A Review of Judgment Analysis Algorithms for Crowdsourced Opinions

Sujoy Chatterjee, Anirban Mukhopadhyay, *Senior Member, IEEE*, and
Malay Bhattacharyya, *Member, IEEE*

Abstract—The crowd-powered systems have been shown to be highly successful in the current decade to manage collective contribution of online workers for solving different complex tasks. It can also be used for soliciting opinions from a large set of people working in a distributed manner. Unfortunately, the online community of crowd workers might involve non-experts as opinion providers. As a result, such approaches may give rise to noise making it hard to predict the appropriate (gold) judgment. Judgment analysis is in general a way of learning about human decision from multiple opinions. A spectrum of algorithms has been proposed in the last few decades to address this problem. They are broadly made up of supervised or unsupervised types. However, they have been readdressed in recent years having focus on different strategies for obtaining the gold judgment from crowdsourced opinions, viz., estimating the accuracy of opinions, difficulties of the problem, spammer identification, handling noise, etc. Besides this, investigation of various types of crowdsourced opinions to solve complex real-life problems provide new insights in this domain. In this survey, we provide a comprehensive overview of the judgment analysis problem and some of its novel variants, addressed with different approaches, where the opinions are crowdsourced.

Index Terms—Crowdsourcing, judgment analysis, dependent judgment analysis, constrained judgment analysis, clustering

1 Introduction

The research studies on judgment analysis is not spanking new. In fact its philosophy, in a formal sense, dates back to several decades ago [1], [2]. Initially, this problem was addressed for the applications like policy making and group decision making in a general sense. In the earlier years, there were some real-life problems that were targeted by human beings as judgment analysis tasks. The examples of such daily-life problems include weather prediction, clinical diagnosis, etc. In most of the judgment analysis tasks, there were various issues like selecting proper feedback providers, correlating among them, and statistical estimation of the important parameters regarding the judgment. Judgments of people in the abstract scenarios have also been investigated in the past. In majority of the earlier studies, judgment making on clinical tasks has been studied [1], [2]. The interesting ideas they incorporated are selecting the person's judgment and the environmental criteria. In such approaches, the main factors are linear and non linear relationship between environment and judges, inter-cue agreement, etc.

The pioneering approach by employing the Lens model by E. Brunswik (1952) [1] has been sufficiently studied for the research on human intelligence. Thereafter, a lot of research have been carried out to accelerate the progress in the field of judgment analysis. In 1955, Hammond incorporated the conceptual framework of Brunswik's Lens model to study the clinical judgment analysis [1]. Judgment analysis [3], [4], [5] is extensively used for the type of tasks where we have no prior experience about the ground truth label. So by this judgment analysis, new decision making policies can be made in the abstract situation also.

In some cases, it has been established that bootstrapping can provide better judgment than the individual judgments. This situation happens when the judge is expert enough with valid linear knowledge, and the environment is predictable. In these instances, it is shown that bootstrapping can give better prediction when individual judgments are replaced with a linear model. Unfortunately, in these earlier judgment analysis models, there was no concept of including non-expert self-contributing feedback providers, likewise used in crowdsourcing. It is already established in various complex real-life problems that it is advantageous to obtain public opinions and thus crowdsourcing can help us to solve these problems very efficiently in terms of time and cost. In spite of having numerous benefits, there are multiple challenges to be resolved in order to derive proper and feasible judgment from the crowdsourced opinions. The judgment analysis tasks in this new paradigm pose several additional issues that need research attention toward developing robust models. In this review paper, we study this new dimension

- *S. Chatterjee is with the Université Côte d'Azur, CNRS, I3S, France. E-mail: schatter@i3s.unice.fr.*
- *A. Mukhopadhyay is with the Department of Computer Science & Engineering, University of Kalyani, Nadia, West Bengal 741235, India. E-mail: anirban@klyuniv.ac.in.*
- *M. Bhattacharyya is with the Machine Intelligence Unit, Indian Statistical Institute, Kolkata, West Bengal 700108, India. E-mail: malaybhattacharyya@isical.ac.in.*

Manuscript received 18 Apr. 2018; revised 5 Feb. 2019; accepted 3 Mar. 2019. Date of publication 8 Mar. 2019; date of current version 3 June 2020.
(Corresponding author: Malay Bhattacharyya.)
Recommended for acceptance by L. Chen.
Digital Object Identifier no. 10.1109/TKDE.2019.2904064

of judgment analysis when the opinions are collected from the crowd.

Judgment analysis has gained lots of popularity with the growth of World Wide Web (WWW) because the general people can express their opinions and judgment through the Web media. Let for example, a blog acts as a useful web medium where the users can record their views, opinions in diverse fields related to sports, politics, any new products launched by the companies, etc. As the different opinions registered for a particular topic or any particular product is obtained from a pool of diverse people (who may not be experts or do not have adequate knowledge about the item), the collected data are very much noisy.

The rest of this article is organized as follows. The background details of judgment analysis are introduced in Section 2. Section 3 defines the basic terminologies that have been used throughout and describes the problem formulation. Section 4 represents the types of the different state-of-the-art methods. The categorization of different approaches from different perspectives is described in Section 5. A comparative discussion over several methods is reported in Section 6. Finally, we make conclusion providing a spectrum of future scope of study in Section 7.

2 Basics of Judgment Analysis

For the judgment analysis problem, appropriate definition of the judgment analysis task should be described. The judgment analysis task comprises five important phases. It can be described as follows.

1): *Understanding the judgment of interest* The judgment should be defined clearly and it should be clearly understandable the annotators. The rating scale should also be predefined. The rating scale can be either numeric or categorical. In some cases, let for weather based judgments depending on tweets or financial market prediction based on Twitter opinions, the judgment depends upon some single word sentiment like ‘Yes’, ‘No’, ‘Unsure’. But in some other cases, the judgment can be any numerical value within of some specified range.

2): *Identifying the annotators* Identifying the annotators is one of the most crucial stages in judgment analysis. In different judgment analysis tasks the competent annotators are initially selected through some methods and then that task is distributed among the selected annotators. For the prediction of financial markets, some online annotators are primarily selected to forecast about market. It is advantageous to do some preliminary selection on the annotators so that there is a lesser chance of having some malfunctioning provoked by any user. On the other hand, when the selection is done, it might happen that some incompetent annotators are included whereas some important annotators get excluded. In numerous research problems the judgment analysis task is distributed among the crowd workers without any primary selection to avoid such issues. After collecting the opinions, the filtering of noisy data and the identification of the spammers are carried out.

3): *Context of judgments* When an annotator gives some opinions over a particular question, it should follow some priorly defined assumptions. These assumptions should be followed in the same manner for all the questions and it should be invariant.

4): *Distribution of various options* In this stage, the notation of actual opinion value is defined. Normally, it varies between a predefined numeric range. Otherwise, for categorical cases like the tweet based sentiment analysis, it can have some sentential form containing a particular word that should be highly matched from a dictionary of words. The dichotomous annotators judge the sentiment by setting their opinion within two values (Yes, No) but for any other case multiple values are possible (Yes, No, Unsure, I am sure, etc.). The distribution of opinion values should commensurate with the actual judgment analysis problem but in some abstract scenarios a normal distribution is followed.

5): *Relation between the annotators' opinions* Designing inter-correlation among different annotators is very important to estimate the incompetent annotators. Sometimes, the total correlation is totally unknown. The opinions taken from the diverse pool of opinion providers can be used as an effective tool and business policy for different business organizations when they launch a new product in the market. Then they wait for feedback from different users. The business organizations analyze the data for their business purpose to find some concluding feedback about the product. But as the opinions contain lots of noisy information, so integrating those individual opinions to find a concluding and summarizing judgment is a real challenge.

In this article, we are more interested to investigate the approaches that solve the judgment analysis problem when it is outsourced to the crowd. For this purpose, some basic descriptions of crowdsourcing have been briefly introduced. Crowd-powered systems came into the limelight in the year 2005 by two editors of Wired Magazine, Jef Howe and Mark Robinson. In 2006, Jef Howe formally defined the term crowdsourcing in an article appearing in the same magazine. In 2008, a new direction is given by Brabham [6], when he provided the formulation of how crowd wisdom can be made more effective in solving various real-life problems. Computation in a bounded time might be very challenging for getting a judgment over a particular task. But if these tasks are distributed among the crowd (not necessarily experts) then this problem can be solved easily and in efficient manner. Amazon Mechanical Turk (MTurk) is the first formal crowd-powered platform that is used to distribute simple labeling tasks to hundreds of workers. Thereafter, many crowdsourcing platforms came into the light like OpenIDEO, ioby, Sparked, www.change.org that raise awareness about important causes. Other than these, different crowdfunding companies like Kickstarter, GoFundMe or biotechnology company 23andMe, WikiProjects, etc. have also appeared. Some of these platforms are non-profit companies to brainstorm with real social challenges, whereas, other for-profit organizations are geared up to collect fund for artistic design, alternative capital investment, etc. Again in recent years, some machine learning based crowdsourcing applications in computer vision field have also helped to speed up the annotation task with dramatically improved performance [7], [8], [9]. Before the advent of crowdsourcing, the task has to be completed statically by person at a much lower scale and it was too much time consuming, as well as tedious, expensive and inappropriate.

Crowdsourcing can therefore act as a powerful approach as the collaborative brainstorming of crowd workers makes the task easier to complete in specified time. Many practical applications require to manage the labeling of images belonging to very large databases of image elements. Hence, crowdsourcing can be used to utilize the vast human resources through the internet. The task is totally distributed among the crowd through online and they solve that particular task in parallel. Now, as the crowd workers are not necessarily experts, so there may be some wrong judgment labeled by some annotators. Therefore, it is more obvious that multiple labels per task should be collected.

Many consensus algorithms have already been proposed in recent times [8], [10], [11], [12], [13], [14], [15], [16], [17], [18], [19], [20], [21], [22], [23], [24], [25], [26] to tackle the different types of challenges evolving from the different aspects of crowdsourcing problem. These problems are due to the distributed and anonymous nature of solving the task. Many researchers focus on getting the best consensus method to be used for a given task.

Opinion based judgments problem has been successfully resolved by crowd-based systems [20]. Research in this direction helps in obtaining accurate judgment based on annotation for the image processing, NLP tasks, etc. Consider other type of tasks, for example let there are a number of questions that are to be marked by the crowd annotators. The label marked by an annotator is termed as the opinion and the variation of the opinion can be of three types, namely, 'Yes', 'No', and 'Unsure'. The annotator can assign his/her opinion as any one of the labels taken from this set. Now, if a number of opinions are collected from the crowd workers then it is challenging to retrieve the most accurate solution. Generally, the problem is solved using the majority voting. The majority voting assumes that opinion providers in majority give the final judgment. In case there is a tie between any two options, they are broken by random choice or any other standard tie breaking strategies. Another thing is that, majority voting cannot be applied when the worst case scenario happens, i.e., if the number of opinions taken is small but the possibilities of opinions is higher. To resolve these types of challenges, semantic majority voting approach has been proposed in recent years [27]. Some approaches include Bayesian based probabilistic model to simultaneously measure the worker accuracy and the true answer [28], [29]. Here, features are initially constructed with majority voting and EM algorithm is applied. Then prediction of the true answer is done by using the trained decision tree.

