Separate or Joint? Estimation of Multiple Labels from Crowdsourced Annotations

Lei Duan, Satoshi Oyama, Haruhiko Sato, and Masahito Kurihara

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

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.

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

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; Whitehill et al., 2009; Welinder et al., 2010; Oyama et al., 2013) 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 & 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 et al., 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 & 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.

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

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

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

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

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

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, an instance with *fear* may co-exist with a certain degree of *anger* and *surprise*, *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.

3.1. Label-dependent DS (D-DS) model

As a first step, we focus on the assumption that labels depend on each other. To depict the dependency relationships among candidate labels, we introduce the concept of *conjoint-label*. Let J be the set of candidate labels. A conjoint-label represents a subset of J. For example, in the affect annotation experiment, the two conjoint-labels {*happiness*, *relief*} and {*happiness*, *excitement*} express two different kinds of "happiness": one is comparatively mild while the other is strong.

Let $T_i = \hat{J}\left(i \in I, \hat{J} \subseteq J\right)$ denote that $\hat{J}$ is the true conjoint-label for instance i, and let $n_{i\hat{L}}^k \in \mathbb{N}\left(k \in K, i \in I, \hat{L} \subseteq J\right)$ be the number of times that worker k annotated instance i with conjoint-label $\hat{L}$, which can be directly calculated from the crowdsourced annotations. Our goal is to estimate the set of true conjoint-labels $\{T_i\}_{i\in I}$, i.e., the multiple true labels for each instance, given the set of multi-label annotations $\left\{n_{i\hat{L}}^k\right\}_{k\in K, i\in I, \hat{L}\subseteq J}$. Similar to that of the *DS* model, the true conjoint-label for instance i is the one that achieves the maximum likelihood given the annotations for instance i:

$$\arg\max_{\hat{J}\subseteq J} P\left(T_i = \hat{J} \mid \left\{n_{i\hat{L}}^k\right\}_{k\in K, \hat{L}\subseteq J}\right). \tag{3}$$

Therefore, the *D-DS* model is an implementation of the *LP* method.

It is self-evident that $2^{|J|}$ conjoint-labels can be derived from J. We now describe the estimation of the $2^{|J|}$ posterior probabilities in formula (3) for each instance in I . By Bayes' Theorem, we have

$$P\left(T_i = \hat{J} \mid \left\{n_{i\hat{L}}^k\right\}_{k\in K, \hat{L}\subseteq J}\right) = \frac{P\left(\left\{n_{i\hat{L}}^k\right\}_{k\in K, \hat{L}\subseteq J} \mid T_i = \hat{J}\right) P\left(T_i = \hat{J}\right)}{P\left(\left\{n_{i\hat{L}}^k\right\}_{k\in K, \hat{L}\subseteq J}\right)}. \tag{4}$$

In the *DS* model, each clinician's predilections, which are called *error rates*, are captured in confusion matrix π, where π_{jl}^k specifies how likely clinician k will declare a patient to be in state l when the patient is actually in state j. In the *D-DS* model, $\pi_{\hat{J}\hat{L}}^k\left(k \in K, \hat{J} \subseteq J, \hat{L} \subseteq J\right)$ represents the probability that worker k assigns conjoint-label $\hat{L}$ when the true conjoint-label is $\hat{J}$. The

numbers of times that worker k annotated instance i with different conjoint-labels $\hat{L} \subseteq J$ when $\hat{J}$ is the true conjoint-label are distributed according to a multinomial distribution, i.e.,

$$P\left(\left\{n_{i\hat{L}}^k\right\}_{\hat{L}\subseteq J} \mid T_i = \hat{J}\right) = \frac{\left(\sum_{\hat{L}\subseteq J} n_{i\hat{L}}^k\right)!}{\prod_{\hat{L}\subseteq J}\left(n_{i\hat{L}}^k!\right)} \prod_{\hat{L}\subseteq J} \left(\pi_{\hat{J}\hat{L}}^k\right)^{n_{i\hat{L}}^k}.$$

We assume that, given the true conjoint-label, each worker's capability to make the correct judgment is conditionally independent of that of other workers, i.e.,

$$\begin{aligned} P\left(\left\{n_{i\hat{L}}^k\right\}_{k\in K,\hat{L}\subseteq J} \mid T_i = \hat{J}\right) &= \prod_{k\in K} P\left(\left\{n_{i\hat{L}}^k\right\}_{\hat{L}\subseteq J} \mid T_i = \hat{J}\right) \\ &= \prod_{k\in K}\left(\frac{\left(\sum_{\hat{L}\subseteq J} n_{i\hat{L}}^k\right)!}{\prod_{\hat{L}\subseteq J}\left(n_{i\hat{L}}^k!\right)} \prod_{\hat{L}\subseteq J} \left(\pi_{\hat{J}\hat{L}}^k\right)^{n_{i\hat{L}}^k}\right). \end{aligned} \tag{5}$$

Let $p_{\hat{J}}\left(\hat{J}\subseteq J\right)$ be the prior probability that an instance drawn at random has true conjoint-label $\hat{J}$, i.e.,

$$P\left(T_i = \hat{J}\right) = p_{\hat{J}}. \tag{6}$$

Different conjoint-labels being true for instance i are mutually exclusive events. From the law of total probability, we have

$$\begin{aligned} P\left(\left\{n_{i\hat{L}}^k\right\}_{k\in K,\hat{L}\subseteq J}\right) &= \sum_{\hat{J}\subseteq J} P\left(\left\{n_{i\hat{L}}^k\right\}_{k\in K,\hat{L}\subseteq J} \mid T_i = \hat{J}\right) P\left(T_i = \hat{J}\right) \\ &= \sum_{\hat{J}\subseteq J}\left(\prod_{k\in K}\left(\frac{\left(\sum_{\hat{L}\subseteq J} n_{i\hat{L}}^k\right)!}{\prod_{\hat{L}\subseteq J}\left(n_{i\hat{L}}^k!\right)} \prod_{\hat{L}\subseteq J} \left(\pi_{\hat{J}\hat{L}}^k\right)^{n_{i\hat{L}}^k}\right) p_{\hat{J}}\right). \end{aligned} \tag{7}$$

