Expert Systems with Applications 41 (2014) 5723–5732

Contents lists available at ScienceDirect

Expert Systems with Applications

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

Separate or joint? Estimation of multiple labels from crowdsourced annotations

Lei Duan [*], Satoshi Oyama, Haruhiko Sato, Masahito Kurihara

Graduate School of Information Science and Technology, Hokkaido University, Kita 14, Nishi 9, Kita-ku, Sapporo, Hokkaido 060-0814, Japan

ARTICLE INFO

Keywords:
Multi-label estimation
Crowdsourced annotation
Label dependency
Quality control
Human computation

ABSTRACT

Artificial intelligence techniques aimed at more naturally simulating human comprehension fit the paradigm of multi-label classification. Generally, an enormous amount of high-quality multi-label data is needed to form a multi-label classifier. The creation of such datasets is usually expensive and time-consuming. A lower cost way to obtain multi-label datasets for use with such comprehension–simulation techniques is to use noisy crowdsourced annotations. We propose incorporating label dependency into the label-generation process to estimate the multiple true labels for each instance given crowdsourced multi-label annotations. Three statistical quality control models based on the work of Dawid and Skene are proposed. The label-dependent *DS* (*D-DS*) model simply incorporates dependency relationships among all labels. The label pairwise *DS* (*P-DS*) model groups labels into pairs to prevent interference from uncorrelated labels. The Bayesian network label-dependent *DS* (*ND-DS*) model compactly represents label dependency using conditional independence properties to overcome the data sparsity problem. Results of two experiments, "affect annotation for lines in story" and "intention annotation for tweets", show that (1) the *ND-DS* model most effectively handles the multi-label estimation problem with annotations provided by only about five workers per instance and that (2) the *P-DS* model is best if there are pairwise comparison relationships among the labels. To sum up, flexibly using label dependency to obtain multi-label datasets is a promising way to reduce the cost of data collection for future applications with minimal degradation in the quality of the results.

1. Introduction

Given the complexity of human thinking, several artificial intelligence systems aimed at simulating human comprehension, including affect prediction and intention inference, have one thing in common: They more naturally fit the paradigm of multi-label classification than that of single-label classification since one instance may evoke more than one "comprehension" at the same time. Generally, an enormous amount of multi-label data is needed to form a multi-label classifier. Moreover, the data quality directly affects the performance of machine learning techniques. Obtaining high-quality data from both experts and large crowds can be expensive and time-consuming. We investigated ways to obtain at low cost reliable multi-label datasets for use with aforementioned comprehension–simulation techniques.

On-line crowdsourcing services provide a means for outsourcing various kinds of tasks to a large group of people, and labeling is one of the main crowdsourcing tasks. Although multi-label data can be obtained from a crowdsourcing service at very low cost (time and expense), crowdsourcing workers are rarely trained and generally do not have the abilities needed to accurately perform the offered task. Moreover, some workers may simply submit random responses as a means to earn easy money. Therefore, ensuring the quality of the results submitted by workers is one of the biggest challenges in crowdsourcing.

A promising approach to improving the quality of crowdsourced annotations is to introduce redundancy, which involves asking several workers to work on each task, and then aggregating their results to obtain a more reliable result. The simplest aggregation strategy, *majority vote*, is valid only if the number of workers is large enough. It is based on the implicit assumption that all workers have the same probability of making an error. If the number of workers is less than a certain unknown number, the detrimental effect of the noisy responses is significant, and treating responses given by different workers equally produces poor quality results. However, collecting data from a large number of workers is almost impossible due to the high cost (time and expense). In view of this, several sophisticated statistical techniques (Dawid & Skene, 1979; Oyama, Baba, Sakurai, & Kashima, 2013; Welinder,

[*] Corresponding author.
E-mail address: duan@ec.hokudai.ac.jp (L. Duan).

http://dx.doi.org/10.1016/j.eswa.2014.03.048
0957-4174/© 2014 Elsevier Ltd. All rights reserved.

Branson, Belongie, & Perona, 2010; Whitehill, Wu, Bergsma, Movellan, & Ruvolo, 2009) have been proposed for obtaining reliable results from annotations provided by a handful of crowdsourcing workers. However, these techniques simply handle the problem of estimating a single true label for each single-labeled instance. Nowak and Rüger (2010) investigated the agreement between experts and crowdsourcing workers (non-experts) for multi-label image annotation. They found that the quality of crowdsourced annotations is similar to the annotation quality of experts. However, they did not determine how many crowdsourcing workers are needed to obtain comparable quality. To the best of our knowledge, the problem of multi-label estimation has not been effectively solved. Therefore, our aim here is to determine the best way to estimate multiple true labels for each instance from multi-label annotations provided by a handful of crowdsourcing workers. The aim is to reduce the cost of creating high-quality multi-label datasets for future applications with minimal degradation in the quality of the results.

