

Bayesian Aggregation of Categorical Distributions with Applications in Crowdsourcing

Paper #1903

Abstract

A key problem in crowdsourcing is the aggregation of judgments of proportions. For example, workers might be presented with a news article or an image, and be asked to identify the proportion of each topic, sentiment, object, or colour present in it. These varying judgments then need to be aggregated to form a consensus view of the document's contents. Often, however, these judgments can be skewed by workers who provide judgments randomly (i.e. they are spammers). Spammers make the cost of acquiring judgments more expensive and degrade the accuracy of the aggregation. For such cases, we provide a new Bayesian framework for aggregating these responses (expressed in the form of categorical distributions) that for the first time accounts for spammers. We elicit 796 judgments about proportions of objects and colours in images. Experimental results on three real-world datasets show comparable aggregation accuracy when 60% of the workers are spammers, as other state of the art approaches do when there are no spammers.

1 Introduction

The emergence of crowdsourcing platforms such as Amazon Mechanical Turk (AMT), Crowdfunder and oDesk, has impacted a number of domains. In particular in areas such as sentiment analysis, citizen science, and digital humanitarianism, it is now possible to collect low-cost judgments rapidly with reduced reliance on domain experts. These judgments usually take the form of a single or multiple discrete labels, but can be, in general, any metadata attached to an item. A key problem of interest in this area is the aggregation of judgments of proportions by each worker [17, 7]. Text categorisation offers an example scenario. In this context, the task is to assign categories for a text document whose content may cover multiple topics (e.g. sport and politics) or evoke various sentiments (e.g. joy and surprise) [4]. The documents are assigned to humans who read the text and provide a judgment about the proportion of each category. For example, if a worker believes that

3/4 of a document evokes joy and 1/4 surprise, he will submit the following judgment {joy:75%,surprise:25%}. Another domain of interest is digital humanitarianism. Satellite imagery is increasingly being used to map deforestation and natural disasters rapidly by volunteers. However, image quality is often poor (e.g. cloud obstruction or low resolution) with features that are difficult to identify precisely by the volunteers. A volunteer who believes 3/4 of the image consists of rubble and 1/4 of undamaged buildings would provide the following judgment {rubble:75%,undamaged:25%}. In a different domain, the ability to accurately measure colour proportions is critical to industries that work with textiles, paints, food, design and marketing. However, accurately quantifying human perception of colours is surprisingly challenging in practice due to our implicit interpretation of illumination [22]; and adding memory demands compounds the difficulty [3]. A worker who remembered that half of an item was of colour indigo, 1/4 cyan, and 1/4 blue would provide the following judgment {indigo:50%,cyan:25%,blue:25%}.

While collecting judgments from individual workers is appealing, such methods raise issues of reliability, due to the unknown incentives of the participants. There are likely to be malicious workers (i.e. spammers) who provide judgments randomly. For example, in the sentiment analysis and colour proportions domain, workers may provide as many random judgments as possible for quick financial gain [5]. Similarly, in the digital humanitarian domain, maliciously volunteers may provide misleading reports to divert the attention of ecologists or emergency services. It has been estimated that up to 45% of workers on crowdsourcing platforms may fall into this category [19]. Thus, reliably aggregating multiple judgments from crowd workers is non-trivial.

To address these challenges, we regard the problem of judging proportions as one of probabilistic modelling where the judgments from the workers are categorical distributions [14]. Now, a number of approaches have been proposed to aggregate probability distributions. Perhaps the most common fall into the category known as *opinion pools*. In particular, the *linear opinion pool* (LinOp) [2] and the *logarithmic opinion pool* (LogOp) [20] are the most popular class of methods to aggregate distributions. They aggregate individual workers' judgments using a weighted arithmetic average and a

weighted geometric average respectively. They are simple, and frequently yield useful results with a moderate amount of computation. Alternatively, approaches based on *confusion matrices*, such as IBCC [11] and CBCC [18], have been shown to produce more accurate aggregations when faced with spammers compared to single weight methods [9]. A confusion matrix is a square stochastic matrix where each row represents the worker’s accuracy of a given category. It is this breakdown of the workers’ accuracy per category that overcomes the limitation of using single weights. However, to date, these models only address settings where documents have a single category chosen among multiple alternatives. When faced with documents with multiple categories, such models infer inaccurate aggregation and workers’ ability.