Finally, by substituting Equations (5), (6), and (7) into Equation (4), we can

estimate the posterior probabilities in formula (3) using

$$
\begin{aligned}
P\left(T_i=\hat{J}\mid\left\{n_{i\hat{L}}^k\right\}_{k\in K,\hat{L}\subseteq J}\right)=&\frac{\left(\prod_{k\in K}\left(\frac{\left(\sum_{\hat{L}\subseteq J}n_{i\hat{L}}^k\right)!}{\prod_{\hat{L}\subseteq J}\left(n_{i\hat{L}}^k!\right)}\prod_{\hat{L}\subseteq J}\left(\pi_{\hat{J}\hat{L}}^k\right)^{n_{i\hat{L}}^k}\right)\right)p_{\hat{J}}}{\sum_{\hat{J}\subseteq J}\left(\prod_{k\in K}\left(\frac{\left(\sum_{\hat{L}\subseteq J}n_{i\hat{L}}^k\right)!}{\prod_{\hat{L}\subseteq J}\left(n_{i\hat{L}}^k!\right)}\prod_{\hat{L}\subseteq J}\left(\pi_{\hat{J}\hat{L}}^k\right)^{n_{i\hat{L}}^k}\right)p_{\hat{J}}\right)}\\
=&\frac{\left(\prod_{k\in K}\prod_{\hat{L}\subseteq J}\left(\pi_{\hat{J}\hat{L}}^k\right)^{n_{i\hat{L}}^k}\right)p_{\hat{J}}}{\sum_{\hat{J}\subseteq J}\left(\left(\prod_{k\in K}\prod_{\hat{L}\subseteq J}\left(\pi_{\hat{J}\hat{L}}^k\right)^{n_{i\hat{L}}^k}\right)p_{\hat{J}}\right)}.
\end{aligned}
\tag{8}
$$

3.2. Label pairwise DS (P-DS) model

It is generally agreed that if two labels are similar or opposite, they are strongly correlated. Let us consider an attendance intention annotation task with four candidate labels, *have attended*, *plan to attend*, *want to attend*, and *no intention of attending*. It is obvious that the first two labels are strongly correlated, as are the last two, because someone who has already attended an activity (like an annual festival) would not likely plan to attend again, and someone who has no intention of attending would also be unlikely to want to attend. However, the four labels are not so generally correlated. Unfortunately, neither the "independent assumption" of the *DS* model nor the "dependent assumption" of the *D-DS* model can properly depict the relationships among these four labels. In view of this, we propose grouping candidate labels into pairs and then estimating the states of the two labels within each pair separately using $|J|/2$ independent models, each of which can be seen as a "two-label version of the *D-DS* model", in order to prevent interference from uncorrelated labels.

The crucial problem with the *P-DS* model is how to pair the candidate labels. Let $H(a,b)\,(a\in J,b\in J,a\neq b)$ be the joint entropy of labels a and b. The optimal pairing pattern S is the one that minimizes the sum of the joint entropies of all label pairs:[2]

$$
\arg\min_{S}\sum_{(a,b)\in S}H(a,b).
$$

[2]For a detailed proof of this, see AppendixA.

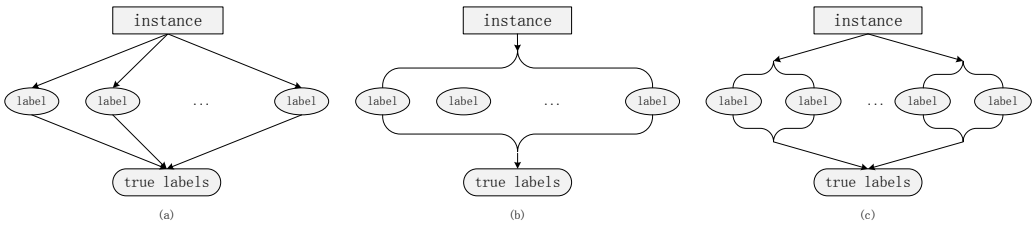

Figure 1: Multi-label estimation models: (a) *DS*, (b) *D-DS*, and (c) *P-DS*

Our experiments demonstrated that most label pairs contain synonymous or antonymous labels. In fact, the pairing pattern described above for the four intention labels is the optimal one for handling the intention annotation experiment, as described in Section 6.

The differences among the *DS*, *D-DS*, and *P-DS* models are illustrated in Figure 1.

4. Inference algorithm

Let us take the *D-DS* model as an example because the *P-DS* model can be considered to consist of several two-label *D-DS* models. Let $q_{i\hat{J}}\left(i \in I, \hat{J} \subseteq J\right)$ represent the posterior probability in formula (3), which means

$$q_{i\hat{J}} = P\left(T_i = \hat{J} \mid \left\{n_{i\hat{L}}^k\right\}_{k\in K, \hat{L}\subseteq J}\right).$$

Similar to the approach for the *DS* model, we use an EM-based algorithm to obtain the maximum likelihood estimates of model parameters $\left\{\pi_{\hat{J}\hat{L}}^k\right\}_{k\in K,\hat{J}\subseteq J,\hat{L}\subseteq J}$ and $\{p_{\hat{J}}\}_{\hat{J}\subseteq J}$, with the probabilities of true conjoint-labels $\{q_{i\hat{J}}\}_{i\in I,\hat{J}\subseteq J}$ as latent variables.

We then proceed as follows:

(1)[3] Obtain the initial estimates of unobserved variables $\{q_{i\hat{J}}\}_{i\in I,\hat{J}\subseteq J}$.

(2) **M-step:** Estimate the maximum likelihood estimates of parameters $\left\{\pi_{\hat{J}\hat{L}}^k\right\}_{k\in K,\hat{J}\subseteq J,\hat{L}\subseteq J}$ and $\{p_{\hat{J}}\}_{\hat{J}\subseteq J}$ by using the current expectations of

[3]This step is discussed in more detail in section 5.