In many NLP tasks the human annotation has been proven to be less time consuming. MTurk has already established itself as a powerful platform for collecting annotation from the crowd workers for various purposes. Luis von Ahn gave the idea of collecting online annotation in the form of a game [9]. ESPGame for image labeling is one of such game. Other than this, a lot of approaches have been made to attract crowd workers to annotate data [10], [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], [41], [42], [43], [44], [45], [46], [47], [48], [49], [50], [51], [52], [53]. In a different work, five different types of tasks: affect recognition, identifying word similarity, recognizing textual entailment, event temporal ordering and word sense disambiguation [20] have been introduced. This study tries to demonstrate whether a non-expert labeler can give reliable judgment. They suggested that training different machine learning algorithms with non-expert labels can lead to good annotation. In addition to that, they try to correct the annotator bias to improve the quality of annotator's judgment.

It has been seen that MTurk is frequently used for getting annotations for the NLP datasets, computer vision tasks, etc. Unfortunately, some workers try to maximize their (profit) by choosing random answers that might not be related to the true label. A major limitation of obtaining opinions through a crowdsourcing platform is that we do not have any control over the quality of annotators. They are usually from a pool of crowd workers. The skill diversity in crowd is because of the combined presence of genuine experts, novices, biased annotators and also spammers. This skill diversity of the crowd annotators causes difficulty in deriving the actual consensus. Therefore, identifying the spammers and removing them is one of the major challenges in the judgment analysis task. In [54] an algorithm to identify spammers and excluding them has been proposed. It provides an approach to consolidate annotation that eliminates spammers automatically. It simultaneously estimates the consensus ground truth. This type of approach can also be extended easily to handle the missing annotations (which is a more realistic scenario). It has been observed that the algorithm (SpEM) [54] shows a better accuracy with respect to AUC and accuracy than majority voting and EM based approaches.

Some recent studies have also been done in an unsupervised manner to identify the reliable annotators and as well as the spammers. These attempts are based on item-response model. Multi Annotator Competence Estimation (MACE) is an unsupervised model that simultaneously estimates the trustworthy annotators using some probabilistic rules [55]. Here, posterior entropy is considered for the identification of spammers and prediction of ultimate label is achieved with better accuracy.

In many major Web search engines, training of search rankers is done using human rating based on the query pairs. To rely on the crowdsourced rating, they need to continuously train and monitor the human judges. As these judges are well trained, rating a task is economically very costly. Hence, researchers try to control the quality of judgment by introducing a set of surveillance on the judges' work secretly. A probabilistic solution is discussed to model the confusion of their answers and it is basically a generalized version of the work introduced in [56].

As described earlier, MTurk is a powerful platform to experiment the different crowdsourcing tasks that aim to annotate data in exchange of payments. However some annotators unfortunately try to earn quickly by choosing a random label. These labelers are called the spammers. Many researchers focus on the area of how to identify the spammer and adopt some filtration criteria in order to achieve the best and consistent consensus. So to reconstruct the correct label, the most trusted label should be chosen. To address this problem, generally unsupervised model called item-response model is chosen [57]. Among all the proposed models, Sheshardi et al. [19] provides a benchmarking study for the researchers in judgment analysis community.

3 Basic Terminologies and Formal Representation

Let us define some basic terminologies that will be used throughout the paper.

Definition 3.1 (Annotator). *Annotator is the crowd worker from the online community who gives his judgment over a particular question.*

It is basically the crowd worker who gives the opinion by choosing a label. Annotation is a function that returns one of the available labels according to some procedure. Better annotators have a smaller chance of guessing a label at random.

Definition 3.2 (Opinion). *An annotation marked by an annotator.*

There is a finite set of annotations for a given question.

Definition 3.3 (Domain of Opinion). *The possible values of opinions.*

For example, let there be three types of opinions. These opinions are 'Yes', 'No' and 'Unsure'. So here the annotator should choose any single option from the opinion set. So the domain of the opinion is {'Yes', 'No', 'Unsure'} and here the cardinality of the domain is three.

Definition 3.4 (Gold judgment). *The true label for each question.*

If the questions are distributed among the crowd soliciting answers then the aggregation is done after obtaining the answers from them. Aggregating these answers predicts the gold judgment and the deviation from the true label can be computed.

Definition 3.5 (Item difficulty). *The item (or question) difficulty is the hardness or toughness of the item (or question).*

Any ambiguity in the question can also be defined in terms of the item difficulties.

Let us consider an annotation process containing n questions. Each question has two possible options (Yes/ No). There are m annotators who decide the judgment (Yes/No) for each question.

The difficulty of an item (or question) j is represented as $1/B_j \ \in [0, \infty]$, where B_j is constrained to be positive. Here $1/B_j = \infty$ means the question has high ambiguity so the competent annotator has minimal chance of labeling it properly and $1/B_j = 0$ means the question is so easy that even the worst annotator will always label it correctly.

Definition 3.6 (Worker accuracy). *The accuracy of worker means how reliable the worker is.*

A competent annotator provides the accurate judgment whereas an unskilled worker provides some inconsistent judgment. Different models calculate the worker accuracy in different ways. In some algorithms, the worker accuracy is calculated by confusion matrix or any other ways like F1 score, by setting likelihood or sensitivity, etc.

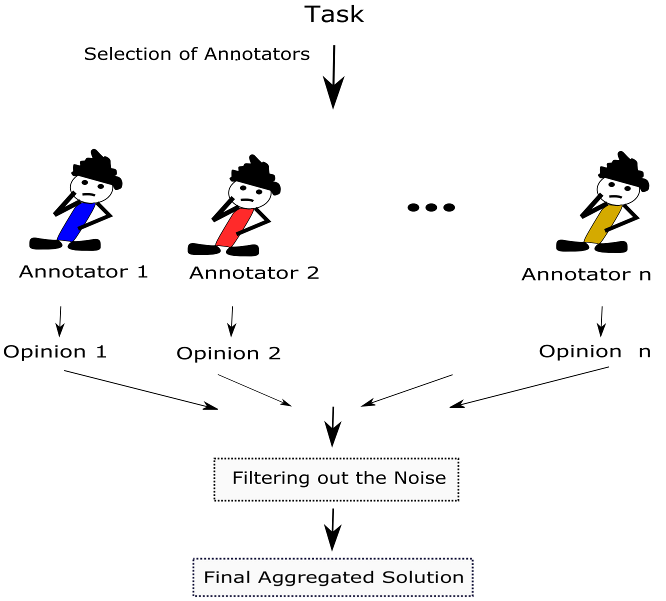


Fig. 1. Different stages of the judgment analysis process.

Here, the terms question, problem, task, and item are used synonymously. On the other hand, annotator and worker are also same.

Definition 3.7 (Multidimensional annotator). *The annotator whose ability can be modeled with multidimensional entity.*

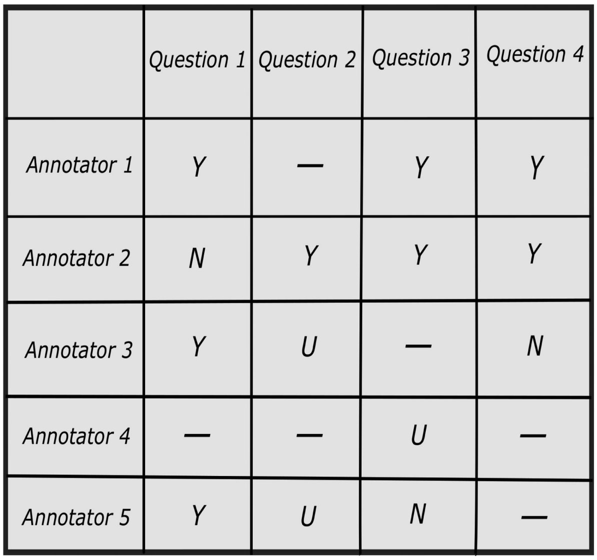
All the annotators cannot be equally efficient with respect to the entire relevant task. The ability of an annotator can be multidimensional. That means an annotator can be modeled with multidimensional entity with variables representing competence, expertise, biasness, etc. An annotator may be good in some task but not in others.

Definition 3.8 (Adversarial labeler). *The expertise of each annotator i is modeled by the parameter $\alpha_i \in (-\infty, \infty)$. Here, $\alpha_i = +\infty$ means that the annotator always annotate the images correctly and $\alpha_i = -\infty$ means the annotator always annotates the images incorrectly. He can distinguish between the two classes perfectly but always inverts the label, in the latter case either maliciously or because of a consistent misunderstanding. If in any case $\alpha_i \ < \ 0$ then the labeler is said to be adversarial.*

Definition 3.9 (Spammer). *The low quality annotators who provide random opinions that are within the option range but deviate from the true label by a significant amount.*

Generally, annotators tag the item or question by a certain label that varies within some range. Spammers can be treated as low quality annotators who provide random opinions. It may be due to the reasons that the annotator does not understand the labeling range, does not consider the instances, or pretending as an expert for making quick income. If the presence of number of spammers becomes too large then the total task of collecting the information becomes costly (as the annotators are to be paid for making their contribution).

Now let us formally present the problem of basic judgment analysis (as shown in Fig. 1) of crowdsourced opinions.



	Question 1	Question 2	Question 3	Question 4
Annotator 1	Y	—	Y	Y
Annotator 2	N	Y	Y	Y
Annotator 3	Y	U	—	N
Annotator 4	—	—	U	—
Annotator 5	Y	U	N	—

Fig. 2. A subset of the annotators giving their opinions over a particular set of questions. 'Y' denotes Yes, 'N' denotes No, 'U' denotes Unsure, and '-' denotes did not attempt the question.

In the crowd based judgment analysis model, each annotator is indexed as a_i where i is called crowdworker ID. Let the question instance is denoted by q_i (i.e., ith question) and opinion instance is denoted by o_i.

An annotation process can be formally represented as a quadruplet (Q, A, O, τ) that consists of the following:

1) A finite set of annotators $A = \{a_1, a_2, \ldots, a_n\}$,
2) A finite set of questions $Q = \{q_1, q_2, \ldots, q_m\}$,
3) A finite set of opinions $O = \{o_1, o_2, \ldots, o_k\}$,

where $\tau : Q \times A \rightarrow O$ is a mapping function.