To address this shortcomings, we extend IBCC to deal with settings where documents have multiple categories. Specifically, we seek to aggregate multiple workers’ judgments of the proportion of those categories. We elicit these judgements in the form of categorical distributions from each worker, and use these distributions as input to our model. We then sample and weight (through the workers’ confusion matrices) each distribution repeatedly to obtain multiple discrete observations for a single document by the same worker. This sampling step is crucial to enable the use of confusion matrices to identify spammers, while avoiding the constraint of classifying documents into a single category as found in IBCC. A by-product of using confusion matrices is that it enables the classification of spammers from diligent workers. This is key to improving the aggregation accuracy as spammers can be removed from the original dataset; and can further be excluded from the pool of available workers in future crowdsourcing campaigns. Furthermore, our approach is sufficiently flexible to learn the workers’ ability in both unsupervised and semi-supervised settings. In particular, the unsupervised approach learns the workers’ ability and the documents’ proportion simultaneously, making it especially suitable when the ground truth is not available. On the other hand, if the true proportion is known for some of the documents, performance can be improved on new documents using this limited training data.

In more detail, we make the following contributions to the state of the art: (i) we define a probabilistic Bayesian model that jointly learns for the first time the per category accuracies of individual workers, together with the aggregated distribution of categories associated with each document, (ii) we empirically show that our model outperforms existing aggregation methods on three real-world datasets; achieving a comparable level of aggregation accuracy when 60% of the workers are spammers, as other approaches do when there are no spammers, (iii) we show a five times improvement in the expected number of misclassified spammers compared to existing aggregation methods.

The remainder of this paper is organised as follows. We first introduce our notation and present IBCC in Section 2. In Section 3, we detail our method. In Section 4, we present the results of our experimental evaluation. We conclude in Section 5.

2 Preliminaries

We denote by $x \perp y$ the fact that the random variable x is independent of the random variable y ; $a \sim$ means “distributed as”; and $|$ expresses conditional probabilities. If x is has categorical distribution we write $x \sim \text{Cat}(\cdot)$. The Kronecker function $\delta(x)$ is 1 if $x = 0$; 0 otherwise.

Given this notation, IBCC assumes that each document has a single unknown category which we want to infer (i.e. the target category). The target category t_i for document i takes a value in $j \in \{1, \dots, J\}$, where J is the number of alternatives. Categories are assumed to be drawn from a categorical distribution with probability

$$t_i \sim \text{Cat}(\boldsymbol{\kappa}). \quad (1)$$

Given a set of K workers, each worker $k \in \{1, \dots, K\}$ submits a judgment $c_i^{(k)} = l$ of the target category $t_i = j$ for document i , where $l \in \{1, \dots, J\}$ is the set of discrete judgments that the worker can make. A judgment $c_i^{(k)}$ from worker k is assumed to have been generated from a categorical distribution

$$c_i^{(k)} \sim \text{Cat}\left(\boldsymbol{\pi}_{t_i}^{(k)}\right). \quad (2)$$

Furthermore, the workers’ judgments are assumed to be conditionally independent given the target category t_i

$$c_i^{(k)} \perp c_i^{\{1, \dots, K\} \setminus k} | t_i, \quad \forall k \in \{1, \dots, K\}.$$

This is the assumption commonly used in *naïve Bayes classifiers* which ignore correlations between workers. It is a reasonable assumption in crowdsourcing since workers do not typically interact with each other. The probabilities $\pi_{j,l}^{(k)}$ for $j \in \{1, \dots, J\}$ and $l \in \{1, \dots, J\}$ are the individual error-rates of the k -th worker. The confusion matrix $\boldsymbol{\Pi}^{(k)} = \{\boldsymbol{\pi}_j^{(k)} : j = 1, \dots, J\}$ for worker k is a square stochastic matrix defined on $\mathbb{R}^{J \times J}$ capturing the probabilistic dependency between the workers’ judgments and the consensus. Each row represents a possible category $j \in \{1, \dots, J\}$, while each column represents the worker’s judgment $l \in \{1, \dots, J\}$ regarding each category. All rows of the confusion matrix are assumed independent within and across workers

$$\boldsymbol{\pi}_i^{(k)} \perp \boldsymbol{\pi}_j^{\{1, \dots, K\} \setminus k}, \quad \forall k \in \{1, \dots, K\} \text{ and } \forall i \neq j.$$

