Machine Learning with Crowdsourcing: A Brief Summary of the Past Research and Future Directions

Victor S. Sheng
Department of Computer Science
University of Central Arkansas
201 Donaghey Avenue
Conway, AR 72035, U.S.A.

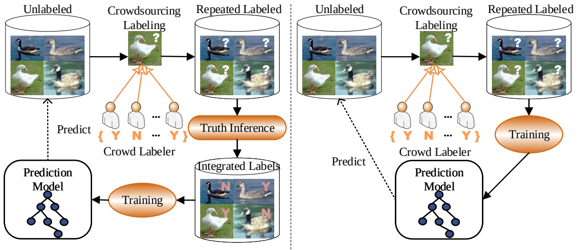
Jing Zhang*
School of Computer Science and Engineering
Nanjing University of Science and Technology
200 Xiaolingwei Street
Nanjing 210094, China

Abstract

With crowdsourcing systems, labels can be obtained with low cost, which facilitates the creation of training sets for prediction model learning. However, the labels obtained from crowdsourcing are often imperfect, which brings great challenges in model learning. Since 2008, the machine learning community has noticed the great opportunities brought by crowdsourcing and has developed a large number of techniques to deal with inaccuracy, randomness, and uncertainty issues when learning with crowdsourcing. This paper summarizes the technical progress in this field during past eleven years. We focus on two fundamental issues: the data (label) quality and the prediction model quality. For data quality, we summarize ground truth inference methods and some machine learning based methods to further improve data quality. For the prediction model quality, we summarize several learning paradigms developed under the crowdsourcing scenario. Finally, we further discuss several promising future research directions to attract researchers to make contributions in crowdsourcing.

Introduction

With the emergence of crowdsourcing systems, such as Amazon Mechanical Turk, Figure Eight (CrowdFlower), etc., our machine learning community first realized that these systems might provide opportunities for machine learning research (Lease 2011). The most common way that crowdsourcing can facilitate machine learning is to collect labels for training models. Starting from our well-known work (Sheng, Provost, and Ipeirotis 2008), supervised learning models can be created in a crowdsourcing scenario. To solve the uncertainty of non-expert workers in crowdsourcing, requesters usually let each instance be labeled by multiple workers, which is called *repeated-labeling*. Machine learning with crowdsourcing is a new learning paradigm, involving human-in-the-loop learning activities and exhibiting significant importance in today's big data era. This paradigm has already been put into some applications, such as relation extraction (Abad, Nabi, and Moschitti 2017), image recognition (Deng, Krause, and Li 2013), etc, to form evolving intelligent systems. These systems usually require continuous learning, and continuous learning requires training samples to be efficiently labeled as needed.



(a) Leaning with truth inference

(b) Leaning without truth inference

Figure 1: Two paradigms (w/ or w/o truth inference) of learning prediction models from crowds.

Since concept learning requires building prediction models from the data with concept classes (in which each instance has a unique class label), a dataset with repeated crowdsourced labels is usually fed into a truth inference algorithm to obtain an integrated label for each instance before building a prediction model. This two-stage learning scenario is illustrated in Fig. 1(a), which currently is a mainstream framework for crowd-sourced learning. Thus, truth inference has received considerable attention in our machine learning community, because its results directly affect the quality of subsequently learned models. Of course, some studies (Kajino, Tsuboi, and Kashima 2012; 2013) followed the learning framework without truth inference as Fig. 1(b) illustrates. Although whether this mode can effectively improve the quality of learned models is still questionable [1], it is still an interesting research topic. One of important reasons that we use crowdsourcing to collect labels is to reduce labeling costs, which is exactly compatible with the goal of active learning. Active learning with crowdsourcing, on the one hand, must consider the instance selection strategies; on the other hand, it optionally considers worker selection strategies (Yan et al. 2011).

[*] Dr. Jing Zhang is the corresponding author of the paper. (Contact email: jzhang@njust.edu.cn) Two authors contributed equally.

[1] There are two reasons that we prefer the separation of data integration and model training. First, we have not observed any meaningful performance improvement for the bundled model. Moreover, when the system has a problem, we usually need to know precisely which part is out of order. The bundled model blurs the boundaries between integration and training.

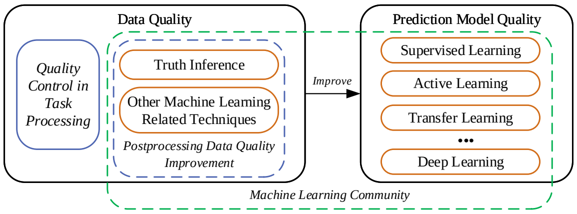

Data Quality
Quality Control in Task Processing
Truth Inference
Other Machine Learning Related Techniques
Postprocessing Data Quality Improvement
Improve
Prediction Model Quality
Supervised Learning
Active Learning
Transfer Learning
•••
Deep Learning
Machine Learning Community

Figure 2: Techniques related to data quality and prediction model quality.

