Crowdsourcing for structured labeling with applications to protein folding

Jian Peng JPENG@CSAIL.MIT.EDU CSAIL, MIT
Qiang Liu QLIU1@ICS.UCI.EDU ICS,UCI
Alexander Ihler IHLER@ICS.UCI.EDU ICS,UCI
Bonnie Berger BAB@CSAIL.MIT.EDU CSAIL,MIT

Abstract

Label aggregation is an important problem in crowdsourcing research. Besides simple classification and regression tasks, many crowdsourcing experiments require workers to return structured outputs as labels. These labels are often inherently constrained, making the aggregation task very difficult. We approach this problem by formulating it as an exemplar-based clustering problem and choosing the most representative cluster centers as the aggregated label. This approach naturally extends the majority voting method and incorporates workers' abilities. We evaluate our method on the most recent protein structure prediction experiment CASP10, where the structured outputs are 3D coordinates of proteins, and find that our method outperforms both majority voting and the single best predictor. These results are very promising since in CASP10 all earlier aggregation methods performed worse than the single best predictor.

1. Introduction

Crowdsourcing has become very popular as a novel human computational paradigm in numerous data processing or problem solving tasks, including image processing (e.g., Welinder et al., 2010), natural language processing (e.g., Snow et al., 2008) and bioinformatics (e.g., Cooper et al., 2010). Human intelligence is able to solve problems that are difficult for computers to solve. Systems, platforms and organized collaborations have been initiated to collect such human intelligence (e.g. labeled data or predictions), from either non-expert crowds or trained experts. However, the quality of data from crowdsourcing experiments is often noisy and unreliable. An important computational challenge is to combine the crowdsourced labels properly to get the most possible accuracy results.

Probably the most widely used method to aggregate labels is majority voting, which simply chooses the labels returned by the majority of workers. However, majority voting treats all the workers uniformly, making it error-prone when there exist adversaries or spammers in the crowd. A line of research has been developed to improve majority voting by building generative probabilistic models that take workers' abilities into account (e.g., Dawid & Skene, 1979; Welinder et al., 2010; Karger et al., 2011; Liu et al., 2012). Essentially, all these methods can be considered as weighted majority votes where different workers are weighted differently according to their abilities.

Most current label aggregation algorithms are designed for simple binary/multi-class labels, or well structured labels such as ranking (e.g., Lee et al., 2012) and sequential labels (e.g., Wu et al., 2012). However, many real-world crowdsourcing experiments generate labels with complicated structures where it is intractable to build generative models. For instance, in protein structure prediction experiments (Critical Assessment of Techniques for Protein Structure Prediction or CASP in short), the outputs or predictions gathered from each worker are collections of three-dimensional coordinates which are invariant subject to translations or rotations and must satisfy a set of complicated geometric constraints. Therefore,

ICML Workshop: Machine Learning Meets Crowdsourcing, Atlanta, Georgia, USA, 2013.

it does not make sense to simply average or vote the coordinates element-wisely, and it is difficult to adapt methods based on generative probabilistic modeling. It is an important challenge to build efficient consensus algorithms for handling these complicated labels.

One key insight is that the structures of the labels are well characterized by their similarity measurements or kernels, which are relatively easier to access. In this paper, we develop a general approach to aggregate complicated structured labels using only similarity matrixes, which avoids dealing with the complicated inner structures directly. Our method can be treated as a kernelized weighted majority voting, where the workers are weighted by how well they can represent the other workers. Finally, we demonstrate the efficiency of our approach on the protein structure prediction problem.

2. Basic Settings

We assume that there are K tasks and N workers in the crowdsourcing experiment. Each edge E_{ik} in a bipartite graph $G = (\{V_i, V_k\}, \{E_{ik}\})$ indicates that worker i makes a prediction for task k. For any task k, a similarity matrix S^k denotes the similarity between predictions made by all workers. In this matrix, each entry S^k_{ij} measures the similarity between the predictions made by workers i and j. We assume that many workers can generate reliable predictions. This assumptions often holds in practice, at least in the case of protein structure prediction, where most workers are experts in the field. Further, we also assume that good predictions tend to be similar to each other. This assumption is also usually valid, since the structure space is so large that any two random predicted structures are typically very dissimilar.

Similar to majority voting methods for classification or regression, a variant of this setting is to simply take the most representative member (or the medoid) of all predictions by considering their similarities. For task k, the medoid x^*_k of all predictions $\{x^k_i\}$ can be calculated by

$$x^*_k = x_{i^*_k}, \qquad i^*_k = \arg\max_i \sum_j S^k_{ij}, \tag{1}$$

where i^*_k is the index of the medoid. Surprisingly, this baseline method works very well in practice, especially when tested on the recent two CASP experiments. The assessors noticed that it outperformed most quality assessment programs which utilized substantial domain knowledge. On the other hand, all consensus methods (Kryshtafovych et al., 2011), including this baseline method, were still worse than the best worker in terms of the prediction accuracy. Furthermore, workers in the experiment often have diverse abilities on different types of proteins. All of these observations indicate the need to take workers' abilities into account.