This means that a worker’s ability to identify a given category is not dependent on their ability to identify the other alternatives. In line with Bayesian inference, the parameters $\boldsymbol{\kappa}$ and $\boldsymbol{\Pi}$ are considered random variables. Therefore, a conjugate Dirichlet prior distribution is assigned to the parameter vector $\boldsymbol{\kappa}$, such that

$$\boldsymbol{\kappa} \sim \text{Dir}(\boldsymbol{\nu}) \quad (3)$$

where $\boldsymbol{\nu}$ is a parameter vector of dimensions J – one for each of the possible categories. Intuitively, we can view the hyperparameter $\boldsymbol{\nu}$ as *pseudo-counts* of prior observations; that is, the number of documents in each category that all the workers have already judged. A conjugate Dirichlet prior distribution is similarly introduced over the parameter $\boldsymbol{\pi}_j^{(k)}$ with hyperparameter $\boldsymbol{\alpha}_j^{(k)}$ such that

$$\boldsymbol{\pi}_j^{(k)} \sim \text{Dir}\left(\boldsymbol{\alpha}_j^{(k)}\right). \quad (4)$$

The set of hyperparameters $\alpha_j^{(k)}$ form a matrix $\mathbf{A}^{(k)} = \{\alpha_1^{(k)}, \dots, \alpha_J^{(k)}\}$, where each row j is a point value vector $\alpha_j^{(k)}$. The matrix $\mathbf{A}^{(k)}$ is chosen to represent any prior level of uncertainty in the workers' confusion matrix, and can also be regarded as pseudo-counts of prior observations; that is, the number of documents of each category that worker k has already judged.

Now, IBCC assumes that workers classify documents into a single category among multiple alternatives. This limitation provides the basis for our extension.

3 Aggregating Judgments over Multiple Categories

Our proposed model – referred to hereafter as *multi-category independent Bayesian classifier combination* (MBCC) – is a generalisation of IBCC that deals with settings where documents have multiple categories. In particular, MBCC regards the problem of judging proportions as one of probabilistic modelling, where the proportion of a category is the probability of that category in the document. Thus, we associate each document with a categorical distribution, where each dimension is a distinct category. Workers then provide judgments of the proportion in each document, again using categorical distributions. These judgments are then aggregated by sampling each distribution, while taking into account the workers' ability through their confusion matrix. This sampling step is crucial to leverage confusion matrices to identify spammers, leading to more accurate aggregation. In addition, MBCC can be used in both unsupervised and semi-supervised settings where known proportions for limited number of documents can be used to improve the accuracy of the inferred workers' ability.

In more detail, we introduce a categorical distribution (i.e. the category proportion) with parameter Λ_i over the J categories for each document i (Equation 5) instead of the single categorical distribution κ common to all documents found in IBCC (Equation 1). Moreover, given a set of K workers, each worker $k \in \{1, \dots, K\}$ submits a distribution $\Phi_i^{(k)}$ over the J categories for each document i instead of the single category $c_i^{(k)}$ required by IBCC. The confusion matrix $\Pi^{(k)}$ of each worker k is kept unchanged from IBCC. Therefore in this new setting, to assess the accuracy of each worker through their confusion matrices $\Pi^{(k)}$, we need to first independently sample the aggregated distribution Λ_i to obtain a set of categories $z_{i,n}$ for each document i such that

$$z_{i,n} \sim \text{Cat}(\Lambda_i), \quad (5)$$

for all samples $n \in \{1, \dots, N\}$. The vector \mathbf{z}_i of dimension N can be seen as independent target categories t_i of document i (as in IBCC) drawn from the categories proportion in Equation 5. We subsequently match these samples against samples from each of the workers' distribution $\Phi_i^{(k)}$ to obtain

$$c_{i,n}^{(k)} \sim \text{Cat}(\pi_{z_{i,n}}^{(k)}) \quad (6)$$

Procedure 1 Generative process of MBCC.

```

1: Input: the confusion matrices  $\Pi$  and the category proportions  $\Lambda$ 
2:
3: for each document  $i \in \{1, \dots, I\}$  do
4:   for each sample  $n \in \{1, \dots, N\}$  do
5:     Sample  $z_{i,n} \sim \text{Cat}(\Lambda_i)$ 
6:     for each worker  $k \in \{1, \dots, K\}$  do
7:       Sample  $c_{i,n}^{(k)} \sim \text{Cat}(\pi_{z_{i,n}}^{(k)})$ 
8:     end for
9:   end for
10:
11:   for each worker  $k \in \{1, \dots, K\}$  do
12:     for each category  $j \in \{1, \dots, J\}$  do
13:        $\Phi_{i,j}^{(k)} = \beta_{i,j}^{(k)} \sum_{n=1}^N \delta(c_{i,n}^{(k)} - j)$ 
14:       where  $\beta_{i,j}^{(k)}$  is a normalising constant.
15:     end for
16:   end for
17: end for
18:
19: return  $\Phi$ 

```

for each document i . This is equivalent to running IBCC N times with different values of the workers' judgment $c_i^{(k)}$ drawn from their distribution $\Phi_i^{(k)}$. We perform this sampling until the accuracy no longer increases.