During the past eleven years from 2008 through 2018, significant progress has been made in the field of machine learning with crowdsourcing. Our machine learning community has developed a large number of techniques to deal with inaccuracy, randomness, and uncertainty problems in both truth inference and prediction model learning. This paper does not intend to conduct a comprehensive survey on either truth inference or prediction model learning [2]. Instead, we attempt to show a blueprint for the entire technological development in this field. In particular, with our understanding, we try to classify these techniques into several reasonable categories and explain why we need these techniques. We highlight several core technical unsolved issues and discuss potential research topics in detail, which may help new researchers to enter into this field.

Overview of Techniques

When learning from crowdsourced data, two core issues that must be well-solved are the improvement of both data quality and prediction model quality. Fig. 2 illustrates relationships between these two qualities, techniques involved, and the interests of our machine learning community. The *data* (*label*) *quality* measures the accuracy that crowdsourced (derived) labels match their true values. The (*prediction*) *model quality* measures the generalization performance of a learned model trained from a crowdsourced dataset. Intuitively, the latter improves with the improvement of the former. Therefore, researchers in different domains have made great efforts to improve data quality.

Quality Control in Task Processing aims to introduce some quality control mechanisms (Allahbakhsh et al. 2013), which regulate the behaviors of workers so that they are willing or might be forced to provide better answers. This direction usually attracts researchers in operation, gaming, and management, which is out of the scope of this paper.

Postprocessing Data Quality Improvement aims to use a simple *repeated-labeling* scheme to improve the data (label) quality after tasks are completed. Because each instance in this scheme will obtain a set of multiple noisy labels, we need to develop an algorithm to infer the true label for each instance. Since techniques here have a natural connection with the statistical inference, our machine learning community has developed numerous statistical inference algorithms. There are some other methods, which use machine learning technologies to improve data quality.

Compared with the fruitful results of the truth inference field, studies of building prediction models are still in an early stage. One of the difficulties lies in how to adapt traditional machine learning algorithms to utilize the information provided by crowd annotations to overcome the negative effects of the labeling uncertainty. Another direction is to develop novel model training methods in a crowdsourcing scenario. We are delighted to find that many learning paradigms, such as active learning, cost-sensitive learning, deep learning, etc., have made new progress in recent years.

Truth Inference

Truth inference is defined as a process of estimating the true label of each instance from its multiple label set. The connotation of truth inference in crowdsourcing includes: (1) truth inference must work in an agnostic way, where no other information except that observed labels can be used; (2) the core evaluation metric is the accuracy of the inferred (integrated) labels matching their true values; and (3) other information (parameters), such as the reliability of workers, the difficulty of instances, etc., could be inferred.

Probabilistic Generative Methods

The simplest but effective method is majority voting, which works well in most *binary-labeling* cases but may malfunction in some complicated circumstances, where spammers appear or labeling exhibits bias. Comparably, probabilistic generative methods have more solid theoretical basis, which may perform better. Inspired by the classic Dawid & Skene's model (Dawid and Skene 1979), early work focused on using an EM algorithm for statistical inference. Raykar et al. (2010) introduced the Bayesian estimation to model the sensitivity and the specificity of workers, which improves the performance of the truth inference in binary biased labeling. Besides modeling workers, some methods model the various aspects of crowdsourced labeling. GLAD (Whitehill et al. 2009) introduced the difficulty of instances into their model. Welinder et al. (2010) proposed a more complex multi-dimensional model, in which the feature noises of instances were also introduced while considering the reliability of workers and the difficulty of instances. To refine the modelling of workers, Bi et al. (2014) proposed a method to add the *dedication* of workers to the inference model, and Kurve, Miller, and Kesidis (2015) added the *intention* of workers to the inference model, which can distinguish malicious workers from normal ones.

The advantage of probabilistic generative methods is that the model can be solved by several machine learning techniques, such as sampling (e.g., (Kim and Ghahramani 2012)), EM algorithms, convex optimization (e.g., (Zhou et al. 2012)), and even variational inference (e.g., (Liu, Peng, and Ihler 2012)). Among them, the deterministic EM algorithm is an effective tool to solve the MLE or MAP estimates with latent variables. However, one of the difficulties of EM is how to properly set the initial values of the model parameters, which obviously affects the inference accuracy.