$\{q_{i\hat{J}}\}_{i\in I,\hat{J}\subseteq J}$:

$$\pi^k_{\hat{J}\hat{L}} = \frac{\sum_{i\in I} q_{i\hat{J}} n^k_{i\hat{L}}}{\sum_{\hat{L}\subseteq J}\sum_{i\in I} q_{i\hat{J}} n^k_{i\hat{L}}},$$
$$p_{\hat{J}} = \frac{\sum_{i\in I} q_{i\hat{J}}}{|I|}.$$

(3) **E-step:** Estimate the expected values of $\{q_{i\hat{J}}\}_{i\in I,\hat{J}\subseteq J}$ using Equation (8) with the current estimates of parameters $\left\{\pi^k_{\hat{J}\hat{L}}\right\}_{k\in K,\hat{J}\subseteq J,\hat{L}\subseteq J}$ and $\{p_{\hat{J}}\}_{\hat{J}\subseteq J}$.

(4) Alternately perform steps (2) and (3) until the likelihood for all annotations $P\left(\left\{n^k_{i\hat{L}}\right\}_{k\in K,i\in I,\hat{L}\subseteq J}\right)$ converge. At this point, the $\hat{J}$ with the maximum $q_{i\hat{J}}$ is the true conjoint-label for instance i. Since all instances are annotated independently, from Equation (7), we have

$$\begin{aligned} P\left(\{n^k_{i\hat{L}}\}_{k\in K,i\in I,\hat{L}\subseteq J}\right) &= \prod_{i\in I} P\left(\{n^k_{i\hat{L}}\}_{k\in K,\hat{L}\subseteq J}\right) \\ &= \prod_{i\in I}\left(\sum_{\hat{J}\subseteq J}\left(\prod_{k\in K}\left(\frac{\left(\sum_{\hat{L}\subseteq J} n^k_{i\hat{L}}\right)!}{\prod_{\hat{L}\subseteq J}\left(n^k_{i\hat{L}}!\right)}\prod_{\hat{L}\subseteq J}\left(\pi^k_{\hat{J}\hat{L}}\right)^{n^k_{i\hat{L}}}\right) p_{\hat{J}}\right)\right). \end{aligned}$$

To avoid the "zero frequency problem" in step (2), π is estimated using Lidstone smoothing. Note that if worker k annotated only a certain instance with conjoint-label $\hat{L}$ one time and did not annotate any other instances, for k's *error rate* matrix, $\pi^k_{\hat{J}\hat{L}} = 1$ and $\pi^k_{\hat{J}\hat{L}'} = 0$ $(\hat{J}\subseteq J, \hat{L}'\subseteq J, \hat{L}'\neq\hat{L})$ constantly within iterations. Therefore, to estimate a worker's *error rate* matrix, at least two annotations of that worker must be collected.

5. Discussion: Bayesian network label-dependent *DS* (*ND-DS*) model

Recall that there is an unsolved problem in the first step of the EM algorithm described in Section 4: how to initialize the estimates of unobserved variables $\{q_{i\hat{J}}\}_{i\in I,\hat{J}\subseteq J}$. Let $x^j_{i\hat{J}} \in \{0,1\}\left(j\in J, i\in I, \hat{J}\subseteq J\right)$ be the state of the jth label in conjoint-label $\hat{J}$ for instance i. One possible and intuitive way to initialize the estimates is to assign

$$q_{i\hat{J}} = P\left(\left\{x^j_{i\hat{J}}\right\}_{j\in J}\right) = \frac{\sum_{k\in K} n^k_{i\hat{J}}}{\sum_{\hat{L}\subseteq J}\sum_{k\in K} n^k_{i\hat{L}}}, \tag{9}$$

which is the maximum likelihood estimate of $q_{i\hat{J}}$ and which is indeed used in the *D-DS* model. This is equivalent to estimating a $|J|$-dimensional joint distribution for each instance over the candidate labels. Because the label states are binary-valued, the joint distribution requires the probabilities of $2^{|J|}$ different assignments of values. For all but the smallest $|J|$, the explicit representation of the joint distribution is unmanageable from every perspective. At the practical level, it is too expensive and nearly impossible to acquire a sufficient number of samples from workers to robustly estimate the high-dimensional joint distribution. This means that the *D-DS* model can easily suffer from the sparsity of high-dimensional annotations. To overcome this problem, it is better to represent the distribution more compactly, and to approximate the underlying joint distribution from a finite number of samples by using the conditional independence properties of the joint distribution.

To motivate our discussion, we first assume that all candidate labels are statistically independent. That is, the completely general joint distribution in Equation (9) can be approximated as an independent distribution over candidate labels:

$$q_{i\hat{J}} = \prod_{j \in J} P\left(x_{i\hat{J}}^{j}\right). \tag{10}$$

Intuitively, this simple assumption of ignoring the dependency relationships among candidate labels is unreasonable in most cases, as we explained at the beginning of Section 3. There have been several proposals for approximating high-dimensional joint distributions. Chow & Liu (1968), for example, addressed this problem by approximating an n-dimensional joint distribution as the product of $n-1$ second-order component distributions, where the relationships among random variables are represented by a *dependence tree*. Here we represent label dependency as a Bayesian network and call this extended *D-DS* model the "Bayesian network *D-DS* (*ND-DS*) model". Figure 2 shows an example Bayesian network for our affect annotation experiment. The corresponding approximate product of the joint distribution is

$$\begin{aligned} P(X) = P(x_{su})\,P(x_{sa})\,P(x_{di})\,P(x_{re} \mid x_{su}, x_{sa})\,P(x_{ha} \mid x_{sa})\,P(x_{an} \mid x_{sa}, x_{di}) \\ P(x_{fe})\,P(x_{fo} \mid x_{ha})\,P(x_{sh} \mid x_{fe})\,P(x_{ex}). \end{aligned}$$

Since the number of annotations for one instance is not sufficient for learning a Bayesian network, in the *ND-DS* model, all instances are assumed to share an identical Bayesian network, which is learned from the annotations