To do so, we rename for clarity the hyperparameter ν in IBCC (i.e. the pseudo-count of the categories across the corpus) to ϵ_i (i.e. the pseudo-count of the categories for each document i), such that

$$\Lambda_i \sim \text{Dir}(\epsilon_i). \quad (7)$$

Thus, the joint distribution over all variables is given by

$$p(\mathbf{z}, \mathbf{c}, \Lambda, \Pi) = \prod_{i=1}^I \text{Cat}(\Lambda_i) \text{Dir}(\epsilon_i) \prod_{n=1}^N \text{Cat}(\pi_{z_{i,n}}^{(k)}) \prod_{k=1}^K \prod_{j=1}^J \text{Dir}(\alpha_j^{(k)}). \quad (8)$$

The generative process, that is, the random process by which MBCC assumes the workers' judgment $\Phi_i^{(k)}$ arose, is summarised in Procedure 1. In particular, we start with the confusion matrices Π , and the category proportions Λ sampled from Equation 4 and Equation 7 respectively (Line 1). We then sample N categories $z_{i,n}$ from each category proportion Λ_i from Equation 5 (Line 5). Given each category $z_{i,n}$, we sample judgments $c_{i,n}^{(k)}$ from the workers' $z_{i,n}$ -th row of their confusion matrix $\Pi^{(k)}$ (Line 7). Finally, we find the most likely categorical distributions $\Phi_i^{(k)}$ which generated the samples $\mathbf{z}_i^{(k)}$ for all documents i and workers k (Line 13).

The key inferential problem that needs to be solved in order to use MBCC is that of computing the posterior distribution of the latent variables \mathbf{z} , and the parameters Λ and Π given the data \mathbf{c} . That is

$$p(\mathbf{z}, \Lambda, \Pi | \mathbf{c}) = \frac{p(\mathbf{z}, \mathbf{c}, \Lambda, \Pi)}{p(\mathbf{c})}, \quad (9)$$

where the denominator (i.e. the model evidence) is a constant. To compute the conditional distribution in Equation 9, we can use approximation methods based on statistical sampling (such as *Markov chain Monte Carlo* (MCMC) [6]) or density approximation (such as varia-

tional approximation [10, 1] or Laplace approximation [12]). In particular, our implementation uses the variational message passing algorithm [21] as it has proven itself to be superior in terms of speed and accuracy compared to both sampling and density based approximation [13], but other approaches could be used if appropriate.

4 Empirical Evaluation

To evaluate the efficacy of our model, we use an independently gathered dataset, and introduce two new real-world datasets; all of which include ground truth from expert annotators. We then compare performance against four state-of-the-art benchmarks. The experiments are run in an unsupervised setting, where the ground truth is never exposed to the algorithms, and is only used to measure the accuracy of each model.

4.1 Datasets

The three datasets used consist of: (i) a sentiment proportion annotation in news headlines which requires background knowledge from workers, (ii) an object proportion annotation in images which presents workers with complete information, and (iii) a colour proportion annotation in countries’ flag where no information is presented to the workers other than the countries’ name.

SemEval. This publicly available dataset¹ is provided by Snow et al. [15]. They hired a number of workers from AMT to categorise a set of one hundred news headlines sampled from the SemEval2007 test set [16]. Each worker was presented with a list of headlines, and was asked to give numeric judgments between zero and a hundred for each of six sentiments. Ten judgments were collected for each headline for a total of 1,000 judgments. These judgments were obtained from 38 workers whom provided a minimum of 20 judgments each, and 26 on average. We truncate the total number of judgments per worker to 20 to avoid discrepancies in accuracy of each inferred confusion matrix. We normalise the values submitted by each worker into valid probability distributions by ensuring that the total area is equal to one at any given time. For example, if a worker provided the following judgment $[0, 100, 0, 50, 50, 100]$, his associated normalised distribution will be $[0, \frac{1}{3}, 0, \frac{5}{30}, \frac{5}{30}, \frac{1}{3}]$.

