aberer K (2013) Batc: a benchmark for aggregation techniques in crowdsourcing. In: Proceedings of the 36th international ACM SIGIR conference on research and development in information retrieval, ACM, pp 1079–1080

Nicholson B, Zhang J, Sheng VS, Wang Z (2015) Label noise correction methods. In: IEEE International Conference on, IEEE, data science and advanced analytics (DSAA), 2015. 36678 2015, pp 1–9

Nowak S, Rüger S (2010) How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In: Proceedings of the international conference on multimedia information retrieval, ACM, pp 557–566

Oyen D, Lane T (2015) Transfer learning for Bayesian discovery of multiple Bayesian networks. Knowl Inf Syst 43(1):1–28

Pan SJ, Yang Q (2010) A survey on transfer learning. IEEE Trans Knowl Data Eng 22(10):1345–1359

Parhami B (1994) Voting algorithms. IEEE Trans Reliab 43(4):617–629

Pradhan SS, Loper E, Dligach D, Palmer M (2007) Semeval-2007 task 17: english lexical sample, srl and all words. In: Proceedings of the 4th international workshop on semantic evaluations, association for computational linguistics, pp 87–92

Prati RC, Batista GEAPA, Silva DF (2015) Class imbalance revisited: a new experimental setup to assess the performance of treatment methods. Knowl Inf Syst 45(1):247–270

Prpic J, Shukla P (2013) The theory of crowd capital. In: The 46th Hawaii international conference on system sciences (HICSS), IEEE, pp 3505–3514

Prpic J, Shukla P (2014) The contours of crowd capability. In: The 47th Hawaii international conference on system sciences (HICSS), IEEE, pp 3461–3470

Rätsch G, Schölkopf B, Smola AJ, Mika S, Onoda T, Müller KR (2000) Robust ensemble learning for data mining. Knowledge discovery and data mining. Current issues and new applications. Springer, Berlin, pp 341–344

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

Raykar VC, Yu S, Zhao LH, Jerebko A, Florin C, Valadez GH, Bogoni L, Moy L (2009) Supervised learning from multiple experts: whom to trust when everyone lies a bit. In: Proceedings of the 26th annual international conference on machine learning, ACM, pp 889–896

Raykar VC, Yu S, Zhao LH, Valadez GH, Florin C, Bogoni L, Moy L (2010) Learning from crowds. J Mach Learn Res 11:1297–1322

Rodrigues F, Pereira F, Ribeiro B (2013) Learning from multiple annotators: distinguishing good from random labelers. Pattern Recogn Lett 34(12):1428–1436

Rodrigues F, Pereira F, Ribeiro B (2014) Gaussian process classification and active learning with multiple annotators. In: Proceedings of the 31st international conference on machine learning (ICML-14), pp 433–441

Ross J, Irani L, Silberman M, Zaldivar A, Tomlinson B (2010) Who are the crowdworkers? Shifting demographics in mechanical turk. In: CHI'10 extended abstracts on human factors in computing systems, ACM, pp 2863–2872

Settles B (2010) Active learning literature survey. Univ Wis Madison 52(55–66):11

Settles B, Craven M (2008) An analysis of active learning strategies for sequence labeling tasks. In: Proceedings of the conference on empirical methods in natural language processing, Association for Computational Linguistics, pp 1070–1079

Shah NB, Zhou D (2015) Double or nothing: multiplicative incentive mechanisms for crowdsourcing. In: Advances in neural information processing systems, vol 28. pp 1–9

Shah NB, Zhou D, Peres Y (2015) Approval voting and incentives in crowdsourcing. In: Proceedings of the 32nd international conference on machine learning (ICML)

Sheng VS, Provost F, Ipeirotis PG (2008) Get another label? improving data quality and data mining using multiple, noisy labelers. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, pp 614–622

Sheng VS (2011) Simple multiple noisy label utilization strategies. In: Data mining (ICDM), 2011 IEEE 11th international conference on, IEEE, pp 635–644