As illustrated earlier, in the judgment analysis problem based on crowdsourcing the total environment comprises two entities. One of them is the requester and the other is the set of crowd workers or annotators. A single requester asks the annotators to provide the judgment about some tasks assigned to them. It may so happen that for a single task multiple annotators are giving their opinions simultaneously. Note that, all the annotators are not required to judge all the tasks. This leads to a sparse matrix of opinions for the given set of tasks. In Fig. 2, the snapshot of a particular instance of some annotators giving their opinions over a particular task is shown. Here, the number of options is three, i.e., Yes, No and Unsure. These opinion values have been abbreviated as Y (Yes), N (No) and U (Unsure). We can see that Annotator 4 provides an opinion over the Question 3 only. For Question 1, as there are four opinions the requester has to decide what will be the ultimate judgment for that particular question. Now this opinion set may be different in different scenarios. Even the framework of addressing the annotation problem with crowdsourcing could be different. Therefore, the detailed formulation according to the nature of collecting opinions is required to be changed in different real-life applications appropriately.

4 Types of Judgment Analysis Problems

This review is a comprehensive attempt to evaluate suitable methods for aggregating various types of crowd opinions collected aiming a wide range of applications. The current state-of-the-art approaches deal with the basics of judgment analysis and how analyzing them is useful for solving various real-life problems. These various methodologies can be catergorized into five judgment analysis problems depending on the variety of crowd opinions. These five models are outlined below.

4.1 Judgment on Independent Crowd Opinions

Crowdsourcing a problem and collecting opinions from the general crowd has already been proven to be useful for annotating large-scale datasets for solving numerous real-life problems. To tackle different issues evolved from crowd responses, various methods have already been proposed to aggregate multiple noisy crowd opinions with an objective to obtain consensus judgment [3], [4], [33], [55], [58], [59], [60], [61]. But in some recent studies [3], [4], the problem of aggregating multiple crowd opinions (binary/multiple) are studied from a different angle. In the crowdsourced annotation process it is observed that there are some annotators who are interested to respond certain set of questions. Therefore, the annotator set can be seggregated depending on the attempted questions. In a recent model [3], [4], biclustering approach is used to retrieve similar set of annotators based on similar set of questions. It is occasionally seen numerous annotators attempt a large set of question with little expertise just to gain extra remuneration. To get rid of noisy annotations, a proper aggregation method is proposed to find out noise-free final judgment. Moreover, as the response matrix is very sparse, therefore, probabilistic matrix factorization can play a vital role to predict the missing annotation in deriving better judgment. This intuition has been shown to be effective in another work [60].

4.2 Judgment on Large-Scale Ambiguous Crowd Opinions

It has been already mentioned that the identification of spammers in the annotation process has immense effect in judgment analysis research. In a crowdsourced setting, there are some tweet sentiment analysis tasks that demand for aggregating crowdsourced opinions having some ambiguous options. For example, in the previous paragraph the datasets discussed are basically having binary opinions like 'Yes' or 'No'. But in this type of tweet sentiment analysis task, the datasets contain different ambiguous options like 'Skip', 'I can't tell' and 'Unsure'. Treating these special options in a normal way may lead to a noisy final judgment. Moreover, there are some semantic meaning of the options in these large-scale datasets. The distributions of different options are also highly unbalanced in these datasets that make the problem more tougher and traditional way of aggregation is not suitable for such cases. As the datasets are highly unbalanced, therefore, annotator biasness has a major role in predicting the final judgment. For example, if an annotator randomly provides all his answers as 'Yes' and eventually all the questions have 'Yes' as ground truth then a simple accuracy measure cannot capture the expertise of the annotator. On the other hand, expertise depending upon the difficulty of questions can have a major role to play.

Among the models that tackle these types of complex opinions, probabilistic graphical model based approaches

[28], [29] are found to be very effective. These methods take care of the different parameters like annotator accuracy, question difficulty and annotator biasness to find the final aggregated judgment. These models are applied on two benchmark datasets namely, CrowdFlower and Google dataset, and, the performance is studied. Again, formulation of the proposed model is shown on a relatively different domain of group decision making in another recent work [62].

4.3 Judgment on Dependent Crowd Opinions

The models discussed so far (in previous subsections) basically deal with the independent opinions of the crowd workers. In this setting, there is no chance that a crowd worker can see others' opinions before posting his own opinion. So there is hardly any possibility of inclusion of bias due to the influence by other annotators. But it is seen in numerous real-life applications that the opinions of an annotator remain open to all and one can see other's opinions before posting their opinions. So these type of opinions can be basically treated as dependent opinions and aggregating these type of opinions gives birth of a new direction of judgment analysis problem, termed as Dependent Judgment Analysis problem. The state-of-the-art approaches mainly deal with the independent opinions of the crowd workers, and very limited study has been done that account for dependent opinions [63], [64]. This relevant work is on debiasing task-dependent worker bias and the aim is to model the sequential dependency among worker judgments [63]. MicroTalk is a novel framework that demonstrates debate among the workers instead of hiding their opinions from each other can improve their performance significantly [64]. Therefore, the workers are allowed to reconsider their decisions. Recently, a crowdsourcing model has been proposed that outsources several questions to the crowd workers and obtain the opinions from them in independent as well as dependent manner [65], [66]. Furthermore, a Markov chain based method is proposed to reach into a consensus from a set of independent and dependent crowd opinions. The theoretical proof of convergence of the Markov chain based model followed by discussions on some other variants of the problems are also provided.

4.4 Judgment on Constrained Crowd Opinions

While the previous models dealt with independent opinions and dependent opinions, Chatterjee et al. has recently proposed a different crowdsourcing model that tackles constrained opinions of the crowd workers [67]. In most of the opinion aggregation models, the questions have single component and thus the opinions are unidimensional, either 'Yes', 'No', etc. But in our daily experiences there are so many artificial intelligence problems that comprise some questions having multiple components. For example, in smart city planning, it is often needed to seek public opinions from the crowd to locate/install different facilities in a place. Now effective planning demands that there should be some specific distance between any pair of facilities. This is basically a constraint that is needed to be maintained by the crowd workers while attempting the questions. Although the state-of-the-art approaches deal with the problem of independent crowd opinions, no in-depth study has yet been conducted on analyzing constrained opinions of the crowd workers. In this context, a new judgment analysis problem is introduced that yields another research direction of judgment analysis [67], termed as 'Constrained Judgment Analysis'. it is also shown how a variety of real-life problems including smart-city problem can be solved with this approach [68].

4.5 Judgment on Streaming Crowd Opinions

In the previous models, it is described how aggregated judgment can be improved by introducing different opinion collection frameworks as well as suitable opinion aggregation methods. These opinions are basically independent binary opinions, independent ambiguous complex opinions, and dependent opinions. In these frameworks, the number of crowd opinions are fixed after collecting their opinions. But there are various applications that require to provide judgment over streaming crowd opinions. In sensor network and sensor-based room occupancy inference system the annotation over straming data is very necessary. A recently published research provides a parallel and streaming algorithm to capture relevant judgment from large-scale streaming crowd opinions [69].

Besides the study of different crowd opinion aggregation frameworks, this introduces a new perspective of looking at the judgment analysis problem. It bascially discusss that how quality of judgment can be improved by refining the response matrix initially. Finally, the judgment is infered from the refined response matrix. The refinement can also be done using matrix factorization techniques over the response matrix to predict the missing emtries. In this ontext, probabilistic matrix factorization is found to be very effective [60], [70]. On the other hand, it is also shown that aggregation of different rankings based on multiple features (rather than a single feature) can be very much helpful for crowd judgment analysis. To accomplish this, a novel rank aggregation method is developed and the performance of it with other state-of-the-art rank aggregation methods [71] is compared. Finally, how the judgment analysis can be improved is demonstrated by applying it on different MTurk datasets. The type of various judgment analysis methods are shown altogether in Fig. 3.

5 Approaches to Judgment Analysis

The judgment analysis problem is not a recently emerging topic, rather the philosophy of the problem was introduced in a very different context in 1979 [72]. It was basically the first formalized study on this topic. In this work, for a fixed set of patients repeated clinical judgments were taken for different time instants and then the final clinical judgment is concluded from them. Later this problem has been posed in the same context but the major difference is that repeated judgment from the same annotator for a single question is not taken. In a previous study, estimating an error rate of an individual observer was the main goal, but later the error rate has been analyzed extensively [72]. In the crowd based judgment analysis problem, there are some algorithms that solve the particular problem in supervised manner where the ground truth labels are known to the requester. In some cases, there is no prior knowledge about the dataset, hence these are to be

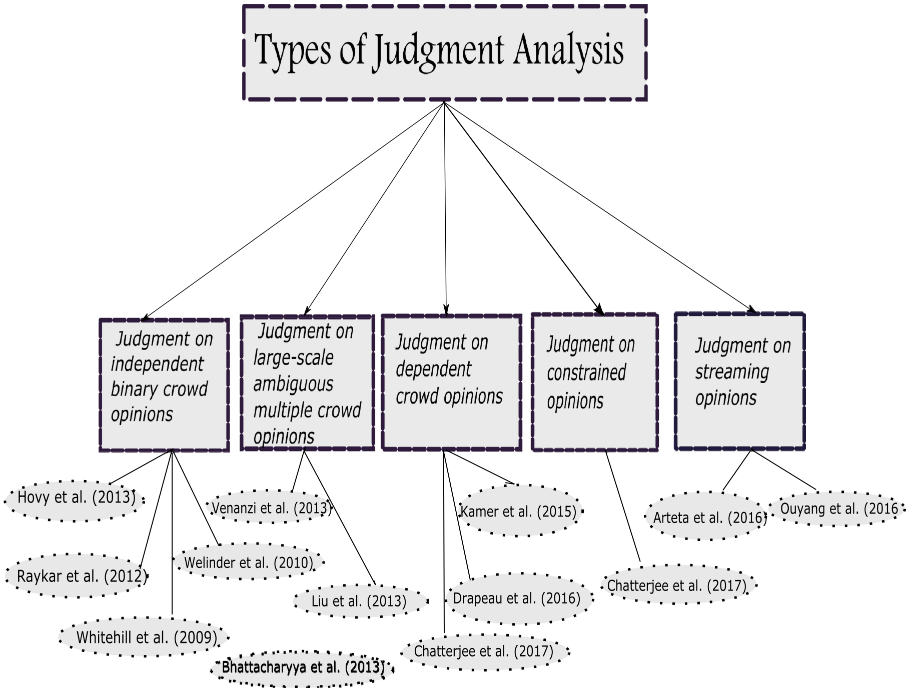

Types of Judgment Analysis
Judgment on independent binary crowd opinions
Judgment on large-scale ambiguous multiple crowd opinions
Judgment on dependent crowd opinions
Judgment on constrained opinions
Judgment on streaming opinions
Hovy et al. (2013)
Raykar et al. (2012)
Whitehill et al. (2009)
Welinder et al. (2010)
Venanzi et al. (2013)
Liu et al. (2013)
Bhattacharyya et al. (2013)
Kamer et al. (2015)
Drapeau et al. (2016)
Chatterjee et al. (2017)
Chatterjee et al. (2017)
Arteta et al. (2016)
Ouyang et al. (2016)

