Information Sciences 375 (2017) 138–154

Contents lists available at ScienceDirect

Information Sciences

journal homepage: www.elsevier.com/locate/ins

Judgment analysis of crowdsourced opinions using biclustering

Sujoy Chatterjee[a], Malay Bhattacharyya[b,*]

[a] Department of Computer Science and Engineering, University of Kalyani, Nadia – 741235, India
[b] Department of Information Technology, Indian Institute of Engineering Science and Technology, Shibpur, Howrah – 711103, India

ARTICLE INFO

Article history:
Received 13 October 2015
Revised 1 July 2016
Accepted 14 September 2016
Available online 28 September 2016

Keywords:
Judgment analysis
Opinion ensemble
Majority voting
Biclustering

ABSTRACT

Annotation by the crowd workers serving online is gaining focus in recent years in diverse fields due to its distributed power of problem solving. Distributing the labeling task among a large set of workers (may be experts or non-experts) and obtaining the final consensus is a popular way of performing large-scale annotation in a limited time. Collection of multiple annotations can be effective for annotation of large-scale datasets for applications like natural language processing, image processing, etc. However, as the crowd workers are not necessarily experts, their opinions might not be accurate enough. This causes problem in deriving the final aggregated judgment. Again, majority voting (MV) is not suitable for such problems because the number of annotators is limited and they have multiple options to choose. This might cause too much conflicts among the opinions provided. Additionally, there might exist annotators who randomly try to annotate (provide spam opinions for) too many questions to maximize their payment. This can incorporate noise while deriving the final judgment. In this paper, we address the problem of crowd judgment analysis in an unsupervised way and a biclustering-based approach is proposed to obtain the judgments appropriately. The effectiveness of this approach is demonstrated on four publicly available small-scale Amazon Mechanical Turk datasets, along with a large-scale CrowdFlower dataset. We also compare the algorithm with MV and some other existing algorithms. In most of the cases the proposed approach is competitively better than others. But most importantly, it does not use the entire dataset for deriving the judgment.

1. Introduction

Crowdsourcing is one of the emerging fields that has been shown to have wide applications in the areas of data mining, machine learning, bioinformatics, etc. [17,18,25]. It provides the new opportunities to tackle diverse real-life problems using the united power of independent crowd workers. The recent popularity of online crowdsourcing services has also become useful for managing large-scale labeling tasks. The crowdsourcing model was formally introduced by Howe in 2006 [4,9]. However, Amazon Mechanical Turk (AMT) was the first crowd-powered system that appeared in 2005 to successfully solve diverse problems using the online crowd workers. Other than AMT, various crowdsourcing platforms like WikiProjects, Kickstart, Crowdfyerged, etc. are the most popular tools. These crowdsourcing platforms can be categorized into two main types

* Corresponding author.
E-mail addresses: sujoy@klyuniv.ac.in (S. Chatterjee), malaybhattacharyya@it.iiests.ac.in, malaybhattacharyya@gmail.com (M. Bhattacharyya).

http://dx.doi.org/10.1016/j.ins.2016.09.036
0020-0255/

– collaborative and competitive [33], based on their working behavior. When a large number of people solve a given problem independently with competing interests, it is known as competitive crowdsourcing. If the same problem is solved by a set of people together, then it is called collaborative crowdsourcing. The behavior of collaborative crowdsourcing was first studied by Ipeirotis in 2010 [10], whereas competitive crowdsourcing was first studied by Boudreau et al. in 2011 [3].

With the advancement of crowdsourcing research, it has become easier to label large amount of data in limited cost by a large pool of collaborative crowd workers. The labeling task was usually given to the expert annotators in earlier times. But this might consume high cost in terms of time and money. The set of experts has recently been successfully replaced by a large set of online workers (where no guarantee is there about their expertise) to minimize the overall cost. In this paper, we discuss about an unsupervised model that uses this type of workflow. Though the traditional labeling expects that every annotator should be an expert having background knowledge, but interestingly there exist some annotators who give their opinions without knowing the possible options from the option set. The annotators who give their opinions without using their knowledge over the possible option set are called spammers. They just pretend to be experts to gain extra money. This makes noise in generating the aggregated label when the final consensus judgment is to be desired from the multiple opinions. Sorokin and Forsyth found that sloppy annotation causes some of the error [32].

Taking repeated labeling from the same worker is one of the most common ideas to get rid of spamming strategy and to estimate the original true label from the noisy opinions. Snow et al. [31] found that an annotation given by a small number of non-experts may perform like an expert annotator. Estimating the observer error rate as well as reducing it, can keep the information loss within a manageable limit. Dawid Skene in 1979, addressed a clinical judgment problem that solicited each patient clinical record repeatedly [5]. Here, the error rate has been estimated in supervised and unsupervised way by using the Expectation Maximization algorithm. Recent attempts have been made to derive the accurate judgment from the multiple individual opinions [2,12–14,20–23,29]. Finding the reliability of an annotator and estimating the gold label are the most important challenges. Hovy et al. addressed this problem to find the trustworthiness of an annotator [8].

In this paper, the opinion-based judgment analysis problem has been addressed using the biclustering approach. In crowd-powered labeling, it is not necessary that all the annotators would give their opinions for all the questions. Thus, it becomes different from the class of ensemble problems. Again, we have seen that there are some annotators (basically the spammers) who try to annotate a large number of questions by giving random opinion or they try to pretend that they are the expert annotators to gain extra money. The consistence performance over a large set of questions signifies that the annotator may not be a spammer. Hence, the weighted judgment of that annotator should be given more importance. This idea has been taken into account in our approach. It has been seen that there is a certain set of annotators who are attempting a particular set of questions. So, these annotators can be classified into certain groups based on their attempted questions. Based on this idea, annotator set and question set has been segregated. As a result, we are getting some set of annotators with respect to the some set of questions. Now, we have mixed the concept of majority voting (MV) to compute the accuracy of each annotator and based on the accuracy the annotators have been ranked. For the biclustering approach, we have primarily executed majority voting to measure the accuracy of each annotator. The predicted label that we obtained after majority voting algorithm has been incorporated into the proposed algorithm to find the accuracy of the each annotator. Ultimately, we have integrated the individual solutions to find the consensus labeling. We have executed our algorithm on a large-scale dataset namely CrowdFlower dataset, as well as we have also tested the performance of our algorithm on four publicly available small-scale AMT benchmark datasets. In most of the cases we are getting more than 90% accuracy for our algorithm and it outperforms the MV and some other state-of-the-art consensus algorithms.

2. Basic terminology

In this section, we introduce the basic terminology that will be used throughout the paper. We involve five basic terms, namely question, annotator, opinion, gold judgment and annotator accuracy for background description.

2.1. Question

A question is simply a labeling or annotation task. E.g., it can be the labeling of an image (whether it is scary or not), annotation of a text (where is the noun, verb, etc.), or commenting on something (what the tweet, online post, etc. talks about).

2.2. Annotator

An annotator is basically a crowd worker who gives the opinion (annotates a label) over the given question. Annotation can be thought of as a function that returns one of the available options according to some procedure. An annotator may be good in (labeling) one type of questions and can be bad for some other types of questions.

2.3. Opinion

An opinion is the annotation given by an annotator for a particular question. Generally, the opinion should be chosen from an opinion set.