IAPR-TC12. We crowdsourced a set of 16 images sampled from the publicly available² IAPR-TC12 dataset. This is a collection of 20,000 images of urban and rural scenes manually segmented per regions. Each pixel belongs to one of six region. We gathered a total of 21 judgments per image from 21 workers. Workers were asked to estimate the proportion of each region in the image. The ground truth proportion for each category is calculated by dividing the number of pixels in the region by the total number of pixels in the image. The workers reported their judgments with a pie chart, enabling quick and accurate judgments of proportion [8].

Colours. We crowdsourced a set of 460 judgments of the proportion of colours in the flags of 20 countries.

We asked 23 participants to judge, from memory, the proportion of 10 colours in each country’s flag.

Although these three datasets may already contain spammers, we augment the datasets with additional synthetic spammers to explore the loss of accuracy as they increase in number. There are a number of strategies available for modelling spammers. Vuurens et al. [19] identified four types of spammers: sloppy, uniform, random, and semi-random. While sloppy workers try to the best of their abilities to complete the tasks, they may be insufficiently precise in their judgments. On the other hand, uniform spammers use a fixed uniform judgment pattern across all documents. Finally, random spammers provide unique meaningless answers for each document and semi-random spammers also answer a few questions properly. While the obvious characteristic of repeating judgment patterns can be used to manually filter out uniform spammers, random spammers are the most challenging to detect. For this reason, we use a random spamming strategy where a spammer always provides a random categorical distribution for each document. The distribution of each spammer k for each document i shares a prior Dirichlet distribution such that

$$\Phi_i^{(k)} \sim \text{Dir}(\mathbf{1}), \quad (10)$$

where the pseudo-count of $\mathbf{1}$ ensures that all possible distributions $\Phi_i^{(k)}$ are equally likely.

4.2 Experimental Setting

We set the parameter of the prior probability of each confusion matrix for all workers and spammers to $\mathbf{A}^{(k)} = 100 \times \mathbf{I} + \mathbf{1}^T \mathbf{1}$. This means that workers are initially assumed to be reasonably accurate before seeing any data. Using a different assumption leads to distinct aggregation accuracy profiles that will be discussed in more detail in Section 4.5. Furthermore, to ensure fair comparison between the benchmarks, we do not adjust each parameters ϵ of the prior distributions over categories to reflect the balance of each dataset. Finally, we run all models a hundred times each to achieve statistically significant results at the 99% confidence level.

4.3 Benchmarks

We compare the performance of our model to four state of the art benchmarks methods.

Uniform assigns a uniform distribution for each aggregated distribution (i.e. $p_j = \frac{1}{J}$ for all $j \in \{1, \dots, J\}$), making it independent of the dataset. These particular values of p_j maximise the entropy function of categorical distributions. That is, if one were to guess the aggregated distribution far from the ground truth on average (according to some distance metric), it can be expected to have its error above the uniform model.

LinOp simply averages the distributions provided by the workers. Unlike IBCC and MBCC, it does not sample the workers’ distributions to produce the discrete observations, but directly takes the distributions as input. As there is no training set for assigning informative weights to the workers, we assign equal weights $\omega^{(k)} = \frac{1}{K}$ to each worker.

¹<https://sites.google.com/site/nlpannotations/>

²<http://imageclef.org/SIAPRdata>

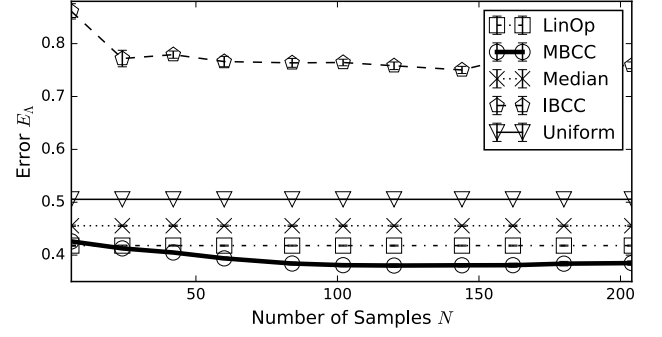
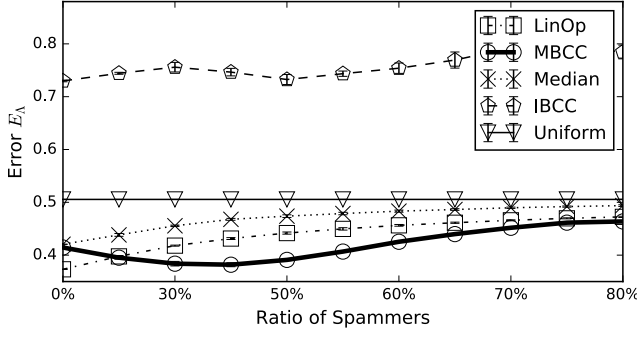


Figure 1: Average error on the aggregated distributions Λ on the SemEval dataset when increasing: (left) the ratio of spammers at $N = 180$ samples, (right) the number of samples at a ratio of spammers of 50%.

