Neurocomputing 160 (2015) 141–156

Contents lists available at ScienceDirect

Neurocomputing

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

Modeling annotator behaviors for crowd labeling

Yunus Emre Kara [a,*], Gaye Genc [a], Oya Aran [b], Lale Akarun [a]

[a] *Department of Computer Engineering, Bogazici University, TR-34342 Bebek, Istanbul, Turkey*
[b] *Idiap Research Institute, Martigny, Switzerland*

ARTICLE INFO

Article history:
Received 15 March 2014
Received in revised form
15 October 2014
Accepted 17 October 2014
Available online 11 February 2015

Keywords:
Label noise
Crowdsourcing
Crowd labeling
Annotator behavior
Annotator quality
Consensus

ABSTRACT

Machine learning applications can benefit greatly from vast amounts of data, provided that reliable labels are available. Mobilizing crowds to annotate the unlabeled data is a common solution. Although the labels provided by the crowd are subjective and noisy, the wisdom of crowds can be captured by a variety of techniques. Finding the mean or finding the median of a sample's annotations are widely used approaches for finding the consensus label of that sample. Improving consensus extraction from noisy labels is a very popular topic, the main focus being binary label data. In this paper, we focus on crowd consensus estimation of continuous labels, which is also adaptable to ordinal or binary labels. Our approach is designed to work on situations where there is no gold standard; it is only dependent on the annotations and not on the feature vectors of the instances, and does not require a training phase. For achieving a better consensus, we investigate different annotator behaviors and incorporate them into four novel Bayesian models. Moreover, we introduce a new metric to examine annotator quality, which can be used for finding good annotators to enhance consensus quality and reduce crowd labeling costs. The results show that the proposed models outperform the commonly used methods. With the use of our annotator scoring mechanism, we are able to sustain consensus quality with much fewer annotations.

1. Introduction

In 1906, statistician Francis Galton observed a contest held in a fair; on estimating the weight of a slaughtered and dressed ox. He calculated that the median guess of 787 people was 1207 pounds which is within 0.8% of the true weight of 1198 pounds [1]. This experiment broke new ground in cognitive science; establishing the notion that opinions of a crowd on a particular subject can be represented by a probability distribution. This is what we today call the wisdom of crowds. A crowd can be any group of people, such as the students of a school, or even the general public. In daily life, when we lack knowledge about a certain concept we inquire those around us to obtain a general idea. A similar approach can also be adapted to scientific research where it is not feasible or possible to observe the phenomenon directly.

Employing the power of a crowd for a task is called crowdsourcing. Many applications in crowdsourcing exist such as fundraising, asking for people to vote their appreciation of movies and books, or dividing up and parallelizing complex tasks to be completed. The microwork concept deals with breaking up very large problem that may or may not be solved by computers. Amazon Mechanical Turk [2] and Crowdflower [3] are examples of microwork platforms where people submit lots of small tasks to be completed by other people all around the world, for a fee.

Ground truth labeling is often considered to be a menial task and consumes the valuable time of researchers acquiring datasets. For labeling tasks that do not require expert opinion, many research centers and universities prefer paying a group of people from the general population for ground truth annotation.

Assume that we have N samples and R annotators where each annotator annotates a randomized subset of N samples and every sample is annotated by a group of annotators. This is a common case for crowdsourced annotation tasks. The aim of our work is to obtain consensus labels for each sample using these annotations.

In this paper, we focus on modeling annotator behavior and incorporating it in four new Bayesian models that we propose for the crowd labeling problem. The models we propose are designed particularly for continuous or ordinal scores, but could be applied to categorical scores as well. Our method is specifically designed for problems where there is no gold standard and we do not include a training step in our approach. We also provide a new annotator scoring mechanism, which may be used to weed out low quality annotators and reduce crowd labeling costs.

We start by addressing related work in the literature in Section 1.1 and emphasizing our contributions in Section 1.2. We investigate annotator behaviors by explaining various annotator types in Section 2. Then, we present the proposed Bayesian models in Section 3, which

[*] Corresponding author
E-mail addresses: yunus.kara@boun.edu.tr (Y.E. Kara), gaye.genc@boun.edu.tr (G. Genc), oaran@idiap.ch (O. Aran), akarun@boun.edu.tr (L. Akarun).

http://dx.doi.org/10.1016/j.neucom.2014.10.082
0925-2312/© 2015 Elsevier B.V. All rights reserved.

are used for simultaneously modeling the behaviors of annotators and finding consensus for each sample. Section 4 describes the measure we propose for scoring the competence of annotators. Since crowdsourced labeling is an expensive process, choosing good annotators is crucial for reducing the costs. That makes annotator competence scoring an important aspect of our work. In Section 5, we present the results of our experiments for evaluating our models. The experiments are performed on two crowdsourced datasets, with and without ground truth information. Finally, we conclude the work in Section 6, with possible future directions.

