Expert Systems With Applications 58 (2016) 155–163

Contents lists available at ScienceDirect

Expert Systems With Applications

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

Quality control of crowdsourced classification using hierarchical class structures[☆]

Naoki Otani[*], Yukino Baba, Hisashi Kashima

Graduate School of Informatics, Kyoto University, 36-1 Yoshida-Honmachi, Sakyo-ku, Kyoto, 606-8501, Japan

ARTICLE INFO

Article history:
Received 28 December 2015
Revised 5 April 2016
Accepted 6 April 2016
Available online 8 April 2016

Keywords:
Crowdsourcing
Quality control
Hierarchical classification
Human computation

ABSTRACT

Crowdsourcing is an emerging approach to utilize a large pool of human workers and execute various intelligent tasks. Repeated labeling is a widely adopted quality control method in crowdsourcing. This method is based on selecting one reliable label from multiple labels collected by workers because a single label from only one worker has a wide variance of accuracy. Hierarchical classification, where each class has a hierarchical relationship, is a typical task in crowdsourcing and used to organize information in many knowledge systems. However, direct applications of existing methods designed for multi-class classification have the disadvantage of discriminating among a large number of classes. In this paper, we propose a label aggregation method for hierarchical classification tasks. Our method takes the hierarchical structure into account to handle a large number of classes and estimate worker abilities more precisely.

Our method is inspired by the steps model based on item response theory, which models responses of examinees to sequentially dependent questions. We considered the hierarchical classification to be a question consisting of a sequence of sub-questions and built a worker response model for hierarchical classification. We conducted experiments using real crowdsourced hierarchical classification tasks for book classification and business classification and demonstrated the benefit of incorporating a hierarchical structure to improve the label aggregation accuracy. Our method also improves the accuracy for multi-class classification task for adult content classification with an implicit hierarchical structure among classes.

1. Introduction

Crowdsourcing is the process of requesting tasks from a large number of workers to obtain needed contents and provides an easy and relatively cheap way to exploit the power of crowds. For example, the construction of annotations for a large dataset, which is essential for the development of natural language processing techniques but expensive, has been crowdsourced (Callison-Burch & Dredze, 2010; Snow, O'Connor, Jurafsky, & Ng, 2008). Crowdsourcing is also being considered as an approach to solving computer science problems that are difficult for computers but comparatively easy for humans. A successful example is protein structure prediction (Cooper et al., 2010), which is computationally difficult because of the huge number of possible patterns but is easily tackled by humans.

An important issue with crowdsourcing is quality control; workers are often unreliable because they have a wide range of expertise and motivation unlike experts. There may be spam workers, who are extremely unreliable. A widely adopted approach to this problem is repeated labeling, which is obtaining labels from multiple workers for each task and aggregating them. Majority voting is the most straightforward way to aggregate labels. However, when there is access to only a few labels, the results may lack reliability. To increase the reliability, labels need to be gathered from more workers, which require more time and money.

In order to reduce such costs, more sophisticated methods have been proposed. These methods incorporate probabilistic models of the worker labeling process and achieve higher accuracy than majority voting even when not many labels are available. Some methods weight skilled workers under the assumption that more competent workers have a higher probability of success (Dawid & Skene, 1979; Demartini, Difallah, & Cudré-Mauroux, 2013; Welinder, Branson, Perona, & Belongie, 2010; Whitehill, Wu, Bergsma, Movellan, & Ruvolo, 2009). While multi-class classification tasks

[☆] Part of this work has been published in preliminary form in the proceedings of the 2015 IEEE International Conference on Data Mining (ICDM).

[*] Corresponding author. Tel.: +81 75 753 5347.

E-mail addresses: otani.naoki.65v@st.kyoto-u.ac.jp (N. Otani), baba@i.kyoto-u.ac.jp (Y. Baba), kashima@i.kyoto-u.ac.jp (H. Kashima).

http://dx.doi.org/10.1016/j.eswa.2016.04.009
0957-4174/© 2016 Elsevier Ltd. All rights reserved.

have been the main focus, other studies have considered quality control methods for ordering tasks (Chen, Bennett, Collins-Thompson, & Horvitz, 2013; Matsui, Baba, Kamishima, & Kashima, 2014; Yi, Jin, Jain, & Jain, 2013), sequence labeling tasks (Wu, Fan, & Yu, 2012), and numerical tasks (Lin, Mausam, & Weld, 2012).

In this paper, we propose a probabilistic labeling model for *hierarchical classification tasks*, where the classes are organized into a class hierarchy. Hierarchical classification is a useful and ubiquitous way to classify information and goods. It includes many important practical tasks, such as book classification in libraries and product classification in marketplaces. ImageNet (Deng et al., 2009) is an online image database and organizes images according to WordNet (Miller, 1995), which is a lexical database that uses a hierarchical structure. Hierarchical classification has also been used to simplify the design of a classification process such as astronomy classification tasks on Galaxy Zoo II (Willett et al., 2013).