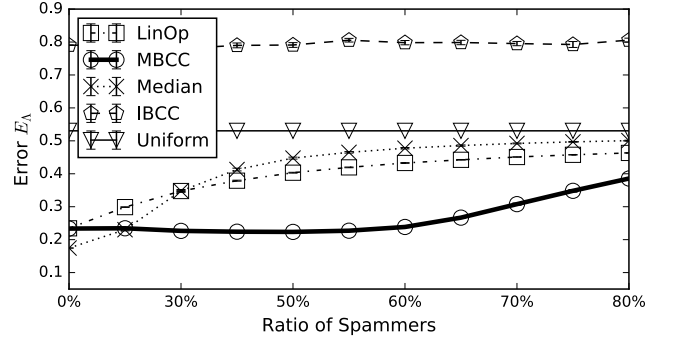
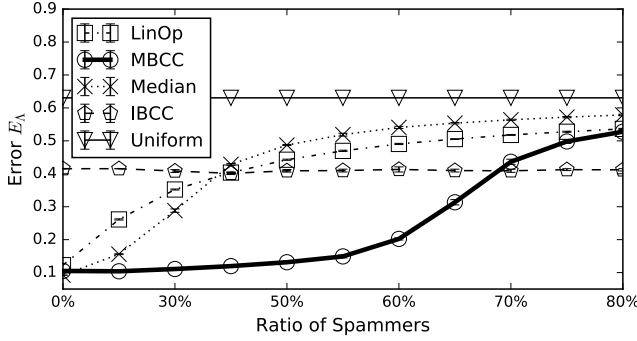


Figure 2: Average error on the aggregated distributions when increasing the ratio of spammers on: (left) the IAPR-TC12 dataset at $N = 180$ samples, (right) the Colours dataset at $N = 330$ samples.

Median estimates the aggregated distribution by arranging the judgments for each document in ascending order and then takes the middle judgment. Each judgment is considered with equal weight for the same reason as for LinOp. As such, this method assumes that all workers have equal abilities. The median is a robust method against extreme judgements since it will not give an arbitrarily large or small result if no more than half of the judgments for a document are incorrect.

IBCC combines discrete judgments from multiple workers and models the ability of each individual worker using confusion matrices [11]. Although IBCC takes a single category as ground truth, the output is a posterior distribution over the categories which can be compared to the ground truth category proportion.

4.4 Accuracy Metrics

To assess the accuracy of the inference, we use the Euclidean distance³. Specifically, we define the average error of the aggregation on the entire dataset by

$$E_{\Lambda} = \frac{1}{I} \sum_{i=1}^I d(\Lambda_i^*, \Lambda_i) \quad (11)$$

³The KL divergence cannot be used since it is not defined for documents with categories having zero probabilities.

where $d(\cdot)$ is the Euclidean distance, I the total number of documents, Λ_i^* the ground truth distribution for document i , and Λ_i the aggregated distribution provided by the model for document i . Furthermore, we define the deviation of the confusion matrix $\Pi^{(k)}$ of worker k to the identity matrix \mathbf{I} by

$$E_{\Pi^{(k)}} = \sum_{j=1}^J d(\mathbf{I}_j, \pi_j^{(k)}). \quad (12)$$

4.5 Results

We now present the results of the empirical evaluation regarding a number of key aspects: (i) accuracy of aggregation and robustness to spammers, (ii) convergence and running time, (iii) classification of spammers. We first consider in detail the results on the SemEval dataset, and briefly discuss the results on the other datasets.