[2] Zhang, Wu, and Sheng (2016) summarized the progress in learning with crowdsourced labeled data. Zheng et al. (2017) provided comparisons among truth inference techniques.

Zhang et al. (2016b) proposed a spectral method to solve this problem in Dawid & Skene's model. Compared with the EM algorithm that seeks a local optimum, Zhou et al. (2012) proposed a global-optimized inference algorithm based on the minimax entropy principle, whose objective function is convex. Tian and Zhu (2015b) first assumed that instances belong to different latent classes where crowdsourced workers' behaviors are different, and then extended the minimax entropy estimator to a non-parametric form to uncover these latent classes when performing inference.

Discriminative Methods

To overcome the weaknesses of probabilistic generative methods, researchers in this field have already proposed several discriminative methods that are based on different techniques such as matrix factorization and convex optimization. The early KOS algorithm (Karger, Oh, and Shah 2011) incorporates singular value decomposition (SVD) of a low-rank matrix with a belief propagation-like procedure to achieve inference. KOS works well when the noisy label matrix is full. The method proposed by Dalvi et al. (2013) also relies on SVD, but relaxes the above prerequisite. Max-margin majority voting (Tian and Zhu 2015a), inspired by the margin calculation in multi-class SVM, is a discriminative model that directly finds the most likely label for each instance by maximizing the margin. In addition, this method also has a Bayesian version, which can by solved by Gibbs sampling or variational inference. In fact, several weighted voting based methods are all discriminative (Aydin et al. 2014; Li et al. 2014), which simply optimize objective functions to obtain the estimates to the labels. Zhou and He (2016) recently proposed two structured methods based on tensor augmentation and completion. The two methods use tensor representation for the labeled data, augment it with a ground truth layer, and estimate the true labels via low rank tensor completion. For biased labeling, we proposed a discriminative method PLAT that can automatically adjust the decision threshold between the inferred positive and negative instances (Zhang, Wu, and Sheng 2015). We also proposed a clustering-based method GTIC for multi-class label inference (Zhang et al. 2016a).

In summary, generative methods are usually probabilistic, and discriminative methods are usually non-Probabilistic. The performance of a generative method is usually more sensitive to data sparsity.

Multi-Class and Multi-Label Cases

Although several methods (Dawid and Skene 1979; Karger, Oh, and Shah 2011; Demartini, Difallah, and Cudré-Mauroux 2012; Zhou et al. 2012; Kurve, Miller, and Kesidis 2015; Zhang et al. 2016b) are naturally suitable for multi-class labeling, they are seldom optimized for multi-class cases. Some other studies have made progress on this issue. Karger, Oh, and Shah (2013) proposed an inference method that combines low-rank matrix approximation and majority voting to obtain estimates. By reducing the K-ary classification tasks into a series of $K-1$ simple binary classification tasks, the performance of the K-ary estimator is improved. We also proposed an inference algorithm GTIC (Zhang et al. 2016a), which generates conceptual features for instances from repeated labels, and uses a K-Means algorithm to cluster all instances into K classes. GTIC can capture the tendency of labeling bias with respect to groups of instances. Ordinal labels sometimes can be treated as a special kind of multi-class labels. Zhou et al. (2014) adapted their minimax conditional entropy method to infer the ordinal labels.

Multi-label learning for crowdsourcing started with the creation of hierarchies of concepts (Bragg, Mausam, and Weld 2013). In their work, a multi-label naive Bayes (MLNB) model was proposed to infer the true values for all labels. Duan et al. (2014) directly extended the classic Dawid & Shene's model to multi-label scenario. Our work introduces a mixture of multiple independently multinoulli distributions to capture the correlation among both true labels and crowdsourced labels, which improves the accuracy of truth inference (Zhang and Wu 2018). Agnostic inference algorithms for multi-label annotation are likely to become a popular research focus in the future.

Other Techniques for Data Quality

Besides the agnostic inference methods, our machine learning community also has developed some other techniques to improve data quality.

Information Injection Methods

Intuitively, if we know more information, it is possible to infer more accurately. Therefore, many studies have attempted to add more information available during the inference process. Although this violates the unsupervised nature of truth inference, it is more practical to obtain more accurate results for specific applications. This paper does not summarize algorithms that rely on domain knowledge, but rather introduces several domain-independent generic algorithms. We classify them as *information injection* methods.