Fig. 3. Different directions of judgment analysis research.

solved in an unsupervised manner. Again, as the total task is outsourced to the crowd worker, there is a high probability to have some non-expert crowd workers who will also give their judgment to that particular question. So here lies the responsibility of the requester to select the best consensus when multiple opinions are collected from several annotators for a particular question. Researchers have investigated several factors (refer to Fig. 4) that are very much responsible to improve the performance of the entire system. Hence, various methodologies have been proposed over the years in order to tackle this problem from different angles (as shown in Fig. 5). It can be seen from Fig. 5 that initially the utility of human intelligence in cheap and faster way was observed in different applications like image annotation, object recognition, etc. Motivated by DAWIDSKENE model, various probabilistic models using Bayes rule proposed the solution to infer the final judgment from multiple noisy opinions. Thereafter, probabilistic graphical model based approaches are found to be effective in deriving better judgment. Mainly in these methods, EM algorithm, variational Bayes, etc. are employed to infer the most probable judgment. However, till 2015, most of the approaches treat all the annotations as independent types and these models focus over some typical characteristics of the datasets. Thereafter, the dependency among tasks along with the annotators opinions are being observed. Therefore, advanced models dealing with the task dependency and annotators opinions are proposed. Moreover, very recently, deep learning revolutionized this area by introducing some generalized frameworks applicable for many domains and these are merely dependent on any special characteristics of datasets. In addition to that, these deep learning based models overcome the additional computational burden evolved in the EM style approaches while handling large-scale datasets.

Here, we introduce a spectrum of influential approaches in judgment analysis that infer better judgment from multiple noisy crowd opinions. These approaches can be classified from different perspectives and broadly categorized into several groups depending upon their working principle. Since the last decade, several works on judgment analysis over crowdsourced opinions have emerged and the overall aim in most of the approaches is to reduce noise in the final judgment.

5.1 Majority Voting Approaches

Majority voting is considered to be the simplest and straightforward way of deriving final judgment from multiple opinions. In crowdsourcing domains there is a high possiblilty of existence of spammers. Unfortunately, majority voting cannot track these spammers effificiently as all anntoators are considered having same expertise level.

5.2 Focusing on Item Difficulty

A common practice in crowdsourcing for finding better judgment is to solicit multiple feedback from crowd workers for same question. To estimate the worker ability and final judgment, the knowledge over inherent difficulty of questions has become necessary [58]. In some works it is proposed that workers annotation cannot be modeled independently without considering the question difficulty. GLAD (Generative model of labels, Abilities and Difficulties), proposed in this work, is a probabilistic method [73] that uses some standard inference tools [58]. This inference method is used to simultaneously infer the expertise of each labeler, the difficulty of each image and the most probable label of the image. It demonstrates more robustness to both adversarial and noisy labelers when applied to the synthetic and real-life datasets. This is the first model to simultaneously estimate the true label, item difficulty and coder expertise in an unsupervised and efficient manner.

So the earlier models integrated the probability, that an annotator would rate an item correctly, as a logistic function

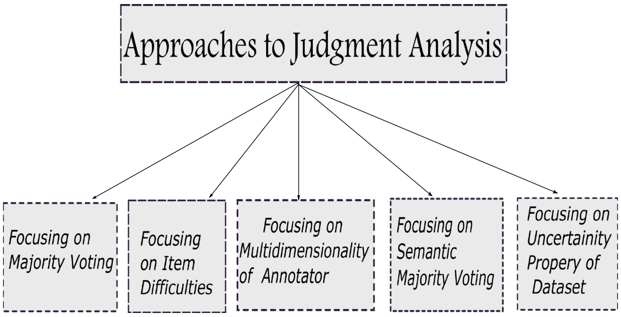

Approaches to Judgment Analysis
Focusing on Majority Voting
Focusing on Item Difficulties
Focusing on Multidimensionality of Annotator
Focusing on Semantic Majority Voting
Focusing on Uncertainity Propery of Dataset

Fig. 4. Different research methodologies for judgment analysis problem.

of the product between quality of the annotator and difficulty of the item. Another recent model not only rediscovers the true label but estimates the judge's quality and item difficulty at the same time [56]. This approach focuses on models under the "True label + confusion" paradigm or diagnostic insights into judges' confusion. The proposed models, SINGLECONFUSION and HYBRIDCONFUSION (Hierarchical Bayesian model), remove the overfitting drawback of the well-known DAWIDSKENE model [72] under the same paradigm. This approach assumes that every annotator has a confusion matrix when considering multiple options (Yes, No and Uncertain) instead of two (Yes and No). Each element of the confusion matrix defines the probability value that how much the opinion of an annotator matches with the gold judgment. The ultimate goal is to prepare a robust and generalized model that is applicable toward any kind of human opinions and for devising an appropriate prediction model. Most of these models are based on Bayesian Statistical approaches that employ the Markov Chain Monte Carlo (MCMC) sampling techniques and are computationally expensive. Nevertheless, the works discussed above consider a passive learning where the data is assumed to be trained apriori. Therefore, to fully understand the dynamic behaviour of crowd, active learning over the systems is necessary. While numerous methods have been developed to aggregate the opinions of crowd workers, limited works are available that consider the sequence of crowd annotations [74].

Subsequently, a line of theoretical research has been proposed to estimate how many questions are considered to be sufficient in order to produce accurate judgment [75], [76]. Obviously, the number of questions solicited from the crowd workers cannot be same for tough question as well as for easy question. Perhaps it is natural to ask as many people as possible when the questions are estimated as very hard. On the other hand, another popular technique is using trap question in order to track the spammers. In this work, the theoretical proof over the count of optimal number of questions to be asked to crowd workers is provided [75], [76].

5.3 Focusing on Multidimensionality of Annotators

Inspired by the work in [58], modelling each annotator as a multidimensional entity has proven to be effective while deriving the ground truth label. Mainly, the annotator accuracy, question difficulty and annotator bias in multidimensional decision surface are considered as important parameters, while annotator bias was not treated as an input parameter in [58]. The performance of this model is demonstrated on two real-life datasets namely, Greeble and Waterbird datasets. However, this sophisticated crowdsourcing approach is more specific to image classification task. So, there is a scope of improvement to devise tools applicable toward other types of tasks.

5.4 Semantic Majority Voting Approaches

Different recent studies have highlighted how the different aspects related to crowd based judgment analysis can be addressed by considering the accuracy of the annotators, their probabilistic judgment, noise factor, etc. Again, recent research have shown how the total consensus can be influenced if the simple concept of majority voting is employed to integrate the individual solutions. Moreover, the consensus can get affected if the number of questions is limited but the options are more than two [27]. To mitigate this problem, a new approach has been incorporated in the majority voting in a very recent study to break the tie between the positive and uncertain answer (Yes and Unsure) and as well as negative

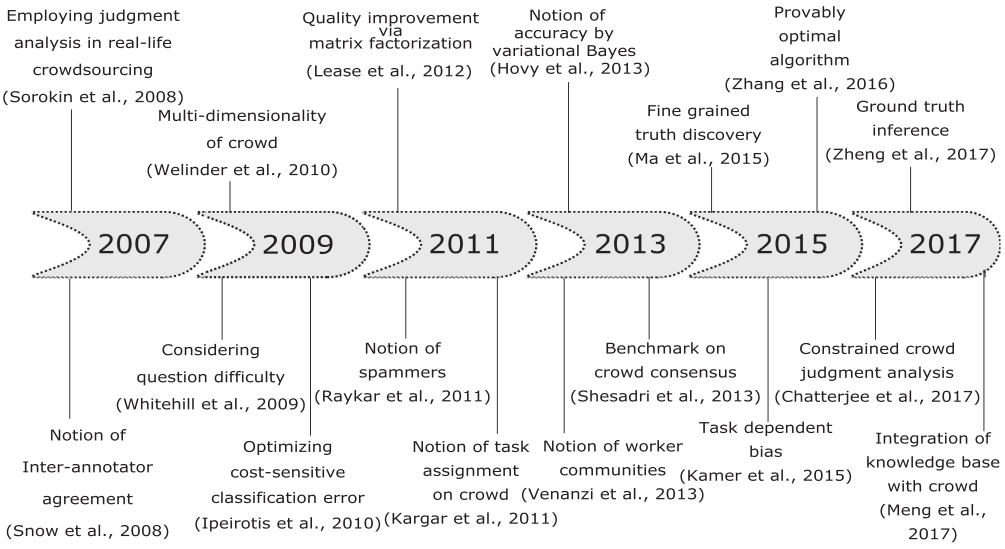

Employing judgment analysis in real-life crowdsourcing (Sorokin et al., 2008)
Multi-dimensionality of crowd (Welinder et al., 2010)
Quality improvement via matrix factorization (Lease et al., 2012)
Notion of accuracy by variational Bayes (Hovy et al., 2013)
Fine grained truth discovery (Ma et al., 2015)
Provably optimal algorithm (Zhang et al., 2016)
Ground truth inference (Zheng et al., 2017)
2007
2009
2011
2013
2015
2017
Notion of Inter-annotator agreement (Snow et al., 2008)
Considering question difficulty (Whitehill et al., 2009)
Optimizing cost-sensitive classification error (Ipeirotis et al., 2010)
Notion of spammers (Raykar et al., 2011)
Notion of task assignment on crowd (Kargar et al., 2011)
Notion of worker communities (Venanzi et al., 2013)
Benchmark on crowd consensus (Shesadri et al., 2013)
Task dependent bias (Kamer et al., 2015)
Constrained crowd judgment analysis (Chatterjee et al., 2017)
Integration of knowledge base with crowd (Meng et al., 2017)

Fig. 5. Evolvment of judgment analysis methods for crowdsourced opinions over the years (only breakthrough papers are highlighted).

and uncertain answer (No and Unsure). Here the approach is to revise the individual solutions (that are collected from different crowd workers) using the semantic majority voting in an unsupervised manner. As no gold level is present in unsupervised cases, each annotator's accuracy is computed by comparing their individual opinions with the opinion that is computed by semantic majority voting. The effectiveness of this approach has been experimented on the basis of precision and average recall. But this study neglects noising and worker bias in the proposed model. In another recent work, an optimal Bayesian voting strategy is proposed to solve the Jury Selection Problem (JSP) [78]. This study solves this problem by an efficient approximation algorithm based on simulated annealing. In this context, the computation of accuracy for different annotators is usually done through the confusion matrix. However, as there is a common problem regarding data sparsity in the crowdsourcing domain, therefore it becomes hard to properly estimate the confusion matrix for some workers who attempts (annotate) very limited questions. In a few other cases, the topic-aware accuracies play a vital role. In these scenarios, a crowd worker may be good with respect to a particular topic, whereas bad in the other [48]. So, a graph based similarity matrix among the microtasks is constructed to probabilistically estimate the accuracies in a better way. For this purpose, Jaccard similarity and Latent Dirichlet Allocation (LDA) are used for the computation of task similarity and topic distribution.