Figure 1 (left) shows the average error as the ratio of spammers increases on the SemEval dataset. To illustrate, a ratio of spammers of 50% means half the workers in the dataset are spammers. As can be seen, MBCC achieves a comparable aggregation accuracy when 50% of the workers are spammers, as LinOp does when no spammers are added. However, the added complexity of detecting spammers in MBCC comes at a cost when the level of spammers is low. Indeed, the highest precision is

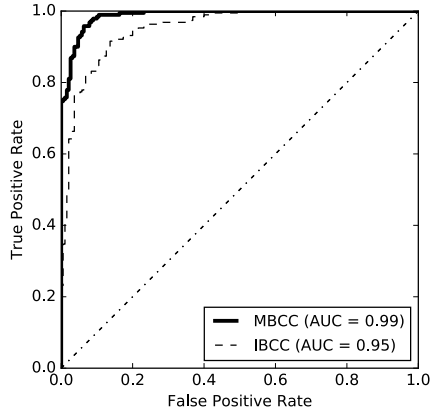


Figure 3: ROC curves on the SemEval dataset. The ratio of spammers is set to 50%.

initially produced by LinOp, but as we increase the ratio of spammers we notice a crossover point after which MBCC is the most precise. The value of this crossover point depends on our assumption regarding the workers (i.e. the choice of prior assigned to each worker), and can be adjusted using the pseudo-counts of prior observations $\mathbf{A}_{ii}^{(k)}$ on the diagonal of the workers' confusion matrix. For high values of pseudo-counts, MBCC assumes all workers are truthful. This is the assumption underpinning LinOp, where all workers have an implicit identity matrix as their confusion matrix. In fact, for high pseudo-counts, the performance of MBCC matches perfectly with LinOp regardless of the number of spammers. On the other hand, with lower values of pseudo-counts, MBCC allows greater flexibility to learn the spammers, at the cost of having a greater error when there are fewer spammers in the dataset. Therefore, the added degrees of freedom of MBCC lead to a tradeoff in accuracy at different ratio of spammers. Furthermore, since sampling both the documents' priors and the workers' distributions is a necessary feature of our approach to enable the use of confusion matrices, it raises the question regarding the number of samples required to accurately infer the aggregated distribution of each document. Figure 1 (right) shows the average error of the aggregated distribution when increasing the number of samples at a ratio of spammers of 50%. As the number of samples increases, we observe convergence of the error at 100 samples to values of 0.75 and 0.37 for IBCC and MBCC respectively. That is, the errors no longer decrease beyond this threshold. Inevitably, the running time also increases as the number of samples increases. In particular, running time of LinOp and Median is typically 3ms, while that of IBCC and MBCC ranges from 12s and 13s, to 6min and 28min respectively, across the range of samples shown in Figure 1 (right).

We now evaluate the accuracy of the confusion matrix-based models at classifying workers from spammers on the SemEval dataset. To do this, we first collect the inferred confusion matrix $\mathbf{\Pi}^{(k)}$ for each worker. We then

compute the deviation $E_{\mathbf{\Pi}^{(k)}}$ of each confusion matrix to the identity matrix using Equation 12. The identity matrix represents the confusion matrix of a perfect worker, that is, one that always gives judgments in accordance with the consensus. We then set a threshold on the error, above which a worker is classified as a spammer. The receiver operating characteristic (ROC) curves in Figure 3 captures the effect of varying the threshold of the error. As can be seen, MBCC has an area under the curve (AUC) of 0.99 showing a 5 times improvement compared to IBCC in terms of the expected number of misclassified spammers. The AUC eventually decreases for both models at higher ratios.

We now briefly discuss the results on the two other datasets. On the IAPR-TC12 dataset (Figure 2 (left)), LinOp, Median and MBCC achieve equal accuracies when no spammers are added. However, since each image in this dataset is dominated by a single category, that is, at least one category has more than 50% coverage in each image, IBCC performs reasonably well even with a high ratio of spammers. This is because IBCC always assigns a probability of one to the most likely category, which emphasises its assumption that documents have exactly one category. In fact, the more the documents' proportion regress to Kronecker functions, the lower the error from IBCC. On the Colours dataset (Figure 2 (right)), the accuracy of MBCC remains a lower bound of LinOp throughout the range of ratios of spammers. However, the Median initially achieves the lowest error with a crossover point with MBCC at 15% spammers. This is because the judgments provided by the workers do include sufficient outliers, making the Median a good measure of central tendency. Finally, convergence of the error on the aggregation is reached at 100 samples for the IAPR-TC12 dataset and 150 for the Colours dataset. Since the Colour dataset has 4 more categories than the IAPR-TC12 datasets, it requires more samples to achieve convergence.

