Inferring Missing Relevance Judgments from Crowd Workers via Probabilistic Matrix Factorization

Hyun Joon Jung
Dept. of Electrical and Computer Engineering
University of Texas at Austin
hyunJoon@utexas.edu

Matthew Lease
School of Information Science
University of Texas at Austin
ml@ischool.utexas.edu

ABSTRACT

In crowdsourced relevance judging, each crowd worker typically judges only a small number of examples, yielding a sparse and imbalanced set of judgments in which relatively few workers influence output consensus labels, particularly with simple consensus methods like majority voting. We show how probabilistic matrix factorization, a standard approach in collaborative filtering, can be used to infer missing worker judgments such that all workers influence output labels. Given complete worker judgments inferred by PMF, we evaluate impact in unsupervised and supervised scenarios. In the supervised case, we consider both weighted voting and worker selection strategies based on worker accuracy. Experiments on crowd judgments from the 2010 TREC Relevance Feedback Track show promise of the PMF approach merits further investigation and analysis.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Design, Experimentation, Performance

Keywords

Crowdsourcing, matrix Factorization, label aggregation

1. INTRODUCTION

Crowdsourced relevance judging offers potential to reduce time, cost, and effort of relevance judging [1] and benefit from greater diversity of crowd judges. However, quality of judgments from non-workers continues to be a concern, motivating continuing work in quality assurance methods based on statistical label aggregation methods or greater attention to human factors. A common approach is to collect multiple, redundant judgments from workers and aggregate them via methods like majority voting (MV) or expectation maximization (EM) to produce consensus labels [4].

Because each crowd worker typically judges only a small number of examples, collected judgments are typically sparse and imbalanced, with relatively few workers influencing output consensus labels. MV is completely susceptible to this problem. EM addresses this indirectly: while only workers labeling an example vote on it, global worker judgments are used to infer class priors and worker confusion matrices.

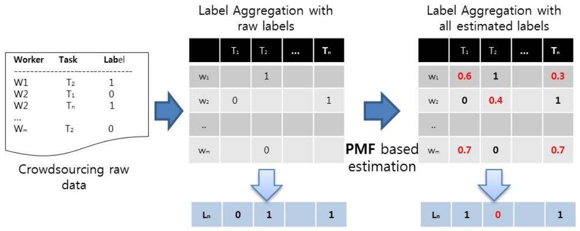

Figure 1: Crowdsourcing workers judgments (Left) are copied to a sparse worker-task matrix (Middle). Missing judgments are inferred via PMF (Right).

We propose to tackle this issue directly by adopting a collaborative filtering approach, which routinely deals with the issue of each user rating only a small number of items (e.g., movies, books, etc.) vs. the complete set. In particular, we employ probabilistic matrix factorization (PMF), which induces a latent feature vector for each worker and example [6] in order to infers unobserved worker judgments for all examples. **Figure 1** depicts our approach graphically.

We are not familiar with any prior work investigating PMF, or collaborative filtering approaches more generally, toward crowdsourcing quality assurance. Related prior work has investigated other ways to infer bias corrected labels in place of raw labels [4], as well as inference of missing labels by estimating a unique classifier for each worker [3].

Probabilistic Matrix Factorization (PMF). Suppose we have M tasks (examples to be labeled), N workers, and a label matrix R in which R_{ij} indicates the label of worker i for task j. Let $U \in \mathbb{R}^{D*M}$ and $V \in \mathbb{R}^{D*N}$ be latent feature matrices for workers and tasks, with column vectors U_i and V_j representing D-dimensional worker-specific and task-specific latent feature vectors, respectively. The conditional probability distribution over the observed labels $R \in \mathbb{R}^{N*M}$ is given by Equation 1. Indicator I_{ij} equals 1 *iff* worker i labeled task j. We place zero-mean spherical Gaussian priors on worker and task feature vectors (Equations 2 and 3).

$$p(R|U,V,\sigma^2) = \prod_{i=1}^{N}\prod_{j=1}^{M}[\mathcal{N}(R_{ij}|U_i^T V_j, \sigma^2)]^{I_{ij}} \tag{1}$$

$$p(U|\sigma_U^2) = \prod_{i=1}^{N}[\mathcal{N}(U_i|0, \sigma_U^2 \mathrm{I})] \tag{2}$$

$$p(V|\sigma_V^2) = \prod_{j=1}^{M}[\mathcal{N}(V_j|0, \sigma_V^2 \mathrm{I})] \tag{3}$$

SIGIR'12, August 12–16, 2012, Portland, Oregon, USA.
ACM 978-1-4503-1472-5/12/08.

Method	Supervised	Worker Labels	Label Aggregation	ACC	Rank	RMSE	Rank	SPE	Rank
1	No	raw (sparse)	MV	0.603	4	0.63	4	0.332	6
2	No	raw (sparse)	EM	**0.644**	**3**	**0.596**	**3**	0.418	**4**
3	No	PMF (complete)	MV	**0.643**	**3**	**0.598**	**3**	**0.440**	5
4	Yes	raw (sparse)	WV	0.642	3	0.598	3	**0.900**	**1**
5	Yes	raw (sparse)	Filtering(α=0.67)	**0.752**	**1**	**0.498**	**1**	0.838	2
6	Yes	raw (sparse)	WV & Filtering(α=0.67)	**0.750**	**1**	**0.500**	**1**	0.848	2
7	Yes	PMF (complete)	WV & Filtering(α=0.7)	0.673	2	0.571	2	0.542	3

Table 1: Results of PMF-based inference of missing worker labels. For the unsupervised case, majority voting (MV) with PMF (Method 3) is compared to MV and EM approaches using input (sparse) worker labels (Methods 1-2). With supervision, we compare weighted voting (WV) and/or filtering with and without PMF. Ranks shown indicate statistically significant differences at $p <= 0.05$ using a two-tailed paired t-test.