Tang and Lease (2011) proposed a semi-supervised truth inference algorithm based on Dawid & Skene's model. In addition to the noisy labeled dataset, the algorithm needs another dataset with known true labels involved in inference. Another algorithm ELICE (Khattak and Salleb-Aouissi 2011) also focuses on the optimization of the truth inference by injecting expert labels into the crowdsourced dataset. The injected instances will make estimations of the reliability of workers and the difficulty of instances more accurate. Liu, Ihler, and Steyvers (2013) provided some theoretical results on how many control items should be used under different scenarios, and provided a simple rule of thumb for crowdsourcing practitioners. Some work introduced supplementary information when labeling tasks. Oyama et al. (2013) proposed a method that requires workers to provide their confidence levels when labeling the instances, and these confidence scores are utilized during inference. Not only the information about data can be injected but also does the information about workers. A recent study (Bonald and Combes 2017) shows that if the reliability of a small portion of workers can be known, the reliability of all workers can be accurately inferred, and the lower bound on the minimax estimation error can be calculated. Liu et al. (2017)

introduced a semi-supervised learning algorithm that selects the most informative instances and maximizes the influence of expert labels injected. They developed a complete uncertainty assessment for instance selection. The expert labels are propagated to similar instances via regularized Bayesian inference. Based on our previous GTIC algorithm (Zhang et al. 2016a), our subsequent work (Zhang, Sheng, and Li 2017) demonstrated that the physical features of instances can further improve the accuracy of the inference results of GTIC.

Model-Prediction Methods

Another recent train of thought is to use prediction models to improve data quality. This approach can be carried out before or after label collection.

Mo, Zhong, and Yang (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 knowledge sparsity of a particular domain. Their method is based on a probabilistic graphical model, and its true inference procedure utilizes Markov Chain Monte Carlo and gradient descent algorithms. In a recent paper (Wang et al. 2017), authors use a small number of instances with high-quality labels to train a classification model that can classify unlabeled instances into two classes: *hard* and *easy*. The instances belonging to class *hard* (*easy*) are allocated to the workers with high (low) reliability. Assigning tasks properly between highly reliable and low reliable workers can improve label quality. Abad, Nabi, and Moschitti (2017) proposed a human-machine co-training method before collecting labels. In their work, a prediction model with a acceptable quality is first trained from the data with high-quality labels. Then, the model is used to train workers so that their reliability can be improved. As more difficult instances are labeled and added to the training set, the prediction model is updated and the workers are re-trained. The co-training procedure runs iteratively.

In contrast, our AVNC method (Zhang et al. 2018) builds prediction models after data collection. In this method, after filtering out the instances with noisy labels, the remaining cleansed dataset is used to create multiple weak classifiers, based on which a powerful ensemble classifier is induced to correct errors in the inferred labels. Two other interesting studies (Gaunt, Borsa, and Bachrach 2016; Yin et al. 2017) constructed deep learning models for truth inference.

Model-prediction methods can use the power of machine learning to a large extent, and it is foreseen that this kind of techniques will appear more frequently in the future.

Prediction Model Learning

In this section, we summarize the techniques of building prediction models from crowdsourced data. Because prediction model learning is highly relevant to application domains, here we still only focus on domain-independent techniques.

Supervised Learning

Usually, we can directly use a suitable learning algorithm (such as decision tree, SVM, or neural networks) to build a prediction model from a dataset with inferred labels, which is called a *two-stage* learning scheme. For example, in the early work (Sheng, Provost, and Ipeirotis 2008), a random forest is built after performing majority voting, and in (Raykar et al. 2010; Bi et al. 2014), logistic regression models are built after true labels are inferred.

Using integrated labels to build learning models may lose information, because noisy labels may reflect the uncertainty of an instance belonging to some class. To overcome this shortage, we proposed five label utilization strategies for model learning (Sheng 2011). In this work, we utilized the fact that some learning algorithms such as cost-sensitive decision tree and neural networks can accept weights for training examples. Thus, we generate weights from repeated labels, calculating them as a frequency or the tail of a Beta distribution.

Several methods directly build prediction models from crowdsourced data without truth inference. Kajino et al. proposed two methods Personal Classifier (Kajino, Tsuboi, and Kashima 2012) and Clustered Personal Classifier (Kajino, Tsuboi, and Kashima 2013) to learn logistic regression models with convex optimization. In their work, each labeler is treated as an independent classifier, and all classifiers can be modeled by a multi-task learning paradigm with an objective function that can be globally optimized. Proactive learning (Donmez and Carbonell 2008) is another method that does not include inference, but it merely works under the simple scenario where two labelers are present. In (Sheng 2011), we also proposed a pairwise training strategy, where each instance has a positive and a negative copy with different weights. Learning without inference is an interesting topic that needs further research.