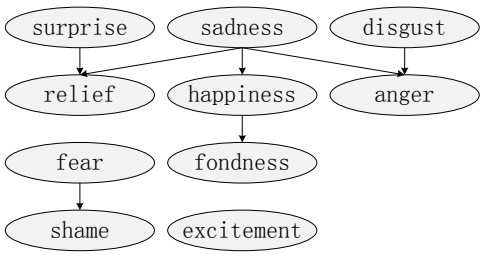

Figure 2: Example Bayesian network for affect annotation experiment

for all instances. This is reasonable because the relationships among candidate labels are independent of instances. We build the network structure of candidate labels using the "PC" algorithm (Spirtes et al., 2000), which is based on hypothesis testing. To test whether two labels x_a and x_b are conditionally independent given a subset of other labels $\bar{X}$, we compute the conditional mutual information of these two labels,

$$\mathrm{CMI}\left(x_a; x_b \mid \bar{X}\right) = \sum_{\bar{X}} P\left(\bar{X}\right) \sum_{x_a, x_b} P\left(x_a, x_b \mid \bar{X}\right) \log \frac{P\left(x_a, x_b \mid \bar{X}\right)}{P\left(x_a \mid \bar{X}\right) P\left(x_b \mid \bar{X}\right)},$$

by using the maximum likelihood estimates on annotations for all instances. Under the independence assumption, $2m * \mathrm{CMI}\left(x_a; x_b \mid \bar{X}\right)$ follows a χ^2 distribution with degrees of freedom equal to $2^{|\bar{X}|}$, where m is the sample size $|I| * |K|$.

Although we take the p-value for rejecting the null hypothesis that two labels are conditionally dependent as 0.1, it is worth mentioning that if the p-value is 1, all labels are determined to be unconditionally independent of each other, and the approximation strategy of the *ND-DS* model is the same as Equation (10). Likewise, the *ND-DS* model is equivalent to the *D-DS* model if the p-value is 0, which means that the network structure is a complete directed acyclic graph, and the depicted approximate product of the joint distribution is the chain rule for Equation (9).

In summary, we proposed two models, *D-DS* and *P-DS*, for estimating multiple true labels for each instance given crowdsourced multi-label annotations. Moreover, we extended the *D-DS* model to create the *ND-DS* model, using the Bayesian network to approximate the joint distribution over the candidate labels.

6. Empirical study

6.1. Affect annotation for lines in story

To create a first test bed for the proposed models containing actual annotations obtained from the Lancers crowdsourcing service, we asked crowdsourcing workers to read some story lines and spontaneously indicate the character's affects (emotions) generated by each line and then estimated the true affects for each line on the basis of the obtained multi-label annotations. To simplify the task, we needed stories in which the lines express clear affects. Since children typically have an elementary level of psychological development, stories written for them usually have vibrant affection tint, distinct character personalities, and higher proportion of lines than other types of stories. The aim is to better attract the attention of children. Therefore, children's stories and fairy tales are commonly used in affect analysis (Alm et al., 2005; Mohammad, 2011). We thus chose two Japanese children's stories, "Although we are in love"[4] ("Love" for short) and "Little Masa and a red apple"[5] ("Apple" for short), as the texts to be annotated. As the source of the texts we used the Aozora Library[6]. While "the Big Six" affects (i.e., *happiness*, *fear*, *anger*, *surprise*, *disgust*, and *sadness*) and the related affect sets are typically used in affective computing research (Alm et al., 2005; Alm, 2010; Trohidis et al., 2008), we used ten affects as the candidate labels in order to provide more choices for the workers and thereby enable us to perform a more in-depth study on multi-label estimation. The affects were taken from the "Emotive Expression Dictionary" (Nakamura, 1993). An example task input screen is shown in Figure 3. If none of the listed affects was felt, the worker could check *neutral*.

People have different tendencies when detecting subjective feelings, so two people may be affected differently by the same line. This means that, for the affect labels to be reliable, they should be in accord with the general consensus of large crowds. The *majority vote* strategy most objectively reflects the general consensus if the number of workers is large enough. Therefore, we obtained gold standards by having each line annotated 30 times and then taking the *majority vote*. That is, the most often annotated conjoint-label for a line was used as the gold standard for that line.

[4]http://www.aozora.gr.jp/cards/001475/files/52111_47798.html
[5]http://www.aozora.gr.jp/cards/001475/files/52113_46622.html
[6]http://www.aozora.gr.jp

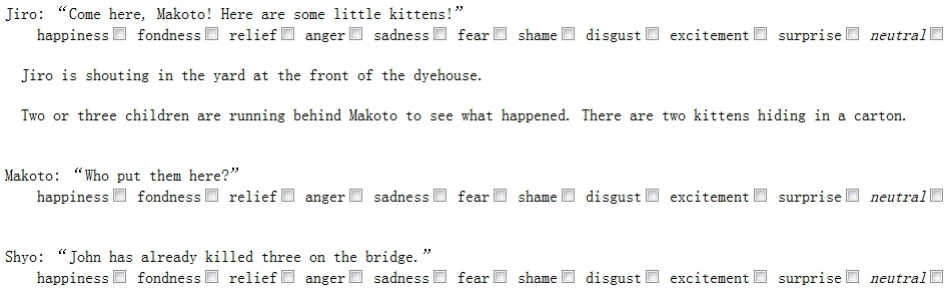

Figure 3: Example task input screen (translated from original Japanese)

For the “Love” story, we asked each of 30 workers to annotate each line one time, which ensured that each worker annotated the complete set of lines. For the “Apple” story, the workers were not specifically selected, so the 30 annotations for every line were provided by arbitrary workers, and few, if any, of them annotated all the lines. This is a more realistic situation since it is not a good idea to submit a very large task to a crowdsourcing service because a large task tends to diminish worker enthusiasm or even cause workers to avoid the task. We conducted the “Apple” task in this way simply to examine the effects of “arbitrary worker interference” on the model results.

