A Study on Quality Management with Relevance Judgment Data from the Wisdom of Crowds

Hyun Joon Jung
Dept. of Electrical and Computer Engineering
University of Texas at Austin

ABSTRACT

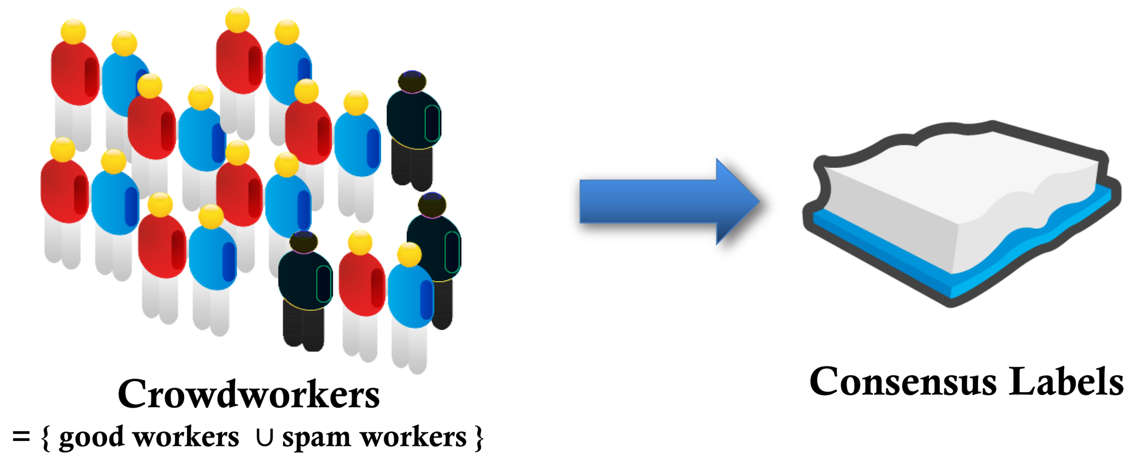
The objective of this project is to study a quality management on crowdsourcing with a large-scale relevance judgment dataset. First, we propose a novel multiple feature weighted majority voting algorithm with seven features of each worker. Next, we try to find outliers by our Z-score based detection algorithm.

Our experiments presents that our proposed weighted majority outperforms the existing single feature weighted majority algorithm. Moreover, Z-score based outlier detection improves the accuracy of consensus with the proposed weighted majority algorithm.

CONTACT

Hyun Joon Jung
The Univ. of Texas at Austin
Email: hyun@mail.utexas.edu
Phone: 512-567-5195

INTRODUCTION



Objective: Improve the Quality of Consensus Labels

Challenge
1. Who are spam workers and What' the features of them?
2. How to filter them out?

Solution
- Use Multiple Feature of Crowdworkers
- Use Multiple Feature based Weighted Majority
- Use Z-score based Spam Worker Detection

DataSet

Relevance Judgment Data(Amazon Mechanical Turk)

- The workers' task: to judge a query / document pairs on a ternary scale: irrelevant, relevant, and strongly relevant
- The number of unique examples: **3,277** examples with gold expert judgments (1,501 non-relevant, 863 relevant, 913 strongly relevant)
- **766** workers annotates 19,232 labels with unique 3,277 examples. (**5.88** labels per example)
- The number of **Broken Link Task** : 1,183 additional tasks

Feature Generation

Feature 1 - Graded accuracy vs. gold (GACG)

Feature 2 - Binary accuracy vs. gold (BACG)

Feature 3 - Graded accuracy vs. majority vote (GACM)