Active Learning

The open, dynamic, and limited budgeting characteristics of crowdsourced labeling make it a natural choice to use the active learning paradigm to build prediction models. Active learning can reduce labeling costs through the design of sample selection strategies that reduce the number of labels required in a learning process.

Our early work (Sheng, Provost, and Ipeirotis 2008) first investigated instance selection strategies in crowdsourcing, proposing the label-uncertainty-based, model-uncertainty-based, and hybrid-uncertainty-based strategies. Yan et al. (2011) believed that a strategy for selecting workers should be added in active learning for crowdsourcing to improve the quality of labels. They proposed a method that selects the labelers that are most beneficial to the performance improvement of the model during the iterative process of active learning. 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, which is essentially using the labels provided by reliable workers to extend the multiple noisy label sets of examples. By adding new reliable labels into the label sets of weak labelers, reliable knowledge can be learned. Rodrigues, Pereira, and Ribeiro (2014) proposed an active learning framework based on a Gaussian process, in which different levels of expertise of workers are modeled, instances and workers are also selected according to their uncertainty and reliability. Long, Hua, and Kapoor (2013) proposed a Bayesian active learning method, which adopts a sorting strategy based on information entropy when selecting instances and workers. Zhong, Tang, and Zhou introduced the labeling confidence into active learning. In our another work (2015), we investigated active learning in biased labeling scenario and proposed three instance selection strategies that incorporate PLAT (Zhang, Wu, and Sheng 2015) and the strategies in (Sheng, Provost, and Ipeirotis 2008). Zhang, Wu, and Shengs (2015) considered an active learning setting where instead of relabeling, one can choose between querying an expert labeler who is more expensive or a noisy labeler who is cheaper. Lin, Mausam, and Weld (2016) tackled the problem of re-active learning, a generalization of active learning that explores the tradeoff between decreasing the noise of the training set via relabeling and increasing the size of the noisy training set by labeling new instances, which introduced two re-active learning algorithms: an extension of uncertainty sampling, and a class of impact sampling algorithms. Huang et al. (2017) observed that it is likely that labelers with a low overall quality can provide accurate labels on some specific instances. Based on this fact, they proposed an active selection criterion to evaluate the cost-effectiveness of instance-labeler pairs, which ensures that the selected instance is helpful for improving the classification model, and meanwhile the selected labeler can provide an accurate label for the instance with a relatively low cost.

Active learning is the most mature learning paradigm under crowdsourcing and will continue to attract enough research attention in the future. It is also highly related to application domains because domain knowledge more or less can help improve the effectiveness of the selection strategies. Domain-dependent crowdsourced active learning techniques are out of the scope of this paper.

Other Learning Paradigms

Besides traditional supervised learning and active learning, several learning paradigms in machine learning have addressed new opportunities and challenges in crowdsourcing circumstances. More and more researchers began to introduce crowdsourcing into different learning paradigms according to their practical application requirements. Because over the past eleven years, general-purpose studies belonging to this category are few, in this section, we just provide some examples to demonstrate the problems and techniques in these studies.

Besides the above mentioned work (Mo, Zhong, and Yang 2013) in which transfer learning can be used for truth inference, transfer learning models can be combined with active learning to provide more plentiful information for instance and labeler selection. Fang, Yin, and Zhu (2013; 2014) proposed a framework that combines knowledge transfer and active learning, where 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. Their method jointly considers the probability distributions of different types of labels in both source and target domains.

In recent years, the rapid development of the deep learning technology has achieved excellent results in many application domains. We were delighted to find that in 2018, there were two studies that combined deep learning with crowdsourcing. Rodrigues and Pereira (2018) addressed the problem of learning deep neural networks from crowds. They first described an EM algorithm for jointly learning the parameters of the network and the reliability of workers, and then proposed a general-purpose crowd layer that can train deep neural networks end-to-end directly from the noisy labels of multiple workers using only backpropagation. Atarashi, Oyama, and Kurihara (2018) addressed the problem of learning from crowdsourced labeled data and unlabeled data under the semi-supervised learning paradigm using deep neural networks. They presented a generative deep learning model in crowdsourcing, which leverages unlabeled data effectively by introducing latent features and data distribution. We believe that there will appear more research work in this direction in the future.

Future Directions

Although in the past eleven years, the research on machine learning with crowdsourcing has made significant progress, there are still some critical issues that have not been well studied. In this section, we enumerate several some of our viewpoints that deserve further study.

Basic Theory of Crowd-sourced Learning

Most of the current research focuses on the *method* (or *algorithm*) level, lacking in-depth discussions of some important basic theoretical issues, for example, the relationship between the performance boundaries of learned models and their various influence factors (variables). Such basic theoretical issues can serve as a guidance in the design of good relabeling schemes, efficient truth inference algorithms, sampling and optimization methods in active prediction model learning.