5.5 Focusing on Uncertainity Property of Dataset

The uncertainity in the collected information from crowd causes several problems in producing noise-free judgment. Scarcity of opinions provided by crowd, presense of ambiguous options in option set, and unbalanced property of possible options in the dataset, etc. are the challenging issue in judgment analysis. In this context, consideration of features, ranking of crowd workers based on their opinions, etc. are employed to yield better judgment.

The unbalanced labeling in the dataset may obstruct to achieve the optimal consensus label from the opinions on an individual task. Suppose there are three different opinion values, let Yes, No, and Skip. Then for a particular dataset, if 90 percent of the questions has positive answer and 10 percent has negative answer, then this unbalanced property may create difficulty in preparing a learning model that can predict the answer. This uncertainty has been taken into consideration for designing algorithms [29]. For different datasets different methods have been adopted for finding gold label from the annotators' decisions. In this context, the value of opinion can be "Yes", "No" and "Skip". In addition to that, the special label "Skip" and "I can't tell" indicate the workers uncertainty about the answers. Treating this as general may cause poor performance. These methods apply some trial and error approaches for the removal of these challenges. But no noise removal technique is adopted to filter out the presence of any spammer [29].

In the study by Venanzi et al. [29], the judgment analysis is done based on some tweet sentiments and estimation of the true value is carried out by considering the possible noise and bias in the opinions. Here, the characteristics of text content (basically the tweets) are also taken into consideration while predicting the actual true label. In this approach, Bayesian model is adopted to infer the true sentiment of the tweets by combining signals from both the crowd labels and words in the tweets. Venanzi et al. have presented the reliability of each annotator using a confusion matrix model. The likelihood of each dictionary word belongs to certain sentiment class using a mixture of bag of words model. Here, the factor graph of the Bayesian Combination Model (referred to as BCCWords) is used and this model can be improved in future because common pattern of each worker's behavior can be exploited for a faster learning over sparse data. In these approaches, aggregation is done after a batch of workers has completed their annotations. Therefore, it cannot dynamically prevent the poor quality workers while they are active in the system (and change their strategies over time). Additionally, these confusion-based techniques does not consider the spatial distribution of each class, so it cannot model properly when the data is sparsely distributed over an area of interest. However, a few methods have recently been proposed to consider these factors [77].

5.6 Non-Iterative Approaches

On the other hand, all the methods can also be classified into the two broad categories i.e., (i) non-iterative and (ii) iterative. In the non-iterative methods, the elimination of the spammers are performed in the preprocessing steps, whereas it can happen sequentially in the iterative algorithms. The discussions over the two categories are provided here.

Among the non-iterative approaches, majority voting, Honeypot project (HP) [41], ELICE model [79], etc. are very popular ones. As majority voting treats all the workers with same expertise, therefore, there is no refinement stage in it. However, the refinement is done in rest of the approaches. The difference between the two models is that spammer identification happens in advance in HP, on the contrary, in ELICE, it is done by considering worker accuracy and question difficulty simultaneously. In both HP and ELICE, the spammer opinions are identified by replying on some trap questions. Basically, the ground truth answers are known for these trap questions. However, some expert annotators can be misinterpreted as spammers if these trap questions become too difficult to answer and these are the few drawback of these methods.

5.7 Iterative Approaches

Iterative algorithms compute the final judgment with several sequences of estimation and updation step. In these methods, mainly the preprocessing step is not used, rather, the spammers are eliminated estimating the posterior probability of true option and updating different parameters like accuracy, question difficulty, bias, and confidence, etc. These processes iterate until the convergence occurs. Probabilistic graphical model [29], [54], EM based methods [55], [59], GLAD [58], ITER [80], etc. belong to this category and are widely used for judgment analysis. These works primarily focus on recognizing unreliable workers depending upon their annotations and then eliminate them. Addressing the adversarial nature of the workers is another important concern because the adversarial workers can change their strategy in the midst of providing opinions. However, none of these models considered this [54], [55]. Recently, an interesting study addresses this issue to identify adversarial workers

TABLE 1
Various Characteristics based on Nature of the Models Considered in Different Major Algorithms

Algorithm	Working principle	Basic model type	Applications	Noise handling
Bhattacharyya et al. [27]	Supervised	Semantic majority voting	General	No
Liu et al. [28]	–	EM based construction of feature with decision tree	Synthetic and real unbalanced data	No
Venanzi et al. [29]	–	Bayesian model with mixture of bag of word model	Tweet sentiment analysis	–
Raykar et al. [54]	–	Bayesian model	General	Yes (by spammer filtering)
Whitehill et al. [58]	Unsupervised	GLAD model based on IRT mode	Computer vision	Yes (adversarial labeler considered)
Welinder et al. [59]	–	Bayesian joint probability distribution	Computer vision	Yes
Dawid et al. [72]	Both	EM algorithm based estimation with individual error rate consideration	Clinical judgment	Yes
Ye et al. [89]	Unsupervised	Hybrid system with maximum a posteriori estimation taking consideration absolute and preference judgment	Images taken from LIVEIQA Dataset	Yes
Snow et al. [20]	Unsupervised	Bias correction on non-expert annotators and estimating worker response likelihood with laplace smoothing	NLP tasks	No
Xu et al. [82]	Unsupervised	Parsimonious mixed-effects model based on HodgeRank estimation	Movie preference prediction, Image quality assessment, Ranking	Yes
Li et al. [91]	Unsupervised	Optimization model (CATD) based on confidence interval of source reliability	Wikipedia history source on city population	–
Zhou et al. [92]	–	Optimization model based on Minimax Principle	Image labeling Web search relevance judging	Yes
Liu et al. [93]	–	Belief propagation and Mean field method	Image labeling NLP datasets	Yes
Li et al. [94]	–	Optimization model (CRH) for truth discovery from heterogenious data	Weather prediction data stock prediction data	Yes
Zheng et al. [78]	Unsupervised	Optimal jury selection based on Bayesian voting	Real-life data on Jury selection	Yes
Yin et al. [83]	Unsupervised	Deep learning based model using label aware auto-encoder	Image data and Web search data	Yes
Zhang et al. [95]	–	Path selection using crowdsourcing	real-world road network of San Francisco and California	Yes
Zhong et al. [96]	–	A support vector machine based active learning approach with unsure option	UCI Machine repository dataset Crowdsourcing dataset	Yes
Zhuo et al. [97]	Unsupervised	A constrained knowledge acquisition model (CAMA)	Dataset containing plan traces	Yes
Shan et al. [98]	Unsupervised	EM algorithm based model (Tcrowd) that handles caterogical and continuous attributes	Real-life dataset containing reviews of crowd	Yes

efficiently [81]. However it is interesting to note that, downweighting the unreliable annotators every time might not be a good strategy, rather the diversity can make the prediction better.

In a very recent study, a mixed-effects model based on HodgeRank estimation has been used for crowdsourced judgment analysis [82]. This work uses a simple iterative algorithm called Linearized Bregman Iterations (LBI) to generate paths of parsimonious models at different sparsity levels. However, a synchronized and parallel version of LBI is finally used for achieving high scalability. Though the aforementioned models demonstrate a better accuracy when compared to majority voting, but these models are specific to some kind of data with typical characteristics. Therefore, generalization of these methods may not always be flexible for generic datasets. A recent study based on autoencoders provides a generalized framework to aggregate the crowd opinions with different characteristics [83]. More specifically, this work is the pioneering study that employs autoencoders to learn hidden data patterns between the source label and inferred labels. In line of this, many other recent approaches leaverage deep learning to infer final judgment from multiple noisy opinions [84], [85], [86]. A disadvantage of the model in [83] is that it can employ the unlabeled data only for learning a classifier. However in [84], the estimation of latent features and data distribution have been taken care of in addition to learning the classifier.

5.8 Preferential and Absolute Judgment

Judgments are often categorized into absolute judgment and preferential judgment. Absolute judgment is the categorical label given for a particular question. Good, bad, poor, etc. may be the different absolute labels. Preferential judgment is like comparing between a pair of questions. In some studies

TABLE 2
Various Characteristics based on Coverage of the Models Considered in Different Major Algorithms

Algorithm	Quantifying questions	Quantifying annotators	Online availability	Data imbalance	Consideration of multidimensional annotator
Bhattacharyya et al. [27]	No	Yes	Yes	No	No
Liu et al. [28]	No	Yes (downweight)	Yes	Yes	No
Venanzi et al. [29]	No	Yes	–	No	No
Raykar et al. [54]	No	–	Yes	No	No
Whitehill et al. [58]	Yes	Yes (bias not considered)	–	No	No
Welinder et al. [59]	Yes	Yes	Yes	No	Yes
Dawid et al. [72]	No	Yes	Yes	No	No
Ye et al. [89]	No	No	Yes	No	No
Snow et al. [20]	No	Yes	–	No	No
Xu et al. [82]	No	Yes	Yes	Yes	Yes
Li et al. [91]	No	Yes	No	No	No
Zhou et al. [92]	No	Yes	Yes	Yes	Yes
Liu et al. [93]	No	Yes	Yes	Yes	Yes
Li et al. [94]	No	Yes	Yes	Yes	Yes
Zheng et al. [78]	No	Yes	No	Yes	No
Yin et al. [83]	Yes	Yes	No	No	No
Zhang et al. [95]	No	Yes	Yes	Yes	Yes
Zhong et al. [96]	No	Yes	Yes	Yes	No
Zhuo et al. [97]	No	Yes	No	No	No
Shan et al. [98]	Yes	Yes	No	No	No

each question is associated with some score instead of assigning labels. The main goal of such works is to find a unified probabilistic model that will combine the two types of judgments. Earlier in the Quality of Experience (QoE) community, most of the works treated the absolute judgment and preferential judgment independently [87]. In this model, due to the variety of tasks, the implementation of global quality measure is not possible. Rather, it is the responsibility of the employer to incorporate the appropriate task design with an aim to produce quality results. Here, the authors utilized the wisdom of crowd to experiment Mean opinion Score (MOS) test for QoE assessment. Chen et al. has recently proposed an approach that integrates different judgments from crowd through pairwise comparison [88]. On the other side, the model in [89] assumes that variance of noise for different annotators for different questions is same. But in practical scenario this appears to be impossible.

