

Computing Crowd Consensus with Partial Agreement

(Extended Abstract)

Nguyen Quoc Viet Hung⁺, Huynh Huu Viet^{*}, Nguyen Thanh Tam[#], Matthias Weidlich[†], Hongzhi Yin^{*}, Xiaofang Zhou^{*}
⁺Griffith University ^{*}The University of Queensland [#]École Polytechnique Fédérale de Lausanne [†]Humboldt-Universität zu Berlin

Abstract—Crowdsourcing has been widely established as a means to enable human computation at large-scale, in particular for tasks that require manual labelling of large sets of data items. Answers obtained from heterogeneous crowd workers are aggregated to obtain a robust result. However, existing methods for answer aggregation are designed for *discrete* tasks, where answers are given as a single label per item. In this paper, we consider *partial-agreement* tasks that are common in many applications such as image tagging and document annotation, where items are assigned sets of labels. Going beyond the state-of-the-art, we propose a novel Bayesian nonparametric model to aggregate the partial-agreement answers in a generic way. This model enables us to compute the consensus of partially-sound and partially-complete worker answers, while taking into account mutual relations in labels and different answer sets. An evaluation of our method using real-world datasets reveals that it consistently outperforms the state-of-the-art in terms of precision, recall, and scalability.

I. INTRODUCTION

Fuelled by the massive availability of Internet users, crowdsourcing has been established as a means for human computation at large-scale [1], [2], [3], [4], [5]. Most crowdsourcing setups are based on questions (aka tasks) that, once posted to a crowdsourcing platform, are answered by users (aka crowd workers) for financial rewards. Aggregation of answers shall complement individual errors, thereby exploiting the ‘wisdom of the crowd’.

In this paper, we focus on a special type of *partial-agreement* tasks, where workers shall provide a set of labels per item. Such tasks received much attention recently in many crowdsourcing applications, such as text categorization, image classification, and medical diagnosis [6].

In addition to the general challenges of answer aggregation, computing crowd consensus with partial-agreement is inherently complex. The labels obtained as part of different answers are often correlated. Identifying the correct set of labels needs to deal with the exponential growth of combinations of labels and dependencies between them. Also, workers no longer either agree or disagree on an answer to a question. Rather, consensus among workers becomes partial.

In this paper, we propose a Bayesian nonparametric model in order to capture the distinct properties of partial-agreement answer aggregation. That is, co-occurrence dependencies between labels are represented by the notion of latent label clusters. Furthermore, partial consensus between workers is modelled

by grouping together workers with similar answers. The resulting model, called *Generic Crowdsourcing Consensus with Partial Agreement (CPA)*, generalises the multi-label setting of answer aggregation and enables incremental learning using the principles of stochastic variational inference. A complete formalisation and evaluation of CPA has been published as [7].

II. PROBLEM AND APPROACH

We consider an illustrative image tagging task, in which workers assign one or more labels to a picture. Table I illustrates an exemplary crowdsourcing result, in which five workers ($u_1 - u_5$) provided their answers to four pictures ($i_1 - i_4$). The correct, yet generally unknown, label assignment is shown in a separate column. A common method to derive an aggregated answer is majority voting [8], which considers all labels separately. Compared to the actually correct assignment, the result obtained in this case has two issues, though: (i) it is partially incorrect (label 4 is not correct for i_1), and (ii) partially incomplete (labels 1 and 3 shall also be assigned to i_4).

TABLE I: Answers provided by five workers for four pictures.

	u_1	u_2	u_3	u_4	u_5	Correct	Majority [8]
i_1	{4,5}	{4,5}	{4}	{1}	{5}	{5}	{4,5}
i_2	{2,3}	{1,4}	{4}	{2}	{3,4}	{3,4}	{4}
i_3	{1,2}	{4}	{4}	{3}	{4,5}	{4,5}	{4}
i_4	{1,2}	{2,3}	{4}	{4}	{1,2,3}	{1,2,3}	{2}

1: sky, 2: plane, 3: sun, 4: water, 5: tree

We capture the setting of partial-agreement answer aggregation by a set of workers \mathcal{U} , identified by their indices, $\mathcal{U} \triangleq \{1, \dots, U\}$ that provide answers for a set of items \mathcal{N} , also identified by their indices, $\mathcal{N} \triangleq \{1, \dots, I\}$. $\mathcal{Z} \triangleq \{1, \dots, C\}$ is the set of all possible labels for these items. Each answer by a crowd worker is a subset of \mathcal{Z} . Formally, answers are modelled as an $I \times U$ answer matrix $\mathcal{M} = [x_{iu}]$, where $x_{iu} \subseteq \mathcal{Z}$ is the set of labels assigned to item i by worker u , or $x_{iu} = \emptyset$ if worker u has not provided an answer for item i .