Although hierarchical classification is a typical crowdsourcing task, no label aggregation methods exist for it. Because hierarchical classification can be considered multi-class classification, previous approaches that focused on multi-class classification can be applied. However, the number of classes in hierarchical classification often becomes very large, and the previous approaches are not designed to handle a large number of classes. By exploiting the information of the hierarchical structure, we propose bundling multiple classes having the same parent to efficiently deal with the large number of classes.

Here we provide a book classification example to explain how the hierarchical structure is important. Assume that we query five workers for a single hierarchical classification task. Two classify it as “social issues”, another two answer “pharmaceutical science”, and the last chooses “medical science”, as shown in Fig. 1(a). If we simply take a majority, “social issues” and “pharmaceutical science” are considered more reliable than “medical science” because they both get two votes. However, “medical science” and “pharmaceutical science” have the same parent (i.e., “science & technology”), while “social issues” does not. When we look at the intermediate classes, “science & technology” has three votes, and “society” has two. Therefore, although “pharmaceutical science” and “social issues” have the same number of votes, the former seems more likely to be correct than the latter. The workers who voted for “pharmaceutical science” can also be considered to classification abilities to some extent because they successfully selected the parent class.

Based on this insight, this paper presents a novel label aggregation method using a hierarchical structure. We employ the idea of a *steps model* based on item response theory to model the labeling process of workers for hierarchical classification (Section 4). We call our approach *Steps-GLAD*, which was inspired by the steps model and *GLAD*, a probabilistic label aggregation method presented by Whitehill et al. (2009).

We regard hierarchical classification to be a question consisting of a sequence of sub-questions, where each sub-question corresponds to a classification at each tier. The probability of successful classification is expressed as a function of the worker ability and sub-question difficulty. While the steps model is designed for a case where the true answers of the test questions are given, we targeted a crowdsourcing situation where we do not know the true labels and aimed to infer them from a set of observed labels. We use the expectation-maximization (EM) algorithm to infer the worker abilities, classification difficulties, and true labels (Section 5). Note that we applied GLAD because of its simplicity and strong connection to item response theory. Our framework is complementary to existing label aggregation models and can be easily combined with them.

We conducted experiments on three real datasets collected with crowdsourcing to demonstrate that Steps-GLAD works better than existing label aggregation methods (Section 6). We performed book classification and business classification, which are both two-tiered hierarchical classifications and have a large number of classes. In addition, we investigated whether our method is applicable to multi-class classification by assuming a hierarchical relationship among classes. We used an adult content classification dataset for this purpose.

Our contribution is summarized as follows:

- We propose a novel label aggregation method for crowdsourced hierarchical classification tasks, where the number of classes often becomes so large that it is difficult for previous methods to deal with.
- We introduce a probabilistic labeling model for hierarchical classification. Our model, which is inspired by the steps model based on item response theory, captures the worker classification abilities in intermediate classes by incorporating a hierarchical structure among classes.
- The experimental results demonstrated that the hierarchical structure is beneficial for obtaining accurate labels in hierarchical classification tasks and is helpful for estimating worker abilities more precisely. We also proved that our method delivers accurate labels for a multi-class classification task where an implicit hierarchy among the classes can be assumed.

2. Related work

A number of methods have been proposed to obtain high quality outputs with less cost by aggregating fewer labels compared to majority voting. These methods probabilistically model the labeling process of workers based on various aspects of the labeling process such as worker abilities (Dawid & Skene, 1979; Demartini et al., 2013; Welinder et al., 2010; Whitehill et al., 2009). Welinder et al. (2010) proposed a multi-dimensional worker ability model that captures the strengths and weaknesses of workers. Some methods incorporate task difficulties by considering that more difficult tasks have a lower probability of success (Welinder et al., 2010; Whitehill et al., 2009). Experiments on real datasets have shown that these probabilistic models can achieve higher accuracy than majority voting.

Our research focused on hierarchical classification, which has been applied to domains such as machine learning, natural language processing, and bioinformatics (Silla & Freitas, 2011). In hierarchical classification, a characteristic issue is that a misclassification at a given tier is propagated downwards. We can consider the classification at each tier as a sub-task and the whole classification process as a pipeline of sub-tasks. Improving the quality of the pipeline process is an important problem and has been the subject of many studies. Darling, Archambeau, Mirkin, and Bouchard (2013) presented a method to predict the root of errors.

A few crowdsourcing studies have focused on hierarchical classification. Chilton, Little, Edge, Weld, and Landay (2013) and Bragg and Weld (2013) presented methods to optimize the workflow to create a taxonomic hierarchy. Kamar and Horvitz (2015) proposed a probabilistic model to balance the output quality and cost by controlling queries to workers. So far, there have been no works on modeling the worker response process in hierarchical classification tasks and using the model to obtain accurate labels.

Several existing studies have addressed quality control for interdependent tasks (Duan, Oyama, Sato, & Kurihara, 2014; Mo, Zhong, & Yang, 2013; Wu et al., 2012). Our work differs because we considered a specific kind of tasks where the success of one task is a prerequisite for the success of another task, and such interdependency is given as prior knowledge coded as a hierarchy.