To estimate model parameters, we maximize the log-posterior over task and worker features with fixed hyper-parameters. Maximizing the posterior with respect to U and V is equivalent to minimizing squared error with L2 regularization:

$$\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{M} I_{ij}(R_{ij}-U_i^T V_j)^2 + \frac{\lambda_U}{2}\sum_{i=1}^{N}\|U_i\|_F^2 + \frac{\lambda_V}{2}\sum_{i=1}^{M}\|V_j\|_F^2$$

where $\lambda_U = \sigma_U/\sigma, \lambda_V = \sigma_V/\sigma$, and $\|.\|_F^2$ denotes the Frobenius Norm. We use gradient descent to find a local minimum of the objective for U and V. Finally, we infer missing worker judgments in the worker-task matrix R by taking the scalar product of U and V. Note that as in [4], we also replace actual labels with bias-corrected inferred labels.

Label Aggregation. Given the complete set of inferred worker relevance judgments in matrix R, we next aggregate worker judgments to induce consensus labels. We consider both unsupervised supervised scenarios. In the former, we consider majority voting with raw (sparse) labels (Method 1), expectation maximization with raw labels (Method 2), and PMF-based MV (Method 3). In the supervised case, we measure each worker's accuracy based on expert judgments, with labels of anti-correlated workers flipped such that accuracy is always $\geq 50\%$. We use supervision in two distinct ways: weighted voting (WV) and worker filtering, in which only workers with accuracy $\geq \alpha$ participate in voting.

2. EVALUATION

Experiments are performed on crowd judgments collected in the 2010 TREC Relevance Feedback Track [2] from Amazon Mechanical Turk. 762 crowd workers judged 19033 query-document tasks (examples), and 89624 judgments were collected. Our worker-task matrix thus has 762 columns (workers) and 19,033 rows (tasks); only 89,624 out of 14,503,146 labels (0.6%) are observed, so data is extremely sparse. 3,275 expert relevance judgments by NIST are partitioned into training (2,275) and test (1,000) sets. The test set is evenly-balanced between relevant and non-relevant classes.

Parameters. For dimensionality of task and worker latent feature vectors, we consider $D \in 10, 30, 50$ and select $D = 30$ based on cross-validation on the entire set of labels (unsupervised). We similarly tune regularization parameter $\lambda \in \{0.001, 0.01, 0.1, 0.5\}$ and select $\lambda = 0.1$. We tune the worker filtering threshold $\alpha \in [0.6, 0.99]$ by cross-validation on the training set using a linear sweep with step-size 0.01.

Metrics and Results. Table 1 reports accuracy (ACC), RMSE, and specificity achieved by each method.

Unsupervised Methods. Method 2 of PMF with majority voting (MV) outperforms the MV baseline (Method 1) and performs equivalently to EM (Method 2).

Supervised vs. Unsupervised Methods. While supervised methods tend to dominate, unsupervised EM and PMF both match performance of the supervised weighted voting (WV) method without filtering or PMF (Method 4).

Supervised Methods. Worker filtering is clearly seen to provide the greatest benefit, and surprisingly performs better without PMF than with PMF (Methods 6 vs. 7). When filtering is used, use of WV is not seen to further improve performance (Methods 5 vs. 6). We do see PMF-based modeling outperform non-PMF modeling when worker filtering is not employed (Methods 7 vs. 4).

3. CONCLUSION

While unsupervised consensus labeling accuracy with PMF only matched EM performance, PMF is advantageous in that once complete worker judgments are inferred, they might be used for a variety of other purposes, such as better routing or recommending appropriate tasks to workers.

Intuitively, an accurate worker's empirical label distribution should resemble the actual class prior. This suggests an alternative, more weakly supervised scenario to consider in which class priors are known while example labels are not. In the unsupervised case, we might instead simply examine the distribution of empirical priors for each worker and detect outliers [5]. In future work, we plan to investigate these ideas further in combination with those described here.

4. REFERENCES

[1] O. Alonso, D. Rose, and B. Stewart. Crowdsourcing for relevance evaluation. *SIGIR Forum*, 42(2):9–15, 2008.

[2] C. Buckley, M. Lease, and M. D. Smucker. Overview of the TREC 2010 Relevance Feedback Track (Notebook). In *Proc. of the 19th Text Retrieval Conference*, 2010.

[3] S. Chen, J. Zhang, G. Chen, and C. Zhang. What if the irresponsible teachers are dominating? a method of training on samples and clustering on teachers. In *24th AAAI Conference*, pages 419–424, 2010.

[4] P. 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.

[5] H. J. Jung and M. Lease. Improving Consensus Accuracy via Z-score and Weighted Voting. In *AAAI Workshop on Human Computation (HComp)*, 2011.

[6] R. Salakhutdinov and et al. Probabilistic matrix factorization. In *NIPS 2008*, volume 20, January 2008.