3. A Convex Clustering Approach

To parameterize workers' abilities into the model, we assign each worker j a weight $w^k_j \in [0, 1]$ satisfying $\sum_j w^k_j = 1$, denoting the probability of his/her prediction as the medoid for task k. This parameter can also be explained as the quality of the worker's prediction. After these weights are obtained, the original majority voting can be naturally extended to a weighted version,

$$x^*_k = x_{i^*_k}, \qquad i^*_k = \arg\max_i \sum_j w^k_j S^k_{ij}. \tag{2}$$

We estimate the weights $\{w^k_j\}$ by solving the following optimization problem proposed by Lashkari & Golland (2010),

$$\max_{\{w^k_j\}} \sum_{k=1}^{K} \frac{1}{N} [\sum_{i=1}^{N} \log[\sum_{j=1}^{N} w^k_j S^k_{ij}]],$$

which can be interpreted as maximizing the log-likelihood of an exponential family mixture model for solving exemplar-based clustering problems (see Lashkari & Golland (2010) for details). This optimization problem is convex and can be optimized by a multiplicative update algorithm with provable convergence and globally optima guarantees.

We also want to make these w^k_j parameters consistent over all tasks while allowing a certain amount of diversity. We do this by adding a regularization term that penalizes the Kullbeck-Leibler divergence between the ability parameters w^k_j and their arithmetic

Algorithm 1 Iterative Optimization

Initialize vectors $w^k := \{w_j^k\}_j$ for all k.
Repeat until convergence:

$$w_j^0 \propto \frac{1}{K}\sum_k w_j^k, \qquad \text{for } \forall\, j$$

$$w_j^k \propto w_j^k \sum_i \frac{S_{ij}^k}{\sum_{j'} w_{j'}^k S_{ij}^k} + \alpha w_j^0, \quad \text{for } \forall\, k \text{ and } j$$

means. We then obtain the optimization problem,

$$\max_{\{w_j^k\}} \sum_{k=1}^{K} \frac{1}{N}[\sum_{i=1}^{N} \log[\sum_{j=1}^{N} w_j^k S_{ij}^k] - \alpha \sum_{k=1}^{K} \mathrm{KL}(\frac{1}{K}\sum_k w^k || w^k),$$

where $w^k = \{w_j^k\}_j$ and α is a hyper-parameter that controls the diversity of the ability parameters, which can be tuned on a validation set. The above optimization problem is also convex and can be solved using a closed form fixed point update shown in Algorithm 1.

4. Applications to Protein Structure Prediction

Critical Assessment of Techniques for Protein Structure Prediction experiments (CASP) have been evaluating the state-of-the-art computational methods for protein structure prediction since 1994. Participants are asked to make predictions for roughly one hundred proteins without publicly known structures. After the prediction season, solved structures for these proteins are made public to evaluate the performance of the state-of-the-art methods. The predictions made by participants typically vary a lot with accuracies over a potentially large range. As a result, reliable estimation of the quality of models is critical for biologists to determine the usefulness of the predictions. Towards this goal, a quality assessment category has been also included in CASP since 2006. Also, these evaluations indicate that the consensus of the predictions by a variant of the majority voting outperforms most the predictions from any single predictor or expert. The dataset and description can be found at http://www.predictioncenter.org/casp10/.

We apply our method to the most recent CASP10 experiment in 2012. To evaluate our method, we choose 46 proteins which were classified as the "All groups" and "Regular". We use exactly the same pool of predictions as the quality assessment category in CASP10. About 150 predictions were made from 66 automatic server groups. In the experiment, each predictor is allowed to submit one to five predictions. Thus the corresponding bipartite graph between workers and tasks is dense but not complete. Besides the three-dimensional structure prediction, a "quality assessment" experiment is also performed after these predictions are submitted. Each assessor is asked to rank these 3D predictions without knowing the native structure (a.k.a. the ground truth). Surprisingly, the naive majority voting method in Equation (1) outperforms most assessors and is comparable to the best assessor (Kryshtafovych et al., 2011). We directly compare our method with this baseline. We use the TM-score program (Zhang & Skolnick, 2004) to calculate the performance of selected predictions. A TM-score is between 0 and 1, and measures the similarity between a prediction and the native structure by optimally superimposing them. The similarity matrices S^k are also calculated by the TM-score program.