The problem of partial-agreement answer aggregation is the construction of a *deterministic assignment* $d: I \rightarrow 2^{\mathcal{Z}}$ assigning a set of labels to each item.

We approach this problem by considering the true labels of items as unobserved random variables and predict their values using a generative process based on a Bayesian network [9], [10], [11].

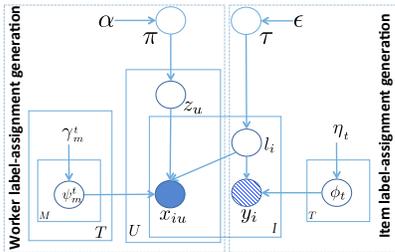


Fig. 1: Graphical representation of the CPA model.

As illustrated in Fig. 1, the model of Generic Crowdsourcing Consensus with Partial Agreement (CPA) is constructed from worker communities, item clusters, and label selection:

Worker Communities. There is a finite set of worker communities π , identified by indices, $\pi \triangleq \{1, \dots, M\}$. $z_u \in \pi$ denotes the community of worker u . We generate π nonparametrically using a Chinese Restaurant Process (CRP)[12].

Item Clusters. There is a finite set τ of clusters, identified by indices, $\tau \triangleq \{1, \dots, T\}$. Here, $l_i \in \tau$ denotes the cluster of item i . Again, τ is generated nonparametrically by a CRP.

Label Selection. Each worker is characterised by a $C \times T$ confusion matrix ψ_m , where m is the community of the worker. We denote by ψ_m^t a column vector of C -dimensions, which contains the probabilities that a worker in community m assigns the respective labels given an item of cluster t . This model has the advantage that, instead of considering exponentially many subsets of labels, it relies on the number of all possible item clusters, which is tractable in practice.

Inferring the parameters of the CPA model is, in fact, the estimation of values of the above priors ($\alpha, \epsilon, \gamma, \eta$). This is equivalent to inferring the posterior distribution of the unobserved variables ($\pi, \tau, z, l, \psi, \phi$) under the observed variables (x, y), which is $p(\pi, \tau, z, l, \psi, \phi | x, y)$.

Instead of computing the intractable posterior distribution directly, we use variational inference and infer an approximation $q(\pi, \tau, z, l, \psi, \phi)$, referred to as variational distribution. To approximate the posterior distributions p by variational distributions q , we minimize the KL -divergence between them, $KL(q | p)$. With $\Theta \triangleq \{\pi, \tau, z, l, \psi, \phi\}$, it is defined as:

$$KL(q | p) \triangleq - \int q(\Theta) \ln \frac{p(\Theta, x, y)}{q(\Theta)} d\Theta + \ln p(x, y) \\ \triangleq -\mathcal{L}(\Theta) + \text{const}$$

$\mathcal{L}(\Theta)$ is called *evidence lower bound* (ELBO) and denotes the variational objective function. Using variational theory [13], taking derivatives of this lower bound with respect to each variational parameter, we can derive the coordinate ascent updates [13] to compute the model parameters until convergence.

The above deterministic variational inference maximises the EBLO function $\mathcal{L}(\Theta)$ using coordinate-ascent for each of the parameters of variational distributions. To realise incremental learning for settings with streaming data, we rely on stochastic variational inference [14] and apply stochastic optimization to the EBLO function based on newly received data. Using stochastic gradient descent, only a small subset of all available data is needed to update the parameters in each iteration.

III. EXCERPT OF EXPERIMENTAL RESULTS

We conducted experiments with real-world datasets from diverse application scenarios: *image annotation*, *topic annotation*, *aspect extraction*, and *entity extraction*. We employed workers to perform item labelling using the CrowdFlower. In total, we spent a budget of 8772 tasks for all datasets and ended up having a repository of 87720 label annotations for 10610 items from 2664 users.

Table II shows an excerpt of our experimental results, illustrating the precision and recall obtained by our CPA model against the three baseline methods: majority vote (MV), expectation maximisation (EM), and Community-based Bayesian Classifier Combination (cBCC). Due to incorporating dependencies between labels, our CPA model, however, significantly outperforms all baseline methods.