Moreover, although our proposed models can handle a line being annotated more than once by a worker, to collect opinions as widely as possible at a fixed cost, it is still best to avoid this situation even though a worker may interpret a line differently at different times. Therefore, in our experiments, all the annotations for a line were obtained from different workers. This means that n_{ii}^k in formula (2) and $n_{i\hat{L}}^k$ in formula (3) are either 0 or 1. The annotation frequencies of the affect labels are shown in Table 1, and other statistics about the datasets are shown in Table 2.

To determine the effect of the number of workers per line on accuracy, we randomly split the workers who annotated a particular line into various numbers of groups of equal size and estimated the reliable affect labels for each line given the annotations within each group. We did this for five different group sizes: 3 (ten groups), 5 (six groups), 10 (three groups), 15 (two groups), and 30 (one group). Since both the estimation result and the gold standard for a line can be regarded as a binary vector, the performance evaluation of the proposed models is the average *simple matching coefficient*, i.e., the average proportion of correct labels between the gold standard and

Table 1: Annotation frequencies of affect labels and *neutral*

Affect label	Freq. in "Love"	Freq. in "Apple"	Total
Relief	516	362	878
Anger	242	623	865
Sadness	522	298	820
Happiness	458	306	764
Fondness	467	226	693
Excitement	379	270	649
Disgust	279	265	544
Neutral	*120*	*352*	*472*
Surprise	190	243	433
Fear	164	107	271
Shame	84	68	152
Total (except *Neutral*)	3301	2768	6069

Table 2: Statistics for affect annotation experiment

	"Love"	"Apple"	Total
No. of workers	30	57	84
No. of lines	63	78	141
No. of annotations	1890	2340	4230
Avg. no. of checked labels per annotation	1.75	1.18	1.43
Avg. no. of annotations per line	30	30	30

the estimation result for all lines. The average accuracies for each group size were obtained with the *majority vote* strategy, the *DS*, *P-DS*, and *D-DS* models, and the *ND-DS* sub-model of the *D-DS* model.

As shown in Figure 4, for the "Love" story, the statistical models achieved better or comparable accuracy than the *majority vote* strategy when the group size was 3, 5, or 10. As shown in Figure 5, for the "Apple" story, the statistical models achieved better accuracy when the group size was 3 or 5, and two of them achieved better or comparable accuracy when the group size was 10. This means that ten workers at most for each line would be a reasonable number. Moreover, the *ND-DS* model consistently outperformed the *D-DS* model. This means that a learned Bayesian network is effective for

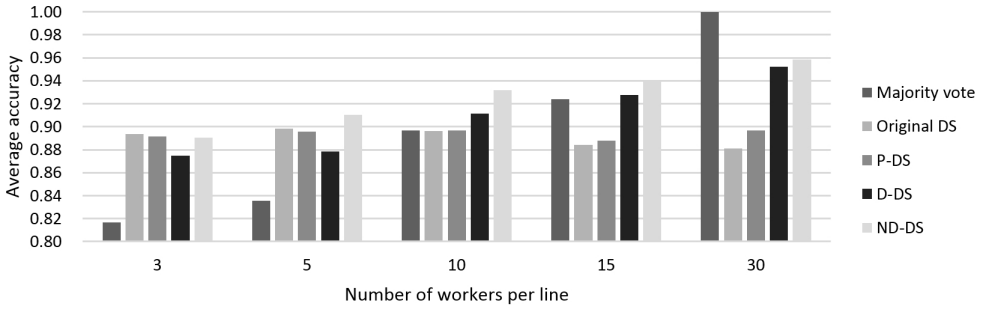

Figure 4: Average affect annotation accuracy for "Although we are in love" story

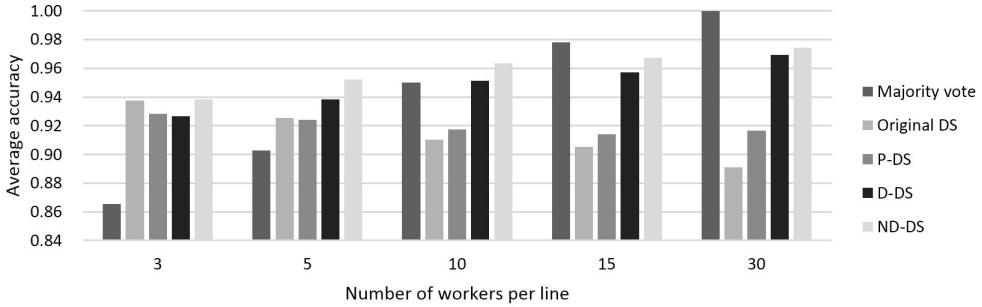

Figure 5: Average affect annotation accuracy for "Little Masa and a red apple" story

approximating the high-dimensional joint distribution over ten affect labels from a finite number of annotations. Although the *DS*, *P-DS*, and *ND-DS* models had virtually the same average accuracy for three workers per line, the *ND-DS* model had significantly better accuracy (greater than 90%) for five or more workers per line. In other words, the *ND-DS* model can most effectively handle the multi-label estimation problem with annotations provided by only about five crowdsourcing workers per instance. One noteworthy result is that the average accuracies of the *DS* and *P-DS* models remained basically unchanged as the group size increased for the "Love" task while they decreased for the "Apple" one. This could be because these two models are more sensitive to the effects of "arbitrary worker interference".

Table 3: Annotation frequencies of intention labels

Intention label	Frequency
have no intention of attending	2521
want to attend	2417
plan to attend	1365
have attended	1200
total	7503

Table 4: Statistics for intention annotation experiment

No. of workers	94
No. of tweets	1398
No. of annotations	6990
Avg. no. of checked labels per annotation	1.07
Avg. no. of annotations per tweet	5

6.2. Intention annotation for tweets

In the second experiment, intention annotation for tweets, we posted 1398 tweets on the Twitter micro-blogging service[7] related to the Sapporo Snow Festival. We again used the Lancers crowdsourcing service and asked workers to infer the attendance intentions of the tweet poster and then select appropriate ones from four intention labels: *have no intention of attending*, *want to attend*, *plan to attend*, and*have attended*. Each tweet was annotated by five arbitrary workers. The annotation frequencies of the intention labels are shown in Table 3, and other statistics about the dataset are shown in Table 4.