A possible scheme is to establish models of the relationship among the influence variables through the Statistical-Query Learnable (SQL) theory (Kearns 1998). For example, we have known that the performance of a classification model has its own upper bound, and it can be tolerant with a certain number of mislabeled instances. Therefore, given the instance features, a classification algorithm, and the reliability of workers, the upper boundary of the number of repeated labels per instance could be estimated according to the SQL theory. The extension of the SQL theory in a crowdsourcing scenario is worth further study.

Fine-Grained Truth Inference

At present, our machine learning community has achieved a lot of results in the general-purpose truth inference. However, developing more fine-grained truth inference methods is an exciting direction in many application domains. Compared with the most general-purpose methods and the most domain-dependent methods, it is more interesting to find a trade-off between the two. That is to say, we will introduce more information in true inference to obtain better results, but the information introduced needs to maintain domain (or application) independence as much as possible.

We list three possible research topics here. The first is to utilize features of instances to help inference. We have made our attempts on this topic by introducing prediction model based label noise correction (Zhang et al. 2018) and bi-layer clustering (Zhang, Sheng, and Li 2017) for inference. The second is the time-series modeling of the reliability of workers. As we know, as the number of completed tasks increases, the experience of a worker will gradually increase, and his/her reliability may increase. On the contrary, as the working time increases, the fatigue may also cause his/her reliability to decrease. Some studies (Donmez, Carbonell, and Schneider 2009; Venanzi et al. 2016) have made some efforts on this topic, but it still needs further investigations. The third topic is topic-model-fused truth inference. It is worth investigating the joint representation of a topic model and a truth inference model and optimized solution methods of the joint models.

Multi-Paradigm Model Learning

The most attractive property of crowdsourcing is to harvest the wisdom of the people. We can give the worker more initiative under the active learning framework. Workers can annotate both the class labels and the sample features from multiple aspects, and finally, use a variety of learning paradigms to build predictive models.

It is worth studying the multi-paradigm model learning in crowdsourcing scenarios. These paradigms include, but are not limited to, active learning, heterogeneous ensemble learning, multi-kernel learning (Gönen and Alpaydın 2011), and more. Specifically, we can study the modeling and integration methods of sample feature annotation, feature selection algorithms and feature evaluation functions under crowdsourcing annotation, multi-kernel and heterogeneous ensemble learning methods (including the selection of base classifiers, tuning the number of base learners, and the objective function of ensemble learners), and so on.

Conclusion

Crowdsourcing systems have provided many opportunities for the development and application of machine learning techniques. This paper summarizes the progress during the past eleven years in the field of machine learning with crowdsourcing. We mainly review the variety of techniques in both truth inference and prediction model learning, and classify them into different categories. This research field is still in its young stage with many theoretical, technical, and application problems that are not well solved. We list some of them that are worthy of being further investigated in our opinions with the hope of attracting more attention from the relevant research communities.

Acknowledgments

This work has been supported by the National Natural Science Foundation of China under grants 91846104, 61603186 and 61728205, the Natural Science Foundation of Jiangsu Province, China, under grant BK20160843, the China Postdoctoral Science Foundation under grants 2017T100370, and the National Natural Science Foundation IIS-1115417.

References

Abad, A.; Nabi, M.; and Moschitti, A. 2017. Self-crowdsourcing training for relation extraction. In *ACL*, 518–523.

Allahbakhsh, M.; Benatallah, B.; Ignjatovic, A.; Motahari-Nezhad, H. R.; Bertino, E.; and Dustdar, S. 2013. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing* 17(2):76–81.

Atarashi, K.; Oyama, S.; and Kurihara, M. 2018. Semi-supervised learning from crowds using deep generative models. In *AAAI*, 1555–1562.

Aydin, B. I.; Yilmaz, Y. S.; Li, Y.; Li, Q.; Gao, J.; and Demirbas, M. 2014. Crowdsourcing for multiple-choice question answering. In *AAAI*, 2946–2953.

Bi, W.; Wang, L.; Kwok, J. T.; and Tu, Z. 2014. Learning to predict from crowdsourced data. In *UAI*, 82–91.

Bonald, T., and Combes, R. 2017. A minimax optimal algorithm for crowdsourcing. In *NIPS*, 4355–4363.

Bragg, J.; Mausam; and Weld, D. S. 2013. Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI conference on human computation and crowdsourcing*, 25–33.

Dalvi, N.; Dasgupta, A.; Kumar, R.; and Rastogi, V. 2013. Aggregating crowdsourced binary ratings. In *ACM WWW*, 285–294.