6 A Comparative Overview of Various Methods

In the introduction part we have elaborately discussed how the judgment analysis process can proceed phase by phase. In this section, different judgment analysis algorithms are discussed based on various performance measures. We have earlier discussed, as there are a number of issues to estimate the gold judgments, various approaches try to address these issues from different perspectives. The algorithms are mainly categorized into two broad classes - supervised and unsupervised. Though, most of the recent approaches focus on a unified probabilistic model (Bayesian joint probabilistic model) and estimate maximum likelihood of some parameters.

Quantifying annotators and questions, deciding the working principle (i.e., whether they are supervised or unsupervised), model type, etc. are some of the measures based on which the algorithms can be analyzed. The comparative analysis of a majority of these methods are reported in Tables 1 and 2. It can be seen from Table 2 that quantification of questions is considered only in [58], [59] and the application domain of these methods is related to the computer vision tasks. So quantifying questions should be different for other types of datasets instead of images. It is also observed that multidimensionality of an annotator is considered in [59]. All of these methods are basically developed with an aim to improve the reliable prediction. A major limitation is that most of these methods consider a general framework of collecting opinions from crowd workers. But different real-life tasks demand to consider different opinion collection models that can solve numerous complex real-life problems efficiently. Due to the discrepancies of human annotations, the goal of this model [59] is to discover the single ground truth for each task. This system predicts on the basis of single ground truth and collects sufficient annotations to remove the differences of crowd workers. However, there are applications (and possibilities) to have multiple ground truth for each image thereby requiring more advanced methods to be developed [90].

7 Conclusion

We have provided a comprehensive overview of aggregating crowdsourced opinions in this manuscript. This study can provide a helpful and in-depth insights over recent cutting-edge research in judgment analysis in crowdsourcing. However, there remain a number of areas open with the scope for future study. We see that clustering approach is effective to find similar type of annotators but similarity can be extended for temporal annotation where the annotators' response over some questions may change over time. Thus triclustering can be used in that scenario to find out similar annotators for a specific set of questions in a particular instant of time.

The graphical model discussed earlier is effective for aggregating large-scale crowd judgment analysis. Previous works judiciously incorporate the crucial factors like annotator accuracy, biasness and question difficulty to predict the final judgment. Now the expression of accuracy, biasness and question difficulty can be quantified in different ways. As a future scope, the proposed approaches can be extended by developing suitable expression for bias and case difficulty to apply on the large-scale unbalanced datasets where the opinions are more ambiguous. Again, in some of the cases the option set can contain continuous values. Therefore, it is needed to define the option set by discretizing it and Bayesian binning approach can also be used to find the optimal number of options.

In most of the crowdsourcing models, the requester posts a job in an online platform to seek the crowd opinions. After the completion of annotation, the opinions of all the annotators are considered to finally derive the aggregated judgment. Now, it is a challenging task for a requester to choose the termination criteria of annotation process. It is an important issue to consider how many responses from the annotators should be considered as enough depending on the difficulty level of questions to derive the final aggregated judgment from multiple opinions [75].

We can see there are different applications where a crowd worker observes just a few recent crowd opinions received while providing his/her own opinion. Rather only a partially observed opinions are available in some cases. So it is interesting to model how human psychology behaves when the partial information is available to them. Again, there should be some logical strategies for designing the online platform to study how many opinions should be revealed to the crowd worker to extract the best possible answers from them in dependent judgment analysis. On the other hand, an important question is that which recent opinions (instead of all the opinions) should be made open while collecting opinions from a crowd worker. One possible way might be revealing the opinions taken from the neighbouring crowd workers based on similarity of answers. Finally, the most interesting challenge is how to make aggregated judgment from these partially observed opinions. There is another scope of devising a more robust model for dependent judgment analysis for extreme-scale datasets.

Utility of considering multiple features of annotators instead of a single feature for accurate judgment is already known [99]. It can also be investigated how weighted rank aggregation method [100], [101] can be employed for this purpose. Due to the uncertainty property, it is not possible to obtain ranking for each of the annotators. Thus there are many situations in which only partial rankings can be obtained from the annotators. In fact, there are some metrics based on which the goodness of few annotators cannot be observed. Therefore, partial rank aggregation can be very useful in this case to produce better aggregated ranking. On the other hand, missing value prediction can have a major role in yielding a better judgment. A few models have been developed employing probabilistic matrix factorization [60], [102] to predict missing values and thereafter performing judgment analysis. But there is a future scope of applying this method on the response matrix when time dimension is also considered in order to produce better judgment. There remains a further scope for improvement as we can produce better missing value estimation if multiobjective optimization of various objective functions are considered simultaneously.

It is also reviewed that various applications of incorporating crowd knowledge can be efficiently used in solving different real-life problems through judgment analysis. In order to establish this, it has been shown that crowdsourcing can solve various types of image clustering tasks having no ground truth regarding the number of clusters [9], [58], [103], [104]. Furthermore, it is shown that ensemble of different clustering solutions is a major challenge. For this purpose, adjusted Rand index is used in general as a suitable cluster validity index to find out the accuracy of annotators. Again further study can be made if the solutions from the crowd workers can be obtained in two phases and confidence score is measured. In a crowd-powerd clustering problem, it is assumed that all the crowd workers annotate all the objects. But the problem arises when partial objects are visible to the crowd workers or all the crowd workers are not interested to cluster all the objects. Therefore, suitable method for label correspondence between any two clustering solutions and proper ensemble of clustering solutions should be designed in order to aggregate diverse multiple crowd solutions. As a whole, several novel and interesting research domains remain open for the researchers interested in judgment analysis of crowdsourced opinions.

ACKNOWLEDGMENTS

The authors would like to thank the anonymous reviewers for their valuable comments that greatly helped to improve the quality of the paper. The work of Malay Bhattacharyya is supported by the Visvesvaraya Young Faculty Research Fellowship 2015-2016 of the Ministry of Electronics & Information Technology, Government of India, being implemented by Digital India Corporation (formerly Media Lab Asia).

REFERENCES

[1] E. Brunswik, *The Conceptual Framework of Psychology. (International Encyclopedia of Unified Science)*, vol. 1, Chicago, IL, USA: The Univ. Chicago Press, 1952.

[2] N. Karelaia and R. M. Hogarth, "Determinants of linear judgment: A meta-analysis of lens model studies," *Psychol. Bulletin*, vol. 134, no. 3, pp. 404–426, 2008.

[3] S. Chatterjee and M. Bhattacharyya, "A biclustering approach for crowd judgment analysis," in *Proc. 2nd ACM IKDD Conf. Data Sci.*, 2015, pp. 118–119.

[4] S. Chatterjee and M. Bhattacharyya, "Judgment analysis of crowdsourced opinions using biclustering," *Inf. Sci.*, vol. 375, pp. 138–154, 2017.

[5] S. Chatterjee and A. M. M. Bhattacharyya, "Judgment analysis based on crowdsourced opinions," in *Proc. 33rd IEEE Int. Conf. Data Eng.*, Apr., 2017, pp. 1439–1444.

[6] D. C. Brabham, "Crowdsourcing as a model for problem solving," *Int. J. Res. Into New Media Technol.*, vol. 14, no. 1, pp. 75–90, 2008.

[7] P. Smyth, U. Fayyad, M. Burl, P. Perona, and P. Baldi, "Inferring ground truth from subjective labelling of venus images," in *Proc. Neural Inf. Process. Syst.*, 1994, pp. 1085–1092.

[8] A. Sorokin and D. Forsyth, "Utility data annotation with amazon mechanical turk," in *Proc. Comput. Vis. Pattern Recognit. Workshops*, 2008, pp. 1–8.

[9] L. von. Ahn and L. Dabbish, "labeling images with a computer game," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2004, pp. 319–326.

[10] Y. Chen, L. Chen, and C. J. Zhang, "Crowdfusion: A crowdsourced approach on data fusion refinement," in *Proc. 33rd IEEE Int. Conf. Data Eng.*, 2017, pp. 127–130.

[11] K. Crammer, M. Kearns, and J. Wortman, "Learning from multiple sources," *J. Mach. Learn. Res.*, vol. 9, pp. 1757–1774, 2008.

[12] A. Slivkins and J. W. Vaughan, "Online decision making in crowdsourcing markets: Theoretical challenges (position paper)," *CoRR*, vol. abs/1308.1746, 2013.

[13] N. Dalvi, A. Dasgupta, R. Kumar, and V. Rastogi, "Aggregating crowdsourced binary ratings," in *Proc. 22nd Int. Conf. World Wide Web*, 2013, pp. 285–294.

[14] G. Demartini, D. E. Difallah, and C. Mauroax, "Zencrowd: Leveraging probabilistic reasoning and crowdsourcing techniques for large scale entity linking," in *Proc. 21st Int. Conf. World Wide Web*, 2012, pp. 469–478.

[15] A. Kittur, J. V. Nickerson, J. V. Bernstein, M. S. Gerber, E. M. Shaw, A. Zimmerman, M. Lease, and J. J. Horton, "The future of crowd work," in *Proc. Conf. Comput. Supported Cooperative Work*, 2013, pp. 1301–1318.

[16] Q. V. H. Nguyen, T. T. Nguyen, N. T. Lam, and K. Aberer, "Batc: A benchmark for aggregation techniques in crowdsourcing," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 1079–1080.

[17] G. Chen, S. Zhang, D. Lin, H. Huang, and P. A. Heng, "Learning to aggregate ordinal labels by maximizing separating width," in *Proc. 34th Int. Conf. Mach. Learn.*, 2017, pp. 787–796.

[18] J. Ross, L. Irani, M. S. Silberman, A. Zaldivar, and B. Tomilson, "Who are the crowdworkers? Shifting demographics in mechanical turk," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2010, pp. 2863–2872.

[19] A. Sheshadri and M. Lease, "SQUARE: A benchmark for research on computing crowd consensus," in *Proc. AAAI Conf. Hum. Comput.*, 2013, pp. 2035–2043.

[20] R. Snow, B. O'Connor, B. Jurafsky, and A. Ng, "Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks," in *Proc. Empherical Methods Natural Language Process.*, 2008, pp. 254–263.

[21] T. Tian and J. Zhu, "Max-margin majority voting for learning from crowds," in *Proc. 28th Int. Conf. Neural Inf. Process. Syst.*, 2015, pp. 1621–1629.

[22] Y. Zheng, G. Li, Y. Li, C. Shan, and R. Cheng, "Truth inference in crowdsourcing: Is the problem solved?" *Proc. VLDB Endowment*, vol. 10, no. 5, pp. 541–552, 2017.

[23] Y. Amsterdamer, Y. Grossman, T. Milo, and P. Senellart, "Crowd mining," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2013, pp. 241–252.