We manually assigned reliable labels to each tweet and used them as the gold standards. The performance of the proposed models was measured in the same way as in the affect annotation experiment. As shown in Figure 6, all the statistical models as well as the *majority vote* strategy performed well due to the simplicity of the task. Particularly noteworthy is that the *P-DS* model had the highest accuracy, followed by the *ND-DS* model. The superior performance of the *P-DS* model is attributed to the fact that the

[7]https://twitter.com

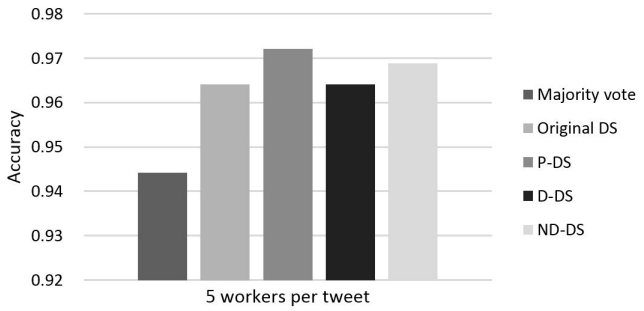

Figure 6: Average accuracies for intention annotation

four intention labels have a typical pairwise characteristic, as explained in Section 3.2.

7. Related work

7.1. Crowdsourcing and quality control

Crowdsourcing is an economical and efficient approach to performing tasks that are difficult for computers but relatively easy for humans. With the recent expansion of crowdsourcing platforms such as Amazon Mechanical Turk[8] and CrowdFlower[9], the concept of crowdsourcing has been successfully applied in various areas of computer science research, including natural language processing (Snow et al., 2008) and computer vision (Sorokin & Forsyth, 2008).

Because there is no guarantee that all workers are sufficiently competent to complete the offered tasks, ensuring the quality of the results is one of the biggest challenges in crowdsourcing. A simple strategy is giving monetary bonuses to high-performance workers and denying payments to low-performance ones. In addition, several approaches geared toward efficient quality control have been applied. For example, Amazon Mechanical Turk provides a pre-qualification system to assess the skill level of a prospective worker, and CrowdFlower enables requesters to inject a collection of tasks with known correct answers into their tasks to measure a worker's performance automatically.

[8]http://www.mturk.com
[9]http://crowdflower.com

Meanwhile, crowdsourcing service researchers have also explored sophisticated statistical strategies for ensuring the quality of crowdsourcing data obtained from noisy responses. Snow et al. (2008) demonstrated that, using an automatic bias correction algorithm, Amazon Mechanical Turk can be used effectively for a variety of natural language annotation tasks. Sheng et al. (2008) explored several methods for choosing which instances should get more labels, and how to include labels' uncertainty information when training classifiers. Whitehill et al. (2009) presented a model for simultaneously estimating the true label of each multi-labeled instance, the expertise of each worker, and the difficulty of each question. Lin et al. (2012) took a decision-theoretic approach to estimating the correct answer for a task that can have a countably infinite number of possible answers. Oyama et al. (2013) investigated the use of not only crowdsourced annotations but also workers' self-reported confidence scores to estimate the true label for each single-labeled instance. Baba et al. (2014) applied quality control techniques to the detection of crowdsourcing tasks considered to be improper by a crowdsourcing service. They showed that the accuracy of detecting improper tasks could be improved by combining non-expert judgments by crowdsourcing workers with expert judgments.

There has also been some research on the problem of multi-label estimation, the focus of this paper. Nowak & Rüger (2010) studied inter-annotator agreement for multi-label image annotation and found that using the *majority vote* strategy to generate one annotation set from several opinions can filter out noisy judgments of non-experts to some extent. However, they did not answer the question of how many crowdsourcing workers are needed to obtain quality comparable to that of expert annotators. Bragg et al. (2013) presented a decision-theoretic approach to taxonomy creation that implements the *BR* method. They showed that, with their approach, 16 workers per instance are sufficient to achieve quality comparable to the general consensus of a large crowd.

7.2. Affect prediction

People, by nature, can be emotionally affected by literature, music, fine art, etc. Predicting how we are affected is an important research direction in artificial intelligence as it is potentially applicable to many further applications, including expressive text-to-speech synthesis (Anderson et al., 2013) and therapeutic education of children with communication disorders (Dias et al., 2013). Many researchers have thus concentrated on this research area.

Alm et al. (2005) investigated the importance of various features for affect analysis and classified the emotional affinity of sentences in the narrative domain of children's fairy tales using the sparse network of winnows ($SNoW$) learning architecture. Trohidis et al. (2008) modeled the automated detection of emotion in music as a multi-label classification task. Alm (2010) analyzed the characteristics of sentences with high-agreement crowdsourced affect annotations. **?** did an experiment on affect analysis of certain characters in narrative text. Kim et al. (2013) introduced a continuous representation, called the *manifold*, for human emotions in sentiment analysis research and constructed a statistical model connecting it to documents and to a discrete set of emotions. A method for identifying emotions in micro-blog posts by using "emotion cause extraction" was proposed by Li & Xu (2014).

7.3. Intention inference

In today's Web 2.0 era, people post descriptions of their various real-world experiences such as visiting places, participating in activities, and shopping to social networking services, such as Twitter and Facebook[10]. Extracting such information from the huge amounts of real-time updated text corpora is important for estimating the popularity of places, activities, and products, and is of great value to navigation and recommendation systems. Lee & Sumiya (2010) developed a method for detecting geo-social events, such as local festivals, by monitoring crowd behaviors indirectly via Twitter. Liao et al. (2012) investigated whether and how micro-messaging technologies could be used to predict attendance trends at the World Expo 2010 in Shanghai. Xu et al. (2012) proposed using a *semidefinite programming* optimization technique for identifying valuable customers from social network services in terms of profit maximization.