TABLE II: Overall accuracy

Dataset	Precision				Recall			
	MV	EM	cBCC	CPA	MV	EM	cBCC	CPA
image	0.65	0.66	0.7	0.81	0.57	0.62	0.63	0.74
topic	0.57	0.60	0.62	0.79	0.54	0.54	0.55	0.70
aspect	0.52	0.61	0.65	0.74	0.53	0.56	0.6	0.64
entity	0.63	0.57	0.60	0.79	0.55	0.50	0.53	0.70

Fig. 2 further illustrates that our approach scales well. Here, static reasoning (*offline*) is compared against incremental inference (*online*), potentially with parallelization using 4 or 16 cores. Furthermore, our methods outperform other baselines except majority voting, illustrating that the above gains in accuracy can be realised efficiently.

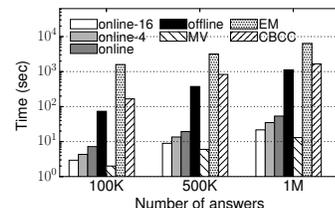


Fig. 2: Runtime of CPA inference and prediction mechanisms

IV. CONCLUSION

In sum, we presented a novel Bayesian nonparametric approach to aggregate partial-agreement crowdsourcing answers. It enables us to capture worker characteristics and dependencies between the labels assigned to items.

REFERENCES

- [1] A. J. Quinn and B. B. Bederson, “Human computation: a survey and taxonomy of a growing field,” in *CHI*, 2011, pp. 1403–1412.
- [2] N. Q. V. Hung, N. T. Tam, L. N. Tran, and K. Aberer, “An evaluation of aggregation techniques in crowdsourcing,” in *WISE*, 2013, pp. 1–15.
- [3] Q. V. H. Nguyen, T. T. Nguyen, N. T. Lam, and K. Aberer, “Batc: a benchmark for aggregation techniques in crowdsourcing,” in *SIGIR*, 2013, pp. 1079–1080.
- [4] N. Q. V. Hung, D. C. Thang, N. T. Tam, M. Weidlich, K. Aberer, H. Yin, and X. Zhou, “Answer validation for generic crowdsourcing tasks with minimal efforts,” *Vldb J.*, pp. 855–880, 2017.
- [5] Q. V. H. Nguyen, C. T. Duong, T. T. Nguyen, M. Weidlich, K. Aberer, H. Yin, and X. Zhou, “Argument discovery via crowdsourcing,” *VldbJ*, pp. 1–25, 2017.

- [6] I. Ounis, C. Macdonald, J. Lin, and I. Soboroff, "Overview of the trec-2011 microblog track," in *TREC*, 2011.
- [7] N. Q. V. Hung, H. H. Viet, N. T. Tam, M. Weidlich, H. Yin, and X. Zhou, "Computing crowd consensus with partial agreement," *TKDE*, 2017.
- [8] J. Deng, O. Russakovsky, J. Krause, M. S. Bernstein, A. Berg, and L. Fei-Fei, "Scalable multi-label annotation," in *CHI*, 2014, pp. 3099–3102.
- [9] H. Yin, L. Chen, W. Wang, X. Du, N. Q. V. Hung, and X. Zhou, "Mobi-sage: A sparse additive generative model for mobile app recommendation," in *ICDE*, 2017, pp. 75–78.
- [10] H. Yin, X. Zhou, B. Cui, H. Wang, K. Zheng, and N. Q. V. Hung, "Adapting to user interest drift for POI recommendation," *TKDE*, pp. 2566–2581, 2016.
- [11] H. Yin, Z. Hu, X. Zhou, H. Wang, K. Zheng, N. Q. V. Hung, and S. W. Sadiq, "Discovering interpretable geo-social communities for user behavior prediction," in *ICDE*, 2016, pp. 942–953.
- [12] P. G. Moreno, A. Artes-Rodriguez, Y. W. Teh, and F. Perez-Cruz, "Bayesian nonparametric crowdsourcing," *JMLR*, pp. 1607–1627, 2015.
- [13] D. Blei and M. Jordan, "Variational inference for dirichlet process mixtures," *Bayesian Anal.*, pp. 121–143, 2006.
- [14] M. D. Hoffman, D. M. Blei, C. Wang, and J. Paisley, "Stochastic variational inference," *JMLR*, pp. 1303–1347, 2013.