Sheshadri A, Lease M (2013) Square: a benchmark for research on computing crowd consensus. In: First AAAI conference on human computation and crowdsourcing, AAAI, pp 156–164

Smyth P, Burl MC, Fayyad UM, Perona P (1994) Knowledge discovery in large image databases: dealing with uncertainties in ground truth. In: KDD workshop, pp 109–120

Smyth P, Fayyad U, Burl M, Perona P, Baldi P (1995) Inferring ground truth from subjective labelling of venus images. In: Advances in Neural Information Processing Systems, vol 7. pp 1085–1092

Snow R, O'Connor B, Jurafsky D, Ng AY (2008) Cheap and fast—but is it good? Evaluating non-expert annotations for natural language tasks. In: Proceedings of the conference on empirical methods in natural language processing, association for computational linguistics, pp 254–263

Sorokin A, Forsyth D (2008) Utility data annotation with amazon mechanical turk. In: Proceedings of the First IEEE Workshop on Internet Vision at CVPR 2008, pp 1–8

Steinwart I, Christmann A (2008) Support vector machines. Springer, Berlin

Strapparava C, Mihalcea R (2007) Semeval-2007 task 14: affective text. In: Proceedings of the 4th international workshop on semantic evaluations, association for computational linguistics, pp 70–74

Sukhbaatar S, Bruna J, Paluri M, Bourdev L, Fergus R (2014) Training convolutional networks with noisy labels. arXiv:1406.2080

Su Q, Pavlov D, Chow JH, Baker WC (2007) Internet-scale collection of human-reviewed data. In: Proceedings of the 16th international conference on world wide web, ACM, pp 231–240

Tang W, Lease M (2011) Semi-supervised consensus labeling for crowdsourcing. In: SIGIR workshop on crowdsourcing for information retrieval, pp 66–75

Tian T, Zhu J (2015) Uncovering the latent structures of crowd labeling. In: Pacific-Asia conference on knowledge discovery and data mining, pp 392–404

Ting KM (2002) An instance-weighting method to induce cost-sensitive trees. IEEE Trans Knowl Data Eng 14(3):659–665

Tong Y, Cao CC, Zhang CJ, Li Y, Chen L (2014) Crowdcleaner: data cleaning for multi-version data on the web via crowdsourcing. In: 2014 IEEE 30th international conference on data engineering (ICDE), IEEE, pp 1182–1185

Urbano J, Morato J, Marrero M, Martín D (2010) Crowdsourcing preference judgments for evaluation of music similarity tasks. In: ACM SIGIR workshop on crowdsourcing for search evaluation, pp 9–16

Vempaty A, Varshney LR, Varshney PK (2014) Reliable crowdsourcing for multi-class labeling using coding theory. IEEE J Sel Top Signal Process 8(4):667–679

Von Ahn L (2009) Human computation. In: The 46th ACM/IEEE design automation conference (DAC'09), IEEE, pp 418–419

Von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of the SIGCHI conference on human factors in computing systems, ACM, pp 319–326

Von Ahn L, Maurer B, McMillen C, Abraham D, Blum M (2008) recaptcha: Human-based character recognition via web security measures. Science 321(5895):1465–1468

Vondrick C, Patterson D, Ramanan D (2013) Efficiently scaling up crowdsourced video annotation. Int J Comput Vis 101(1):184–204

Wainwright MJ, Jordan MI (2008) Graphical models, exponential families, and variational inference. Found Trends Mach Learn 1(1–2):1–305

Wang G, Wang T, Zheng H, Zhao BY (2014) Man vs. machine: practical adversarial detection of malicious crowdsourcing workers. In: 23rd USENIX security symposium, USENIX Association, CA

Watanabe M, Yamaguchi K (2003) The EM algorithm and related statistical models. CRC Press, Boca Raton

Wauthier FL, Jordan MI (2011) Bayesian bias mitigation for crowdsourcing. In: Advances in neural information processing systems, pp 1800–1808

Weiss GM, Hirsh H (1998) The problem with noise and small disjuncts. In: ICML, pp 574–578