8. Conclusion

We focused on crowdsourcing tasks fitting the paradigm of multi-label annotation, which means that an instance can have one or more true label(s). The three statistical quality control models we proposed for the multi-label estimation problem incorporate label dependency into the label-generation process. An EM-based algorithm is used to estimate the multiple true labels

[10]https://www.facebook.com

for each instance as well as the maximum likelihood estimates of the model parameters. Two experiments using Lancers crowdsourcing service showed that two of the models showed promising performance: in most cases, the Bayesian network label-dependent *DS* (*ND-DS*) model most effectively handled the annotations provided by about five crowdsourcing workers per instance. The label pairwise *DS* (*P-DS*) model was the most effective when there were pairwise comparison relationships among candidate labels.

Two widely used methods for multi-label classification are the binary relevance method and the label combination or label power-set method (Tsoumakas et al., 2010), which are the counterparts of the *DS* model and the *D-DS* model in this research. The *DS* model simply decomposes the multi-label estimation problem into several independent binary-label estimation problems, one for each label in the set of candidate labels, and final labels for each instance are determined by aggregating the predictions from all binary estimators. A significant limitation of this method is that they do not take into account any dependency among candidate labels. Since multi-label tasks often have many candidate labels, if we simply incorporate dependency relationships among all candidate labels into the label-generation process, as does the *D-DS* model, we may get data sets with a large number of classes and few samples per class. This means that the *D-DS* model can easily suffer from the sparsity of high-dimensional annotations, which makes the learning process difficult. Therefore, the *D-DS* model performs poorly for high-dimensional data sets.

To address these limitations, we proposed two approaches that flexibly use label dependency. In the first approach, the *P-DS* model is used to group candidate labels into pairs. The states of the two labels within each pair are then estimated separately in order to prevent interference from uncorrelated labels. The crucial problem is how to recognize pairwise comparison relationships among candidate labels. If the labels are pairwise correlated, the optimal pairing pattern should be the one that minimizes the sum of the joint entropies of all label pairs. The reason for this is explained in detail in AppendixA.

In the second approach, the underlying high-dimensional joint distribution over candidate labels is represented more compactly to enable it to be approximated from a finite number of annotations. The *ND-DS* model depicts the properties as a Bayesian network, enabling the joint distribution to be approximated using the product of the conditional distributions of the candidate labels. Because the dependency relationships among candidate labels are independent of instances, the network is learned from the annota-

tions for all instances. The superiority of these two approaches is shown by the experimental results.

Multi-label annotation is crucial for many comprehension-simulation techniques, e.g., affect prediction, intention inference, email analysis, and text, image, music, and movie semantic categorization. The annotation quality directly affects the performance of these techniques. Collecting high-quality annotations from both experts and large crowds can be expensive and time-consuming. The proposed models enable multiple true labels to be effectively estimated using the annotations provided by handful of crowdsourcing workers. This approach to obtaining multi-label datasets with quality approaching that of ones obtained from the general consensus of large crowds or from human experts is a promising way to reduce the cost of data collection for future applications with minimal degradation in the quality of the results.

Our work is an exploration of the human computation issue. Our promising results provide encouragement for further study to overcome the limitations of our present work. For one thing, each worker should label at least two instances because a worker's *error rate* matrix cannot be estimated with only one annotation, as mentioned in Section 4. This requirement may decrease the flexibility of crowdsourcing somewhat. In our research, every instance was assigned an equal number of workers. However, for simple instances, few (one or two) workers may be sufficient, so taking into account the difficulties of instances should further reduce annotation costs.

In view of these considerations, we plan to enhance our research efforts in several ways. First, our experiments were conducted on three small databases, especially the two stories. In future work we will explore the effect of using large datasets on the results. Another possible direction is the design of an effective mechanism for automatically identifying the difficulties of instances such as using the time needed for completing an instance. Other information, such as workers' self-reported confidence scores, which have shown an improvement recently (Oyama et al., 2013), and consistency of story emotionality and character personality for narrative annotation tasks, is also important for the label-generation process and worth studying. In the long-term, we plan to extend our work to other multi-label estimation problems, such as art genre recognition, which is thought to raise more variant opinions among people. Estimating labels not only from crowdsourced annotations but also from user-created content services is also an interesting direction for future work.

AppendixA. Proof of optimal label pairing pattern

Let $P(X)$ be the joint probability distribution over n labels x_1, x_2, ..., x_n, X denoting the n-dimensional vector $(x_1, x_2, ..., x_n)$. Under the condition that labels are pairwise correlated, the joint distribution over all labels takes the following form:

$$P'(X) = \prod_{\substack{i=1 \\ i<j(i)\leq n \\ i\neq j(i'),j(i)\neq j(i')\,(i'=1,\ldots,i-1)}}^{n-1} P\left(x_i, x_{j(i)}\right),$$

where $\left(x_i, x_{j(i)}\right)$ constitutes a label pair. The optimal pairing pattern is the one that minimizes the Kullback-Leibler divergence (Kullback & Leibler, 1951), which measures the difference between two probability distributions over the same event space, between $P(X)$ and $P'(X)$:

$$\begin{aligned} D(P \mid\mid P') &= \sum_X \log \frac{P(X)}{P'(X)} \\ &= -\sum_X P(X) \log \frac{1}{P(X)} - \sum_X P(X) \log P'(X) \\ &= -H(X) - \sum_X P(X) \sum_{\substack{i=1 \\ i<j(i)\leq n \\ i\neq j(i'),j(i)\neq j(i')\,(i'=1,\ldots,i-1)}}^{n-1} \log P\left(x_i, x_{j(i)}\right), \end{aligned} \tag{A.1}$$

where $H(X)$ on the right side is the joint entropy of all labels. Since $P\left(x_i, x_{j(i)}\right)$ is a component of $P(X)$,

$$-\sum_X P(X) \log P\left(x_i, x_{j(i)}\right) = -\sum_{x_i, x_{j(i)}} P\left(x_i, x_{j(i)}\right) \log P\left(x_i, x_{j(i)}\right),$$

where the right side of the equation is the joint entropy of label pair $\left(x_i, x_{j(i)}\right)$. Thus, Equation (A.1) becomes

$$D(P \mid\mid P') = -H(X) + \sum_{\substack{i=1 \\ i<j(i)\leq n \\ i\neq j(i'),j(i)\neq j(i')\,(i'=1,\ldots,i-1)}}^{n-1} H\left(x_i, x_{j(i)}\right).$$