5 Conclusions

This paper introduced a novel model for the aggregation of categorical distributions. The key innovation of our method is the elicitation and sampling of judgments of proportions in the form of probability distributions from the workers, and the use of these samples to improve the accuracy of aggregation. In particular, we showed empirically, on three real-world datasets, that our approach outperforms existing aggregation methods by up to 28% in terms of aggregation accuracy. We have also shown a comparable level of aggregation accuracy when 60% of the workers are spammers, as other approaches do when there are no spammers. Finally, we achieved up to a five time improvement in the expected number of misclassified spammers compared to existing aggregation methods. As future work, we plan to investigate how MBCC can be adapted as an efficient unsupervised multi-category classifier where the workers would provide discrete choices among the alternatives.

References

- [1] H. Attias. A Variational Bayesian Framework for Graphical Models. In *Advances in Neural Information Processing Systems 12*, pages 209–215. MIT Press, 2000.
- [2] M. Bacharach. Normal Bayesian Dialogues. *Journal of the American Statistical Association*, 74(368):837, 1979.
- [3] G.Y. Bae, M. Olkkonen, S.R. Allred, and J.I. Flombaum. Why Some Colors Appear More Memorable than Others: A Model Combining Categories and Particulars in Color Working Memory. *Journal of Experimental Psychology: General*, 144(4):744–763, 2015.
- [4] D.M. Blei, A.Y. Ng, and M.I. Jordan. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.
- [5] D.E. Difallah, G. Demartini, and P. Cudr -Mauroux. Mechanical Cheat: Spamming Schemes and Adversarial Techniques on Crowdsourcing Platforms. In *CrowdSearch*, pages 26–30. Citeseer, 2012.
- [6] S. Geman and D. Geman. Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, (6):721–741, 1984.
- [7] C. J. Ho, R. Frongillo, and Y. Chen. Eliciting Categorical Data for Optimal Aggregation. In *Advances In Neural Information Processing Systems*, pages 2442–2450, 2016.
- [8] J.G. Hollands and I. Spence. Judgments of change and proportion in graphical perception. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 34(3):313–334, 1992.
- [9] P.G. Ipeirotis, F. Provost, and J. Wang. Quality Management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD workshop on human computation*, pages 64–67. ACM, 2010.
- [10] M.I. Jordan, Z. Ghahramani, T.S. Jaakkola, and L.K. Saul. An Introduction to Variational Methods for Graphical Models. *Machine learning*, 37(2):183–233, 1999.
- [11] H.C. Kim and Z. Ghahramani. Bayesian Classifier Combination. In *Proceedings of the 15th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2012.
- [12] P.S. Laplace. Memoir on the Probability of the Causes of Events. *Statistical Science*, 1(3):364–378, 1986.
- [13] T.P. Minka. Expectation Propagation for Approximate Bayesian Inference. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, volume abs/1301.2294, 2013.
- [14] J. Oakley. Eliciting Univariate Probability Distributions. *Rethinking risk measurement and reporting*, 1, 2010.
- [15] R. Snow, B. O’Connor, D. Jurafsky, and A.Y. Ng. 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*, pages 254–263. Association for Computational Linguistics, 2008.
- [16] C. Strapparava and R. Mihalcea. Semeval-2007 Task 14: Affective Text. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 70–74. Association for Computational Linguistics, 2007.
- [17] C.A. Varey, B.A. Mellers, and M.H. Birnbaum. Judgments of Proportions. *Journal of Experimental Psychology: Human Perception and Performance*, 16(3):613, 1990.
- [18] M. Venanzi, J. Guiver, G. Kazai, P. Kohli, and M. Shokouhi. Community-based Bayesian Aggregation Models for Crowdsourcing. In *Proceedings of the 23rd International Conference on World Wide Web*, pages 155–164. ACM Press, 2014.
- [19] J. Vuurens, A.P. de Vries, and C. Eickhoff. How Much Spam Can You Take? An Analysis of Crowdsourcing Results to Increase Accuracy. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR’11)*, pages 21–26, 2011.
- [20] S. Weerahandi and J.V. Zidek. Multi-Bayesian Statistical Decision Theory. *Journal of the Royal Statistical Society. Series A (General)*, 144(1):85, 1981.
- [21] J. Winn and C.M. Bishop. Variational Message Passing. *Journal of Machine Learning Research*, 6:661–694, 2005.
- [22] C. Witzel, C. Racey, and J.K. O’Regan. The Most Reasonable Explanation of "the dress": Implicit Assumptions About Illumination. *Journal of Vision*, 17(2):1, 2017.