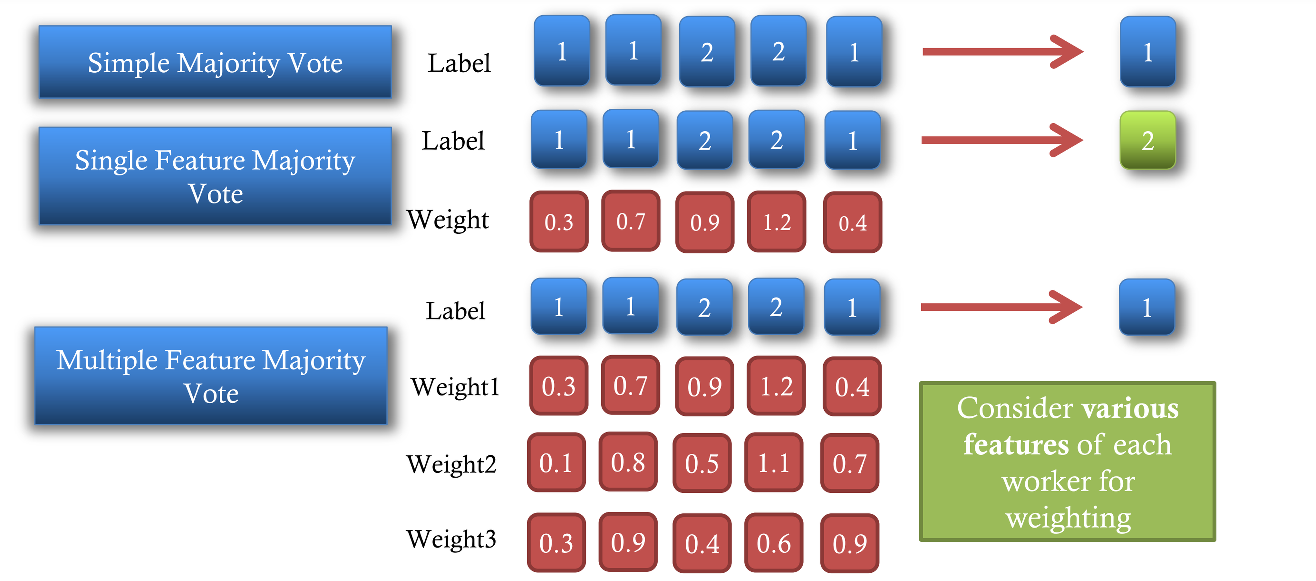
Feature 4 - Binary accuracy vs. majority vote (BACM)

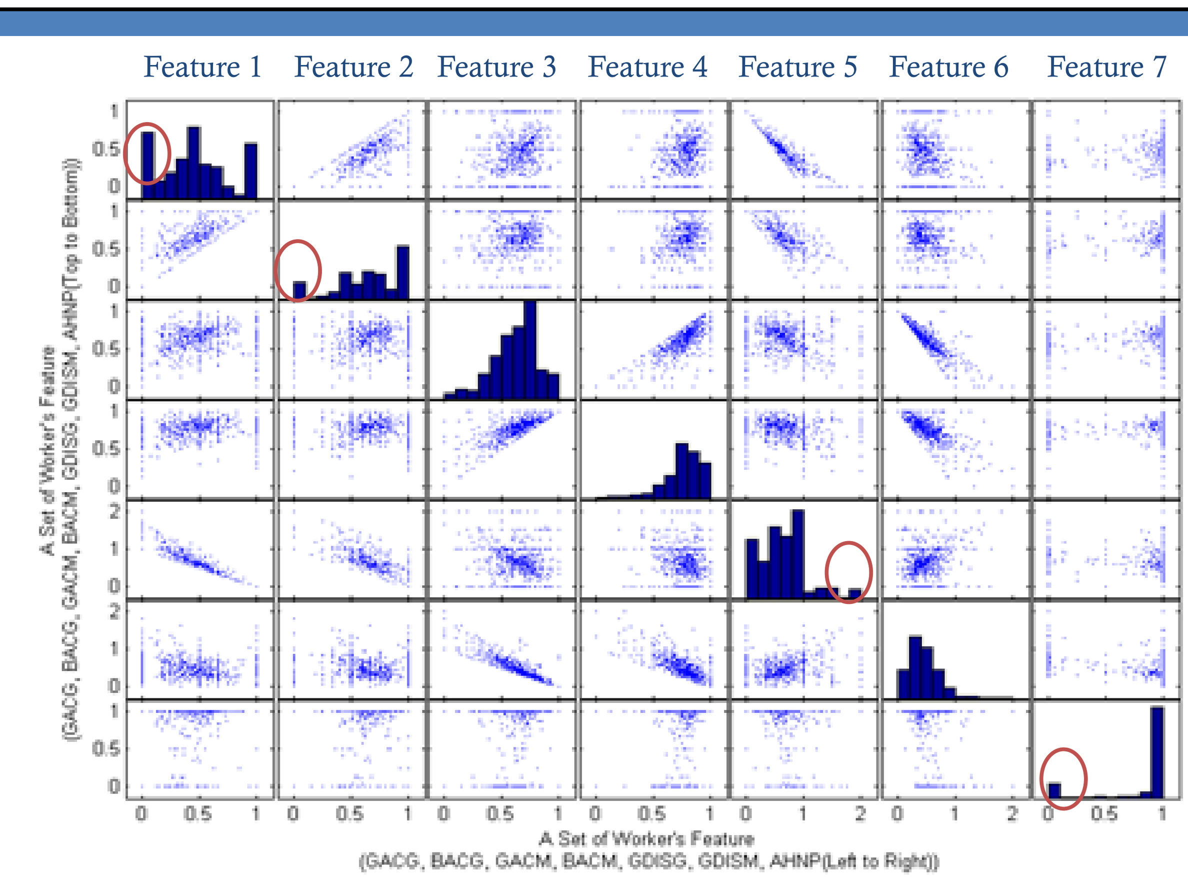
Feature 5 - Graded distance vs. gold (GDSG)