Since $H(X)$ is independent of the pairing pattern, minimizing the Kullback-Leibler divergence $D(P \mid\mid P')$ is equivalent to minimizing the sum of the joint entropies of all label pairs:

$$\sum_{\substack{i=1 \\ i<j(i)\leq n \\ i\neq j(i'),j(i)\neq j(i')\,(i'=1,\ldots,i-1)}}^{n-1} H\left(x_i, x_{j(i)}\right).$$

If we depict the pairing pattern as an undirected graph, where labels are represented by vertices, and the weight of each edge is assigned the joint entropy of the two labels represented by the two vertices of the edge, the sum of the joint entropies of all label pairs can be minimized by finding the minimum-weight perfect matching of the graph. This solution can be achieved by using the Blossom algorithm (Edmonds, 1965).

Acknowledgments

This work was supported in part by JSPS KAKENHI 24650061.

Alm, C. O. (2010). Characteristics of high agreement affect annotation in text. In *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 118–122). Association for Computational Linguistics.

Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 579–586). Association for Computational Linguistics.

Anderson, R., Stenger, B., Wan, V., & Cipolla, R. (2013). Expressive visual text-to-speech using active appearance models. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on* (pp. 3382–3389). IEEE.

Baba, Y., Kashima, H., Kinoshita, K., Yamaguchi, G., & Akiyoshi, Y. (2014). Leveraging non-expert crowdsourcing workers for improper task detection in crowdsourcing marketplaces. *Expert Systems with Applications*, *41*, 2678–2687.

Bragg, J., Weld, D. S. et al. (2013). Crowdsourcing multi-label classification for taxonomy creation. In *First AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*.

Chow, C., & Liu, C. (1968). Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, *14*, 462–467.

Dawid, A. P., & Skene, A. M. (1979). Maximum likelihood estimation of observer error-rates using the em algorithm. *Applied Statistics*, (pp. 20–28).

Dias, M. d. R. D., Faria, S. d. S. B. L. d., Ibrahim, S. C. M. et al. (2013). Im like a river: a health education instrument for stuttering. *Revista de Psicologia da IMED*, *5*.

Edmonds, J. (1965). Paths, trees, and flowers. *Canadian Journal of Mathematics*, *17*, 449–467.

Kim, S., Li, F., Lebanon, G., & Essa, I. (2013). Beyond sentiment: The manifold of human emotions. *International Conference on Artificial Intelligence and Statistics (AISTATS)*, (pp. 360–369).

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*, 79–86.

Lee, R., & Sumiya, K. (2010). Measuring geographical regularities of crowd behaviors for twitter-based geo-social event detection. In *Proceedings of the 2nd ACM SIGSPATIAL International Workshop on Location Based Social Networks* (pp. 1–10). ACM.

Li, W., & Xu, H. (2014). Text-based emotion classification using emotion cause extraction. *Expert Systems with Applications*, *41*, 1742–1749.

Liao, Y., Moshtaghi, M., Han, B., Karunasekera, S., Kotagiri, R., Baldwin, T., Harwood, A., & Pattison, P. (2012). Mining micro-blogs: opportunities and challenges. In *Computational Social Networks* (pp. 129–159). Springer.

Lin, C. H., Mausam, & Weld, D. S. (2012). Crowdsourcing control: Moving beyond multiple choice. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI). Catalina Island, CA, USA. August 2012* (pp. 491–500).

Mohammad, S. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 105–114). Association for Computational Linguistics.

Nakamura, A. (1993). Kanjo hyogen jiten [Dictionary of Emotive Expressions] (in Japanese). *Tokyodo*, .

Nowak, S., & Rüger, S. (2010). How reliable are annotations via crowdsourcing: a study about inter-annotator agreement for multi-label image annotation. In *Proceedings of the International Conference on Multimedia Information Retrieval* (pp. 557–566). ACM.

Oyama, S., Baba, Y., Sakurai, Y., & Kashima, H. (2013). Accurate integration of crowdsourced labels using workers' self-reported confidence scores. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence* (pp. 2554–2560). AAAI Press.

Sheng, V. S., Provost, F., & Ipeirotis, P. G. (2008). Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 614–622). ACM.

Snow, R., O'Connor, B., Jurafsky, D., & Ng, A. Y. (2008). Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 254–263). Association for Computational Linguistics.

Sorokin, A., & Forsyth, D. (2008). Utility data annotation with amazon mechanical turk. *Urbana*, *51*, 820.

Spirtes, P., Glymour, C., & Scheines, R. (2000). *Causation, prediction, and search* volume 81. The MIT Press.

Trohidis, K., Tsoumakas, G., Kalliris, G., & Vlahavas, I. P. (2008). Multi-label classification of music into emotions. In *International Society for Music Information Retrieval (ISMIR)* (pp. 325–330). volume 8.

Tsoumakas, G., Katakis, I., & Vlahavas, I. (2010). Mining multi-label data. In *Data Mining and Knowledge Discovery Handbook* (pp. 667–685). Springer.

Welinder, P., Branson, S., Belongie, S., & Perona, P. (2010). The multidimensional wisdom of crowds. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 2424–2432). volume 10.

Whitehill, J., Wu, T.-f., Bergsma, J., Movellan, J. R., & Ruvolo, P. L. (2009). Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems* (pp. 2035–2043).

Xu, K., Li, J., & Song, Y. (2012). Identifying valuable customers on social networking sites for profit maximization. *Expert Systems with Applications*, *39*, 13009–13018.