Welinder P, Perona P (2010) Online crowdsourcing: rating annotators and obtaining cost-effective labels. In: The 2010 IEEE computer society conference on computer vision and pattern recognition workshops (CVPRW), IEEE, pp 25–32

Welinder P, Branson S, Perona P, Belongie SJ (2010) The multidimensional wisdom of crowds. In: Advances in neural information processing systems (NIPS), vol 23. pp 2424–2432

Wen X, Shao L, Xue Y, Fang W (2015) A rapid learning algorithm for vehicle classification. Inf Sci 295:395–406

Whitehill J, Wu Tf, Bergsma J, Movellan JR, Ruvolo PL (2009) Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In: Advances in neural information processing systems (NIPS), pp 2035–2043

Williams CK, Barber D (1998) Bayesian classification with gaussian processes. IEEE Trans Pattern Anal Mach Intell 20(12):1342–1351

Wu W, Liu Y, Guo M, Wang C, Liu X (2013) A probabilistic model of active learning with multiple noisy oracles. Neurocomputing 118:253–262

Xu Q, Huang Q, Yao Y (2012) Online crowdsourcing subjective image quality assessment. In: Proceedings of the 20th ACM international conference on multimedia, ACM, pp 359–368

Yan T, Kumar V, Ganesan D (2010a) Crowdsearch: exploiting crowds for accurate real-time image search on mobile phones. In: Proceedings of the 8th international conference on mobile systems, applications, and services, ACM, pp 77–90

Yan Y, Rosales R, Fung G, Dy J (2010b) Modeling multiple annotator expertise in the semi-supervised learning scenario. In: Proceedings of conference on uncertainty in artificial intelligence, pp 674–682

Yan Y, Rosales R, Fung G, Schmidt MW, Valadez GH, Bogoni L, Moy L, Dy JG (2010c) Modeling annotator expertise: learning when everybody knows a bit of something. In: International conference on artificial intelligence and statistics, pp 932–939

Yan Y, Fung GM, Rosales R, Dy JG (2011) Active learning from crowds. In: Proceedings of the 28th international conference on machine learning (ICML-11), pp 1161–1168

Zellner A (1996) An introduction to Bayesian inference in econometrics. Wiley, New York

Zhang Z, Pang J, Xie X (2013) Research on crowdsourcing quality control strategies and evaluation algorithm. Chin J Comput 8:1636–1649

Zhang Y, Chen X, Zhou D, Jordan MI (2014) Spectral methods meet em: a provably optimal algorithm for crowdsourcing. In: Advances in neural information processing systems, vol 27. pp 1260–1268

Zhang J, Sheng V, Nicholson BA, Wu X (2015a) Ceka: a tool for mining the wisdom of crowds. J Mach Learn Res 16:2853–2858

Zhang J, Sheng VS, Wu J, Fu X, Wu X (2015b) Improving label quality in crowdsourcing using noise correction. In: Proceedings of the 24th ACM international on conference on information and knowledge management, ACM, pp 1931–1934

Zhang J, Wu X, Sheng VS (2015c) Imbalanced multiple noisy labeling. IEEE Trans Knowl Data Eng 27(2):489–503

Zhang J, Wu X, Shengs VS (2015d) Active learning with imbalanced multiple noisy labeling. IEEE Trans Cybern 45(5):1081–1093

Zhang J, Sheng V, Wu J, Wu X (2016) Multi-class ground truth inference in crowdsourcing with clustering. IEEE Trans Knowl Data Eng 28(4):1080–1085

Zhong J, Tang K, Zhou ZH (2015) Active learning from crowds with unsure option. In: Proceedings of 2015 international joint conference on artificial intelligence

Zhou D, Basu S, Mao Y, Platt JC (2012) Learning from the wisdom of crowds by minimax entropy. In: Advances in neural information processing systems (NIPS), pp 2195–2203

Zhou D, Liu Q, Platt J, Meek C (2014) Aggregating ordinal labels from crowds by minimax conditional entropy. In: Proceedings of the 31st international conference on machine learning (ICML-14), pp 262–270