Feature 6 - Graded distance vs. majority vote (GDSM)

Feature 7 - Accuracy vs. broken-links (AHNP)

Weighted Majority Voting





Z-score based Outlier Detection

Z-Score

Given a set of features of workers, we apply Z-scores based outlier detection which is based on the property of the normal distribution that if $X \sim N(\mu, \rho^2)$, then $Z = \frac{X-\mu}{\sigma} \sim N(0,1)$. Z-scores are a very popular method for labeling outliers and are defined as following:

$$Z_{score}(i) = \frac{x_i - \bar{x}}{s}, \text{where } s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Z-Score based outlier Detection Algorithm

1. Compute Z-score of all features of each worker
2. If one of Z-score > threshold THEN filter out the worker's labels
3. Re-compute the consensus label by using a majority vote with anyone who survives.
4. Re-compute the accuracy of consensus labels over gold.

Evaluation

Objective

- Tuned the Z-score threshold parameter with simple linear sweep from (by 0.1)
- Both tuning and testing use 5-fold cross-validation on the set of 3,277 examples

Configuration

- 5-fold cross validation(Training Set, Testing set)
- Simple Majority(SM), Single Feature Weighted Majority(SWM), Multiple Feature Weighted Majority(MWM)

Accuracy of Consensus Label

	GA +SM	BA +SM	GA+SWM	BA+SWM	GA+MWM	BA+MWM
f1(GACG)	0.4867(0.5)	0.6597(0.4)	0.5041(0.7)	0.6747(1.5)	0.5072(1.5)	0.6811(1.5)
f2(BACG)	0.4944(0.5)	0.6692(0.5)	0.4965(0.5)	0.6769(1.0)	0.5038(1.0)	0.6781(1.0)
f3(GACM)	0.4751(2.0)	0.6546(2.0)	0.4940(0.6)	0.6674(0.7)	0.5066(1.0)	0.6704(1.0)
f4(BACM)	0.4751(3.0)	0.6561(1.8)	0.4901(0.8)	0.6689(0.8)	0.5050(0.8)	0.6750(0.8)
f5(GDISG)	0.4883(0.6)	0.6622(0.6)	0.4995(1.3)	0.6744(1.3)	0.5072(1.3)	0.6805(1.3)
f6(GDISM)	0.4764(1.6)	0.6549(2.2)	0.4986(0.9)	0.6704(1.0)	0.5063(1.1)	0.6708(1.1)
f7(AHNP)	0.5035(1.6)	0.6823(1.7)	0.4974(1.7)	0.6457(1.7)	0.4999(1.6)	0.6515(1.7)
f1+f5+f7	0.5172(0.6)	**0.6866(0.9)**	0.5188(0.7)	**0.6793(1.2)**	**0.5191(1.4)**	0.6827(1.5)
f1+f7	**0.5176(0.6)**	0.6860(0.7)	**0.5194(0.6)**	0.6790(1.3)	0.5191(1.5)	**0.6833(1.5)**
f1+f5	0.5075(0.3)	0.6619(0.6)	0.5011(0.4)	0.6744(1.3)	0.5069(1.5)	0.6805(1.5)
Baseline	**0.4751**	**0.6315**	**0.4822**	**0.6402**	**0.4883**	**0.6488**
ALL	**0.5044(1.8)**	**0.6830(2.0)**	**0.5154(1.1)**	**0.6781(1.5)**	**0.5179(1.3)**	**0.6811(1.9)**

Consensus label accuracy using Z-score worker filtering (with different features) vs. different voting schemes. Bold result indicates best accuracy in the given column. Entries also show the threshold parameter value (γ) used.

CONCLUSIONS

- We present a Z-score based outlier detection algorithm for filtering out low-quality crowd workers.
- We find that filtering in combination with multi-feature weighted majority voting reduces the error of consensus accuracy by 8.94% absolute for graded accuracy and 5.32% for binary accuracy.

REFERENCES

- Ipeirotis, P.G. 2011. The unreasonable effectiveness of simplicity. February 6. http://behind-theenemy-lines.blogspot.com/2011/2/unreasonableeffectivenss-of.html
- Panagiotis G. Ipeirotis, Foster Provost, Jing Wang, Quality Management on Amazon Mechanical Turk, ACM KDD-HCOMP'10
- Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis, Get Another Label? Improving Data Quality and Data Mining Using Multiple, Noisy Labelers, ACM KDD'08