Dawid, A. P., and Skene, A. M. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied statistics* 20–28.

Demartini, G.; Difallah, D. E.; and Cudré-Mauroux, P. 2012. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *ACM WWW*, 469–478.

Deng, J.; Krause, J.; and Li, F.-F. 2013. Fine-grained crowdsourcing for fine-grained recognition. In *CVPR*, 580–587.

Donmez, P., and Carbonell, J. G. 2008. Proactive learning: cost-sensitive active learning with multiple imperfect oracles. In *ACM CIKM*, 619–628.

Donmez, P.; Carbonell, J. G.; and Schneider, J. 2009. Efficiently learning the accuracy of labeling sources for selective sampling. In *ACM SIGKDD*, 259–268.

Duan, L.; Oyama, S.; Sato, H.; and Kurihara, M. 2014. Separate or joint? estimation of multiple labels from crowdsourced annotations. *Expert Systems with Applications* 41(13):5723–5732.

Fang, M.; Zhu, X.; Li, B.; Ding, W.; and Wu, X. 2012. Self-taught active learning from crowds. In *ICDM*, 858–863.

Fang, M.; Yin, J.; and Tao, D. 2014. Active learning for crowdsourcing using knowledge transfer. In *AAAI*, 1809–1815.

Fang, M.; Yin, J.; and Zhu, X. 2013. Knowledge transfer for multi-labeler active learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 273–288.

Gaunt, A.; Borsa, D.; and Bachrach, Y. 2016. Training deep neural nets to aggregate crowdsourced responses. In *UAI*, 242–251.

Gönen, M., and Alpaydın, E. 2011. Multiple kernel learning algorithms. *Journal of machine learning research* 12(Jul):2211–2268.

Huang, S.-J.; Chen, J.-L.; Mu, X.; and Zhou, Z.-H. 2017. Cost-effective active learning from diverse labelers. In *IJCAI*, 1879–1885.

Kajino, H.; Tsuboi, Y.; and Kashima, H. 2012. Convex formulations of learning from crowds. In *AAAI*, 73–79.

Kajino, H.; Tsuboi, Y.; and Kashima, H. 2013. Clustering crowds. In *AAAI*, 1120–1127.

Karger, D. R.; Oh, S.; and Shah, D. 2011. Budget-optimal crowdsourcing using low-rank matrix approximations. In *The 49th Annual Allerton Conference on Communication, Control, and Computing*, 284–291.

Karger, D. R.; Oh, S.; and Shah, D. 2013. Efficient crowdsourcing for multi-class labeling. *ACM SIGMETRICS Performance Evaluation Review* 41(1):81–92.

Kearns, M. 1998. Efficient noise-tolerant learning from statistical queries. *Journal of the ACM (JACM)* 45(6):983–1006.

Khattak, F. K., and 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*, volume 2.

Kim, H.-C., and Ghahramani, Z. 2012. Bayesian classifier combination. In *Artificial Intelligence and Statistics*, 619–627.

Kurve, A.; Miller, D. J.; and Kesidis, G. 2015. Multicategory crowdsourcing accounting for variable task difficulty, worker skill, and worker intention. *IEEE Transactions on Knowledge and Data Engineering* 27(3):794–809.

Lease, M. 2011. On quality control and machine learning in crowdsourcing. In *The 3rd Human Computation Workshop (HCOMP) at AAAI*, 97–102.

Li, Q.; Li, Y.; Gao, J.; Su, L.; Zhao, B.; Demirbas, M.; Fan, W.; and Han, J. 2014. A confidence-aware approach for truth discovery on long-tail data. *Proceedings of the VLDB Endowment* 8(4):425–436.

Lin, C. H.; Mausam, M.; and Weld, D. S. 2016. Re-active learning: Active learning with relabeling. In *AAAI*, 1845–1852.

Liu, M.; Jiang, L.; Liu, J.; Wang, X.; Zhu, J.; and Liu, S. 2017. Improving learning-from-crowds through expert validation. In *IJCAI*, 2329–2336.

Liu, Q.; Ihler, A. T.; and Steyvers, M. 2013. Scoring workers in crowdsourcing: How many control questions are enough? In *NIPS*, 1914–1922.

Liu, Q.; Peng, J.; and Ihler, A. T. 2012. Variational inference for crowdsourcing. In *NIPS*, 692–700.

Long, C.; Hua, G.; and Kapoor, A. 2013. Active visual recognition with expertise estimation in crowdsourcing. In *ICCV*, 3000–3007.

Mo, K.; Zhong, E.; and Yang, Q. 2013. Cross-task crowdsourcing. In *ACM SIGKDD*, 677–685.