To deal with multiple predictions from the same predictor, we modify the clustering model as follows. Assume $u_{j,n}^k$ is the associated weight for the n-th prediction by worker j on task k as a medoid, and $w_j^k = \sum_n u_{j,n}^k$ is the corresponding weight on any of worker j's predictions as a medoid. Let $S_{(i,m),(j,n)}^k$ be the similarity between the m-th prediction by worker i and the n-th prediction by worker j on task k. The clustering model now becomes,

$$\max_{\{w_j^k, u_{j,n}^k\}} \sum_{k=1}^{K} \frac{1}{N}[\sum_{i,m} \log[\sum_{j,n} u_{j,n}^k S_{(i,m),(j,n)}^k]] - \alpha \sum_{k=1}^{K} \mathrm{KL}(\frac{1}{K}\sum_k w^k || w^k),$$
$$s.t. \text{ for all } k \text{ and } j,\ w_j^k = \sum_n u_{j,n}^k.$$

This problem can be easily optimized by a modified version of Algorithm 1. The hyper-parameter α is tuned on an independent validation dataset based on the CASP9 experiment in 2010.

	Our method	Random Selection	Majority Voting	Best Worker	Upper bound
Accuracy	**0.4417**	0.3762	0.4245	0.4315	0.4914

Table 1. Performance comparison on CASP10 dataset

The results are shown in Table 1. We compare the top-selected predictions by different methods. The upper bound is calculated by choosing the prediction with the highest accuracy. It is worth noting that the differences in TM-score here are very significant in protein structure prediction (Kryshtafovych et al., 2011). Majority voting, while outperforming most quality assessors in CASP9 and 10 (Kryshtafovych et al., 2011), is still worse than the best worker (`Zhang-server`) in terms of the average accuracy. Our method is better than both majority voting and the best worker. This result is striking considering that in CASP10, no quality assessors or human group obtained better accuracy than `Zhang-server` even when they were given access to all models generated by all server groups. Moreover, the workers' abilities w_j^k estimated from our approach can choose precisely the best workers.

5. Conclusion and future work

In this paper, we have proposed a clustering-based approach for aggregating structured labels from crowdsourcing experiments. Different from classification or regression, many structured output tasks require certain long-rang constraints on the labels, making simple averaging or EM-like algorithms unsuitable for label aggregation. With an exemplar-based clustering model, we are able to estimate the cluster centers with inferred the workers' abilities. An important feature of this approach is that using the medtriods as aggregated labels ensures the validity of the structured labels, e.g. valid protein structures. Applying this method to protein structure prediction, we showed that it performs very well and outperforms both majority voting and the best single worker in the experiment.

A future direction is how to incorporate of the prior information into the model. For example, in protein structure prediction, there are a variety of domain knowledge-based scores to evaluate the predicted structures. One potential solution is to extend the current model to a generalized exponential family to parametrize these scores as features.

Another issue is how to exploit the correlation within the structured output. For instance, a global protein structures can be decomposed into constrained sub-structures; in natural language processing, global parse trees are also decomposable into smaller sub-trees. These sub-structures can be accurate even though the overall structured labels are not. How to estimate the reliability of these sub-structures and assemble them into globally-valid structured labels is a promising future research direction.

References

Cooper, Seth, Khatib, Firas, Treuille, Adrien, Barbero, Janos, Lee, Jeehyung, Beenen, Michael, Leaver-Fay, Andrew, Baker, David, Popović, Zoran, et al. Predicting protein structures with a multiplayer online game. *Nature*, 466(7307):756–760, 2010.

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

Karger, D.R., Oh, S., and Shah, D. Iterative learning for reliable crowdsourcing systems. *Neural Information Processing Systems (NIPS), Granada, Spain*, 2011.

Kryshtafovych, A., Fidelis., K., and Tramontano, A. Evaluation of model quality predictions in casp9. *Proteins: Structure, Function and Bioinformatics*, 79(S10), 2011.

Lashkari, D. and Golland, P. Convex clustering with exemplar-based models. In *Neural Information Processing Systems Conference (NIPS)*, 2010.

Lee, Michael D, Steyvers, Mark, de Young, Mindy, and Miller, Brent. Inferring expertise in knowledge and prediction ranking tasks. *Topics in cognitive science*, 4(1):151–163, 2012.

Liu, Qiang, Peng, Jian, and Ihler, Alex. Variational inference for crowdsourcing. In *Advances in Neural Information Processing Systems 25*, pp. 701–709, 2012.

Snow, R., O'Connor, B., Jurafsky, D., and Ng, A.Y. 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, 2008.

Welinder, P., Branson, S., Belongie, S., and Perona, P. The multidimensional wisdom of crowds. In *Neural Information Processing Systems Conference (NIPS)*, 2010.

Wu, Xian, Fan, Wei, and Yu, Yong. Sembler: Ensembling crowd sequential labeling for improved quality. In *Twenty-Sixth AAAI Conference on Artificial Intelligence*, 2012.

Zhang, Y. and Skolnick, J. Scoring function for automated assessment of protein structure template quality. *Proteins: Structure, Function and Bioinformatics*, 57, 2004.