1.1. *Related work*

An annotation task completed by crowdsourcing contains vast information along with many interesting challenges. Annotators come from different backgrounds, their experiences vary, and they provide opinions over a large scale. An in-depth survey by Frenay et al. [4] focuses on defining label noise and its sources, and introduces a taxonomy on the types of label noise. Potential drawbacks and related solutions are discussed, including algorithms which are label noise-tolerant, label noise cleansing, and label noise-robust. Srivastava et al. investigate the problem of subjective video annotation and majority opinion is shown to be the most objective annotation for a video [5]. Carpenter [6] utilizes multilevel Bayesian approaches on binary data annotations, and introduces priors on sensitivity and specificity of annotators. Singular opinions of the annotators are unreliable, but the consensus of the crowd provides a strong insight. Finding a reasonable consensus among the annotators is very important, especially in cases where the ground truth (or gold standard) does not exist. Raykar et al. estimate the gold standard and measure the competence of the annotators iteratively in a probabilistic approach [7]. Their results are challenged by Rodrigues et al. in a supervised multiclass classification problem with a simpler probabilistic model [8]. Ground truth estimation is done by annotator modeling by using the annotators' self-reported confidences in [9]. Human personality trait evaluation is also a problem where no quantifiable ground truth exists. Trait annotations collected by crowdsourcing are used in [10] for personality trait classification.

The problem of annotator reliability is a very popular subject and tackled in [11] by using Gaussian mixture models. Liu et al. approach this problem by using belief propagation and mean field methods [12]. Statistical methods are used for estimating annotator reliability and behavior [13], as well as including annotator parameters such as bias, expertise, and competence [14]. Both approaches group annotator behaviors into different “schools of thought”. Deciding on annotator reliability is also accomplished by measuring annotator quality. Wu et al. propose a probabilistic model of active learning with multiple noisy oracles together with the oracles' labeling quality [15]. Dutta et al. also deal with annotator quality in a crowdsourcing case study where the multiple annotators provide high level categories for newspaper articles [16]. Donmez et al. introduce a new algorithm based on Interval Estimation for estimating the accuracy of multiple noisy annotators and select the best ones for active learning [17].

Annotators' varying expertise both among themselves and over different parts of the data are also factors affecting their reliability. Zhang et al. investigate annotator expertise with a combination of ML and MAP estimation [18]. An online learning algorithm weeding out unreliable annotators and asking for labels from reliable annotators for instances which have been poorly labeled has been introduced in [19]. Varying annotator expertise problems are also handled in [20,21] with ground truth estimation, using MAP estimation and EM approach. Whitehill et al. also study annotator expertise, taking noisy and adversarial annotators into account [22].

Detecting spammers/abusers, and biased annotators is also useful for eliminating and/or modifying specific annotations. Spectral decomposition techniques are used for moderating abusive content in [23]. Raykar et al. propose an empirical Bayesian algorithm for iteratively eliminating spammers and estimating consensus labels from good annotators [24]. Wauthier et al. present a new Bayesian model for reducing annotator bias to combine the data collection, data curation and active learning [25].

1.2. *Novelty and contributions*

A straightforward solution for the continuous annotation case might be taking the mean or median of annotations for each sample. For the binary case, majority voting is the first solution that comes to mind. However, a few problems arise with these approaches, such as:

- Annotator errors and outliers have a high impact on the consensus.
- Valuable information on annotator behavior and expertise is disregarded.

Investigating the behaviors of annotators and modeling their aspects would prove useful for utilizing valuable information.

The methods in the literature that we mentioned are mostly designed for binary labeled input [6–8,14,18,21,24,26]. However, in many annotation problems, researchers request continuous or ordinal annotations and map the annotations to binary labels. An example of this is the heart wall segment level ratings where trained cardiologists are asked to rate the samples in the interval 1–5, but the input annotations are binarized as normal (1) and abnormal (2–5) [21,26]. Unfortunately, this binarization process results in the loss of valuable information.

Another approach is to use ordinal annotations, as if they were categories, as input to the categorical models [27,28]. Although it is possible to employ these types of models for ordinal labels, the categorical approach falls short of preserving the ordinal and proportional relations. For continuous or ordinal annotations, it is better to employ models that make use of ordinal and proportional information.

Numerous methods also make use of features extracted from data [7,18,29]. In the case where feature extraction is not possible or feasible, methods such as ours can be used. Moreover, the success of data dependent methods relies heavily on the quality of extracted features. In addition, model performance across different types of problems requiring different types of features is unpredictable.

There are only a handful of works focused on ordinal or continuous annotations. Raykar et al. [7] combined sample classification with label consensus estimation. In addition, they also propose a simple data-independent model for continuous labels. Lakshminarayanan and Teh [30] focus on ordinal labels. They incorporate task difficulty to the discretization of continuous latent variables in their model. These works are pioneering elements in the continuous crowd labeling problems. However, to the best of our knowledge, our work is the first attempt to investigate the effect of diverse annotator behaviors on consensus estimation and annotator scoring mechanism for continuous crowd labeling problems.

The contributions of this study can be summarized as follows:

- We propose four new Bayesian models that model annotator behaviors for continuous or ordinal annotations to estimate the consensus scores. The proposed methods do not require any training step and are particularly designed for problems where there is no ground truth available. As a result, they are suitable to the problems where the ground truth is not available by construct, i.e. subjective annotations of human behavior. We believe that this is the first work that incorporates numerous annotator behaviors in consensus estimation for continuous crowd labeling problems.
- We show that the consensus scores estimated by the proposed models can be converted to categorical scores using simple