[24] J. Deng, O. Russakovsky, J. Krause, M. S. Bernstein, A. Berg, and L. Fei-Fei, "Scalable multi-label annotation," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2014, pp. 3099–3102.

[25] L. von Ahn, R. Liu, and M. Blum, "Peekaboom: A game for locating objects in images," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2006, pp. 55–64.

[26] L. V. Ahn and L. Dabbish, "Labeling images with a computer game," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2004, pp. 319–326.

[27] M. Bhattacharyya, "Opinion ensembling: Learning from dependent judgements of the crowd," in *Proc. Crowdscale Shared Task Challenge*, 2013, p. 1.

[28] Q. Liu, J. Peng, and A. Ihler, "Report of crowdscale shared task challenge 2013, " in *Proc. Crowdscale Shared Task Challenge*, 2013, p. 2.

[29] M. Venanzi, J. Guiver, G. Kazai, and P. Kohli, "Bayesian combination of crowd-based tweet sentiment analysis judgments," in *Proc. Crowdscale Shared Task Challenge*, 2013, p. 3.

[30] Y. Bachrach, T. Minka, J. Guiver, and T. Graepel, "How to grade a test without knowing the answers: A bayesian graphical model for adaptive crowdsourcing and aptitude testing," in *Proc. 29th Int. Conf. Mach. Learn.*, 2012, pp. 819–826.

[31] C. Chai, J. Fan, G. Li, J. Wang, and Y. Zheng, "Crowd-powered data mining," *CoRR*, vol. abs/1806.04968, 2018.

[32] T. Chklovski and Y. Gil, "Towards managing knowledge collection from volunteer contributors," in *Proc. AAAI Spring Symp. Knowl. Collection Volunteer Contrib.*, 2005, pp. 21–27.

[33] Q. V. H. Nguyen, H. V. Huynh, T. T. Nguyen, W. Matthias, Y. Hongzhi, and Z. Xiaofang, "Computing crowd consensus with partial agreement," *IEEE Trans. Knowl. Data Eng.*, vol. 30, no. 1, pp. 1–14, Jan. 2017.

[34] P. G. Ipeirotis, F. Provost, and J. Wang, "Quality management on amazon mechanical turk," in *Proc. ACM SIGKDD Workshop Hum. Comput.*, 2010, pp. 64–67.

[35] F. Rodrigues, F. Pereira, and B. Ribeiro, "Learning from multiple annotators: Distinguishing good from random labelers," *Pattern Recognit. Lett.*, vol. 34, no. 12, pp. 1428–1436, Sep. 2013.

[36] P. Singh, "The public acquisition of commonsense knowledge," in *Proc. AAAI Spring Symp. Acquiring Linguistic Knowl. In. Access*, 2002, pp. 47–52.

[37] D. G. Stork, "Character and document research in the open mind initiative," In *Proc. Fifth Int. Conf. Document Anal. and Recognit.*, pp. 1–12, 1999.

[38] H. Hu, Y. Zheng, Z. Bao, G. Li, J. Feng, and R. Cheng, "Crowdsourced POI labelling: Location-aware result inference and task assignment," in *Proc. 32nd IEEE Int. Conf. Data Eng.*, 2016, pp. 61–72.

[39] L. von Ahn, M. Kedia, and M. Blum, "Verbosity: A game for collecting common-sense knowledge," in *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, 2006, pp. 75–78.

[40] G. Kazai, J. Kamps, and N. Milic-Frayling, "Worker types and personality traits in crowdsourcing relevance labels," in *Proc. 20th ACM Int. Conf. Inf. Knowl. Manage.*, 2011, pp. 1941–1944.

[41] K. Lee, J. Caverlee, and S. Webb, "The social honeypot project: Protecting online communities from spammers," in *Proc. 19th Int. Conf. World Wide Web*, 2010, pp. 1139–1140.

[42] M. Marge, S. Banerjee, and A. I. Rudnicky, "Using the amazon mechanical turk to transcribe and annotate meeting speech for extractive summarization," in *Proc. NAACL HLT Workshop Creating Speech Language Data Amazon's Mech. Turk*, 2010, pp. 99–107.

[43] R. Meng, L. Chen, Y. Tong, and C. J. Zhang, "Knowledge base semantic integration using crowdsourcing," *IEEE Trans. Knowl. Data Eng.*, vol. 29, no. 5, pp. 1087–1100, May 2017.

[44] G. Li, Y. Zheng, J. Fan, J. Wang, and R. Cheng, "Crowdsourced data management: Overview and challenges," in *Proc. ACM Int. Conf. Manage. Data*, 2017, pp. 1711–1716.

[45] G. Li, J. Wang, Y. Zheng, and M. J. Franklin, "Crowdsourced data management: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 9, pp. 2296–2319, Sep. 2016.

[46] A. D. Sarma, A. G. Parameswaran, and J. Widom, "Towards globally optimal crowdsourcing quality management: The uniform worker setting," in *Proc. Int. Conf. Manag. Data*, 2016, pp. 47–62.

[47] H. Garcia-Molina, M. Joglekar, A. Marcus, A. Parameswaran, and V. Verroios, "Challenges in data crowdsourcing," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 901–911, Apr. 2016.

[48] J. Fan, G. Li, B. C. Ooi, K. Tan, and J. Feng, "iCrowd: An adaptive crowdsourcing framework," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2015, pp. 1015–1030.

[49] F. Ma, Y. Li, Q. Li, M. Qiu, J. Gao, S. Zhi, L. Su, B. Zhao, H. Ji, and J. Han, "Faitcrowd: Fine grained truth discovery for crowdsourced data aggregation," in *Proc. 21th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2015, pp. 745–754.

[50] A. R. Khan and H. Garcia-Molina, "Crowddqs: Dynamic question selection in crowdsourcing systems," in *Proc. ACM Int. Conf. Manag. Data*, 2017, pp. 1447–1462.

[51] D. Gao, Y. Tong, J. She, T. Song, L. Chen, and K. Xu, "Top-K team recommendation and its variants in spatial crowdsourcing," *Data Sci. Eng.*, vol. 2, no. 2, pp. 136–150, 2017.

[52] Y. Zheng, J. Wang, G. Li, R. Cheng, and J. Feng, "QASCA: A quality-aware task assignment system for crowdsourcing applications," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2015, pp. 1031–1046.

[53] Y. Zheng, G. Li, and R. Cheng, "DOCS: A domain-aware crowdsourcing system using knowledge bases," *Proc. VLDB Endow.*, vol. 10, no. 4, pp. 361–372, Nov. 2016.

[54] V. C. Raykar and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," *J. Mach. Learn. Res.*, vol. 13, pp. 491–518, 2011.

[55] D. Hovy, T. B. Kirkpatrick, A. Vaswani, and E. Hovy, "Learning whom to trust with MACE," in *Proc. NAACL-HLT*, 2013, pp. 1120–1130.

[56] C. Liu and Y. Wang, "True label + confusions: A spectrum of probabilistic models in analyzing multiple ratings," in *Proc. Int. Conf. Mach. Learn.*, 2012, pp. 225–232.

[57] R. J. D. Ayala, "The theory and practice of item response theory," *Psychometrika*, vol. 75, no. 4, pp. 778–779, 2010.

[58] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan, "Whose vote should count more: Optimal integration of labels from labelers of unknown expertise," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 2035–2043.

[59] P. Welinder, S. Branson, S. Belongie, and P. Perona, "The multidimensional wisdom of crowds," in *Proc. Neural Inf. Process. Syst.*, 2010, pp. 2424–2432.

[60] J. H. Jung and M. Lease, "Improving quality of crowdsourced labels via probabilistic matrix factorization," in *Proc. 4th Hum. Comput. Workshop AAAI*, 2012, pp. 101–106.

[61] J. Zhang, V. S. Sheng, J. Wu, and X. Wu, "Multi-class ground truth inference in crowdsourcing with clustering," *IEEE Trans. Knowl. Data Eng.*, vol. 28, no. 4, pp. 1080–1085, Apr. 2016.

[62] S. Chatterjee and M. Bhattacharyya, "Group decision making with probabilistic graphical model," in *Proc. CHI Extended Abstracts*, pp. 2445–2451.

[63] E. Kamar, A. Kapoor, and E. Horvitz, "Identifying and accounting for task-dependent bias in crowdsourcing," in *Proc. 3rd AAAI Conf. Hum. Comput. Crowdsourcing*, 2015, pp. 92–101.

[64] R. Drapeau, L. B. Chilton, J. Bragg, and D. S. Wel, "MicroTalk: Using argumentation to improve crowdsourcing accuracy," in *Proc. 4th AAAI Conf. Hum. Comput. Crowdsourcing*, pp. 32–41, 2016.

[65] S. Chatterjee, A. Mukhopadhyay, and M. Bhattacharyya, "Consensus of dependent opinions," in *Proc. WiP Track 4th AAAI Conf. Hum. Comput. Crowdsourcing, arXiv preprint arXiv:1609.01408*, 2016.

[66] S. Chatterjee, A. Mukhopadhyay, and M. Bhattacharyya, "Dependent judgment analysis: A markov chain based approach for aggregating crowdsourced opinions," *Inf. Sci.*, vol. 386, pp. 83–96, 2017.

[67] S. Chatterjee, A. Mukhopadhyay, and M. Bhattacharyya, "Smart city planning with constrained crowd judgment analysis," in *Proc. AAAI Spring Symp. AI Soc. Good*, 2017, pp. 16–22.

[68] S. Chatterjee, A. Mukhopadhyay, and M. Bhattacharyya, "Constrained crowd judgment analysis," *ACM SIGWEB Newslett., Autumn*, 2017, Art. no. 4, https://dl.acm.org/citation.cfm?id=3146484

[69] R. W. Ouyang, L. M. Kaplan, A. Toniolo, M. Srivastava, and T. J. Norman, "Parallel and streaming truth discovery in large-scale quantitative crowdsourcing," *IEEE Trans. Parallel Distrib. Syst.*, vol. 27, no. 10, pp. 2984 –2997, Oct. 2016.

[70] J. Dauwels, L. Garg, A. Earnest, and L. K. Pang, "Tensor factorization for missing data imputation in medical questionnaires," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2012, pp. 2109–2112.

[71] S. Chatterjee, A. Mukhopadhyay, and M. Bhattacharyya, "Quality enhancement by weighted rank aggregation of crowd opinion," in *Proc. WiP Track 5th AAAI Conf. Hum. Comput. Crowdsourcing*, arXiv preprint arXiv:1708.09662, 2017.

[72] A. P. Dawid and A. M. Skene, "Maximum likelihood estimation of observer error-rates using the EM algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 20–28, 1979.