Multi-label estimation from crowdsourced annotations can be seen as an unsupervised multi-label classification problem. Two widely used methods for multi-label classification are the binary relevance (*BR*) method and the label combination or label power-set (*LP*) method (Tsoumakas, Katakis, & Vlahavas, 2010). The *BR* method decomposes the multi-label estimation problem into several independent binary-label estimation problems, one for each label in the set of candidate labels. The final labels for each instance are determined by aggregating the predictions from all binary estimators. However, this method does not consider dependency among candidate labels. The *LP* method treats each unique subset of labels in the set of candidate labels as an atomic "label" and considers a new single-label estimation problem, i.e., estimating each member of the power-set of the candidate label set. Although the *LP* method takes label dependency into account, a large number of classes has to be dealt with when the number of candidate labels is large. Simply put, the *LP* method can easily suffer from the sparsity of high-dimensional annotations.

Aiming to address these limitations, we propose flexibly incorporating label dependency into the label-generation process. In particular, we propose three statistical quality control models based on the model of Dawid and Skene (1979) (*DS*), a well-known unsupervised single-label classification algorithm:

- *Label-dependent DS (D-DS) model*
  The *D-DS* model, which is an implementation of the *LP* method, simply takes the dependency relationships among all candidate labels into account.
- *Label pairwise DS (P-DS) model*
  The *P-DS* model groups candidate labels into pairs, and then separately estimates the states of the two labels within each pair, thereby preventing interference from uncorrelated labels.
- *Bayesian network label-dependent DS (ND-DS) model*
  The *ND-DS* model depicts the conditional independence properties of the joint distribution over candidate labels as a Bayesian network and approximates the underlying high-dimensional joint distribution by using the product of the conditional distributions of the candidate labels.

To evaluate the effectiveness of the proposed models for multi-label estimation, we conducted two experiments using Lancers crowdsourcing service.[1] In one, crowdsourcing workers were tasked with annotating the affects (emotions) of lines in a story, and in the other they were tasked with annotating the intentions of tweet posters. The results showed that, with multi-label annotations provided by a handful of crowdsourcing workers, in most cases, the *ND-DS* model handled the multi-label estimation problem more effectively than the other models. However, if there were pairwise comparison relationships among the candidate labels, the *P-DS* model was the most effective.

The remainder of this article is organized as follows. In Section 2, we review the original Dawid–Skene model, which is the basis of our study. Section 3 introduces two of the proposed multi-label estimation models: *D-DS* and *P-DS*. Section 4 describes the use of the expectation maximization (EM) algorithm to infer the results together with the parameters of the model. The drawback of the *D-DS* model is discussed and the *ND-DS* model is presented as the solution in Section 5. Section 6 describes the experimental design and presents the results obtained by applying the *majority vote* strategy, the original *DS* model, and the proposed models to actual crowdsourced annotations. Section 7 briefly introduces related work on quality control in crowdsourcing and provides some background material on the experiments conducted. Finally, Section 8 discusses the strengths of the proposed models, explains the research contributions in theory, discusses the implications of the research, points out the limitations of the proposed models, and suggests several future research directions.

2. Background: original Dawid–Skene (*DS*) model

Our work is based on the well-known Dawid–Skene model (Dawid & Skene, 1979), which is aimed at inferring the unknown health state of a patient given the assessments of several clinicians. Let I be the set of patients, J be the set of health state types, and K be the set of clinicians. That j is the true state of patient i is denoted as $t_i = j(i \in I, j \in J)$. The true state of patient i is estimated as

$$\arg\max_{j\in J} P\left(t_i = j \middle| \{n_{il}^k\}_{k\in K, l\in J}\right), \tag{1}$$

where $n_{il}^k \in \mathbb{N}(k \in K, i \in I, l \in J)$ denotes the number of times that clinician k declared patient i to be in state l.

In our research, instances and crowdsourcing workers are the counterparts of patients I and clinicians K. The state (*true* or *false*) of a particular label for an instance can be considered as the health state of a patient. On the basis of this, the *DS* model can be directly used to estimate the state of a particular label for each instance. Let $t_i = \jmath(i \in I, \jmath \in \{0,1\})$ denote whether a particular label is true $(\jmath = 1)$ or false $(\jmath = 0)$ for instance i, and let $n_{i\imath}^k \in \mathbb{N}(k \in K, i \in I, \imath \in \{0,1\})$ be the number of times that worker k annotated instance i with $(\imath = 1)$ or without $(\imath = 0)$ the label. Similar to formula (1), whether the label is true for instance i can be estimated using

$$\arg\max_{\jmath\in\{0,1\}} P\left(t_i = \jmath \middle| \{n_{i\imath}^k\}_{k\in K, \imath\in\{0,1\}}\right). \tag{2}$$

Simply put, the *DS* model is an implementation of the *BR* method.

3. Proposed models

As described in Section 2, the states of different labels for each instance must be estimated separately using different *DS* models. This is suitable for multi-label estimation only in the extreme case that labels are mutually independent. However, some labels may reveal clues about other labels. For instance, in the affect annotation experiment described in Section 6.1, *fondness*, *happiness*, and *relief* are frequently co-true, *fondness* and *anger* are rarely co-true, and *shame* or *anger* may be false when *relief* is true. To take advantage of this insight, we extended the *DS* model so that it takes label dependency into account to simultaneously estimate multiple true labels for each instance given multi-label annotations.

[1] http://www.lancers.jp.