Oyama, S.; Baba, Y.; Sakurai, Y.; and Kashima, H. 2013. Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *IJCAI*, 2554–2560.

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

Rodrigues, F., and Pereira, F. 2018. Deep learning from crowds. In *AAAI*, 1611–1618.

Rodrigues, F.; Pereira, F.; and Ribeiro, B. 2014. Gaussian process classification and active learning with multiple annotators. In *ICML*, 433–441.

Sheng, V. S.; Provost, F.; and Ipeirotis, P. G. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *ACM SIGKDD*, 614–622.

Sheng, V. S. 2011. Simple multiple noisy label utilization strategies. In *ICDM*, 635–644.

Tang, W., and Lease, M. 2011. Semi-supervised consensus labeling for crowdsourcing. In *SIGIR workshop on crowdsourcing for information retrieval*, 1–6.

Tian, T., and Zhu, J. 2015a. Max-margin majority voting for learning from crowds. In *NIPS*, 1621–1629.

Tian, T., and Zhu, J. 2015b. Uncovering the latent structures of crowd labeling. In *PAKDD*, 392–404.

Venanzi, M.; Guiver, J.; Kohli, P.; and Jennings, N. R. 2016. Time-sensitive bayesian information aggregation for crowdsourcing systems. *Journal of Artificial Intelligence Research* 56:517–545.

Wang, W.; Guo, X.-Y.; Li, S.-Y.; Jiang, Y.; and Zhou, Z.-H. 2017. Obtaining high-quality label by distinguishing between easy and hard items in crowdsourcing. In *IJCAI*, 2964–2970.

Welinder, P.; Branson, S.; Perona, P.; and Belongie, S. J. 2010. The multidimensional wisdom of crowds. In *NIPS*, 2424–2432.

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

Yan, Y.; Rosales, R.; Fung, G.; and Dy, J. G. 2011. Active learning from crowds. In *ICML*, volume 11, 1161–1168.

Yin, L.; Han, J.; Zhang, W.; and Yu, Y. 2017. Aggregating crowd wisdoms with label-aware autoencoders. In *IJCAI*, 1325–1331.

Zhang, J., and Wu, X. 2018. Multi-label inference for crowdsourcing. In *ACM SIGKDD*, 2738–2747.

Zhang, J.; Sheng, V. S.; Wu, J.; and Wu, X. 2016a. Multi-class ground truth inference in crowdsourcing with clustering. *IEEE Transactions on Knowledge and Data Engineering* 28(4):1080–1085.

Zhang, Y.; Chen, X.; Zhou, D.; and Jordan, M. I. 2016b. Spectral methods meet em: A provably optimal algorithm for crowdsourcing. *Journal of Machine Learning Research* 17(1):3537–3580.

Zhang, J.; Sheng, V. S.; Li, T.; and Wu, X. 2018. Improving crowdsourced label quality using noise correction. *IEEE transactions on neural networks and learning systems* 29(5):1675–1688.

Zhang, J.; Sheng, V. S.; and Li, T. 2017. Label aggregation for crowdsourcing with bi-layer clustering. In *ACM SIGIR*, 921–924.

Zhang, J.; Wu, X.; and Sheng, V. S. 2015. Imbalanced multiple noisy labeling. *IEEE Transactions on Knowledge and Data Engineering* 27(2):489–503.

Zhang, J.; Wu, X.; and Sheng, V. S. 2016. Learning from crowdsourced labeled data: a survey. *Artificial Intelligence Review* 46(4):543–576.

Zhang, J.; Wu, X.; and Shengs, V. S. 2015. Active learning with imbalanced multiple noisy labeling. *IEEE transactions on cybernetics* 45(5):1095–1107.

Zheng, Y.; Li, G.; Li, Y.; Shan, C.; and Cheng, R. 2017. Truth inference in crowdsourcing: is the problem solved? *Proceedings of the VLDB Endowment* 10(5):541–552.

Zhong, J.; Tang, K.; and Zhou, Z.-H. 2015. Active learning from crowds with unsure option. In *IJCAI*, 1061–1068.

Zhou, Y., and He, J. 2016. Crowdsourcing via tensor augmentation and completion. In *IJCAI*, 2435–2441.

Zhou, D.; Basu, S.; Mao, Y.; and Platt, J. C. 2012. Learning from the wisdom of crowds by minimax entropy. In *NIPS*, 2195–2203.

Zhou, D.; Liu, Q.; Platt, J.; and Meek, C. 2014. Aggregating ordinal labels from crowds by minimax conditional entropy. In *ICML*, 262–270.