[73] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques - Adaptive Computation and Machine Learning*. Cambridge, MA, USA: MIT Press, 2009.

[74] A. T. Nguyen, B. Wallace, J. J. Li, A. Nenkova, and M. Lease, "Aggregating and predicting sequence labels from crowd annotations," in *Proc. 55th Annu. Meet. Assoc. Comput. Linguistics*, 2017, pp. 299–309.

[75] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, "Spectral methods meet em: A provably optimal algorithm for crowdsourcing," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 3537–3580, Jan. 2016.

[76] Q. Liu, A. T. Ihler, and M. Steyvers, "Scoring workers in crowdsourcing: How many control questions are enough?" in *Proc. 27th Annu. Conf. Neural Inf. Process. Syst.*, 2013, pp. 1–9.

[77] E. Simpson, S. Reece, and S. J. Roberts, "Bayesian heatmaps: Probabilistic classification with multiple unreliable information sources," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*, 2017, pp. 109–125.

[78] Y. Zheng, R. Cheng, S. Maniu, and L. Mo, "On optimality of jury selection in crowdsourcing," in *Proc. 18th Int. Conf. Extending Database Technol.*, 2015, pp. 193–204.

[79] F. K. Khattak and A. Salleb-aouissi, "Quality control of crowd labeling through expert evaluation," in *Proc. 2nd Workshop Comput. Social Sci. Wisdom Crowds*, 2011, pp. 1–5.

[80] D. R. Karger, S. Oh, and D. Shah, "Iterative learning for reliable crowdsourcing systems," in *Proc. Adv. Neural Inf. Process. Syst.*, 2011, pp. 1953–1961.

[81] S. Jagabathula, L. Subramanian, and A. Venkataraman, "Identifying unreliable and adversarial workers in crowdsourced labeling tasks," *J. Mach. Learn. Res.*, vol. 18, no. 1, pp. 3233–3299, Jan. 2017.

[82] Q. Xu, J. Xiong, X. Cao, Q. Huang, and Y. Yao, "From social to individuals: A parsimonious path of multi-level models for crowdsourced preference aggregation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 41, no. 4, pp. 844–856, Apr. 2018.

[83] L. Yin, J. Han, W. Zhang, and Y. Yu, "Aggregating crowd wisdoms with label-aware autoencoders," in *Proc. 26th Int. Joint Conf. Artif. Intell.*, 2017, pp. 1325–1331.

[84] K. Atarashi, S. Oyama, and M. Kurihara, "Semi-supervised learning from crowds using deep generative models," in *Proc. 32nd AAAI Conf. Artif. Intell. 30th Innovative Appl. Artif. Intell., 8th AAAI Symp. Educ. Adv. Artif. Intell.*, 2018, pp. 1555–1562.

[85] D. P. Kingma, D. J. Rezende, S. Mohamed, and M. Welling, "Semi-supervised learning with deep generative models," in *Proc. 27th Int. Conf. Neural Inf. Process. Syst.*, 2014, pp. 3581–3589.

[86] F. Rodrigues and F. C. Pereira, "Deep learning from crowds," in *Proc. 32nd AAAI Conf. Artif. Intell., 30th Innovative Appl. Artif. Intell., 8th AAAI Symp. Edu. Adv. Artif. Intell.*, 2018, pp. 1611–1618.

[87] F. Ribeiro, D. Florencio, Z. Cha, and M. Seltzer, "An approach for crowdsourcing mean opinion score studies," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2011, pp. 2416–2419.

[88] X. Chen, P. N. Bennet, K. Collins-Thomson, and E. Horvitz, "Pairwise ranking aggregation in a crowdsourced setting," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 193–202.

[89] P. Ye, B. Yu, and D. Doermann, "Combining preference and absolute judgments in a crowd-sourced setting," in *Proc. ICML Workshop*, 2013.

[90] D. Gurari, K. He, B. Xiong, J. Zhang, M. Sameki, S. D. Jain, S. Sclaroff, M. Betke, and K. Grauman, "Predicting foreground object ambiguity and efficiently crowdsourcing the segmentation (s)," *Int. J. Comput. Vis.*, vol. 126, no. 7, pp. 714–730, Jul. 2018.

[91] Q. Li, Y. Li, J. Gao, L. Su, B. Zhao, M. Demirbas, W. Fan, and J. Han, "A confidence-aware approach for truth discovery on long-tail data," *Proc. VLDB Endow.*, vol. 8, no. 4, pp. 425–436, Dec. 2014.

[92] D. Zhou, J. C. Platt, S. Basu, and Y. Mao, "Learning from the wisdom of crowds by minimax entropy," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 2195–2203.

[93] Q. Liu, J. Peng, and A. Ihler, "Variational inference for crowdsourcing," in *Proc. 25th Int. Conf. Neural Inf. Process. Syst.*, 2012, pp. 692–700.

[94] Q. Li, Y. Li, J. Gao, B. Zhao, W. Fan, and J. Han, "Resolving conflicts in heterogeneous data by truth discovery and source reliability estimation," in *Proc. ACM SIGMOD Int. Conf. Manag. Data*, 2014, pp. 1187–1198.

[95] C. J. Zhang, Y. Tong, and L. Chen, "Where to: Crowd-aided path selection," *Proc. VLDB Endow.*, vol. 7, no. 14, pp. 2005–2016, Oct. 2014.

[96] J. Zhong, K. Tang, and Z. H. Zhou, "Active learning from crowds with unsure option," in *Proc. 24th Int. Conf. Artif. Intell.*, 2015, pp. 1061–1067.

[97] H. H. Zhuo, "Crowdsourced action-model acquisition for planning," in *Proc. 29th AAAI Conf. Artif. Intell.*, 2015, pp. 3439–3445.

[98] C. Shan, N. Mamoulis, G. Li, R. Cheng, Z. Huang, and Y. Zheng, "T-crowd: Effective crowdsourcing for tabular data," *CoRR*, vol. abs/1708.02125, 2017.

[99] J. Zhang, V. S. Sheng, and T. Li, "Label aggregation for crowdsourcing with bi-layer clustering," in *Proc. 40th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2017, pp. 921–924.

[100] V. Pihur, S. Datta, and S. Datta, "Finding common genes in multiple cancer types through meta-analysis of microarray experiments: a rank aggregation approach," *Genomics*, vol. 92, no. 6, pp. 400–403, 2008.

[101] V. Pihur, S. Datta, and S. Datta, "Weighted rank aggregation of cluster validation measures: A monte carlo cross-entropy approach," *Bioinf.*, vol. 23, no. 13, pp. 1607–1615, 2007.

[102] M. Berry, M. Browne, A. Langville, V. Pauca, and R. Plemmons, "Algorithms and applications for approximate nonnegative matrix factorization," *Comput. Statist. Data Anal.*, vol. 52, no. 2, pp. 155–173, 2007.

[103] A. Dash, S. Chatterjee, T. Prasad, and M. Bhattacharyya, "Image clustering without ground truth," in *Proc. GroupSight Workshop Collocated 4th AAAI Conf. Human Comput. Crowdsourcing, arXiv preprint arXiv:1610.07758*, 2016.

[104] S. Chatterjee, E. Kundu, and A. Mukhopadhyay, "A markov chain based ensemble method for crowdsourced clustering," in *Proc. WiP Track 4th AAAI Conf. Human Comput. Crowdsourcing, arXiv preprint arXiv:1609.01484*, 2016.

Sujoy Chatterjee received the doctoral degree from the Department of Computer Science and Engineering, University of Kalyani, India, in 2018. He is currently a postdoctoral research fellow with the Université Côte d'Azur, CNRS, I3S, France. His research interests include judgment analysis, crowdsourcing, data mining, and machine learning. He has published several research papers in top-tier journals like *Information Sciences* and conferences like ACM SIGCHI, AAAI HCOMP, AAAI Spring Symposium, ICDE, ACM IKDD CoDs, etc. He has received several national and international travel grants and best paper award during his research career.

Anirban Mukhopadhyay received the BE degree in computer science and engineering from the National Institute of Technology, Durgapur, India, in 2002, the ME degree in computer science and engineering from Jadavpur University, Kolkata, India, in 2004, and the PhD degree in computer science from Jadavpur University, in 2009. He is currently a professor of the Department of Computer Science and Engineering, University of Kalyani, Kalyani, West Bengal. He is the recipient of the University Gold Medal and Amitava Dey Memorial Gold Medal from Jadavpur University in 2004 for ranking first class first in ME. He also received the Erasmus Mundus fellowship in 2009 to carry out post-doctoral research at the University of Heidelberg and the German Cancer Research Center (DKFZ), Heidelberg, Germany during 2009-2010. He also visited the I3S laboratory, University of Nice Sophia-Antipolis, Nice, France, in 2011 as a visiting professor, the University of Goettingen, Germany, as a visiting scientist with the DAAD scholarship in 2013, and Colorado State University, Fort Collins, USA, as a visiting researcher with a Fulbright-Nehru fellowship during 2017-18. He has received the Institution of Engineers, India (IEI) Young Engineers Award 2013-14 in Computer Engineering Discipline, and the Indian National Academy of Engineering (INAE) Young Engineer Award 2014. He has coauthored one book and more than 150 research papers in various International Journals and Conferences. He is a senior member of the Institute of Electrical and Electronics Engineers (IEEE), USA, and member of the Association for Computing Machinery (ACM), USA. He is also a member of the IEEE Computational Intelligence Society (CIS) Kolkata Chapter and served in its executive body. He has co-edited special issues in reputed journals and co-organized special sessions in different conferences including IEEE WCCI 2016 and IEEE SSCI 2018. His research interests include soft and evolutionary computing, data mining, multiobjective optimization, pattern recognition, bioinformatics, and crowdsourcing.

Malay Bhattacharyya received the PhD degree in computer science from Indian Statistical Institute, Kolkata, in 2014. He is currently an assistant professor with the Machine Intelligence Unit of Indian Statistical Institute, Kolkata. He is a certified Design Thinker from the MIT Sloan School of Management. He has about six years of teaching and 10 years of research experience. He worked at the Indian Institute of Engineering Science and Technology, Shibpur (2014-2018), University of Kalyani (2012-2014), and the Indian Institute of Science, Bangalore (2014) earlier. He has published more than 70 research papers in various peer-reviewed journals, book chapters, and proceedings of international conferences. His current research interests include crowdsourcing, big data analysis, and computational biology. He received the Young Scientist Award from ISCA (2013-2014), became a Sir Visvesvaraya Young Faculty Research Fellow (2015-2016), received the Young Engineers Award from IEI (2016-17), received the Young Engineer Award from INAE (2018), and won several best paper and best reviewer awards. He visited Stanford University, USA, and the University of Ljubljana, Slovenia, for academic purposes. He is a member of the IEEE, ACM, and AAAI.