SQUARE: Benchmarking Crowd Consensus at MediaEval

Aashish Sheshadri
Department of Computer Science
The University of Texas at Austin
aashishs@cs.utexas.edu

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

ABSTRACT

We extend the SQUARE benchmark for statistical consensus methods to include additional evaluation on two datasets from the MediaEval 2013 Crowdsourcing in Multimedia shared task. In addition to reporting shared task results, we also analyze qualitatively and quantitatively performance of consensus algorithms under varying supervision.

1. ALGORITHMS

We extend SQUARE[1] [5], a benchmark for evaluating statistical consensus algorithms, to include additional evaluation on two datasets from MediaEval 2013 Crowdsourcing in Multimedia shared task[2]. Algorithms are briefly summarized below; the SQUARE paper [5] provides more detailed discussion and comparative analysis. Because crowdsourcing allows rapid generation of datasets for new tasks, where no feature representation of inputs or automatic labeling algorithm may yet exist, we intentionally exclude hybrid methods requiring automatic label generation from features.

Majority Voting (MV) uses random tie-breaking to avoid bias in absence of an informative prior for tie-breaking. ZenCrowd (ZC) [2] implements unsupervised EM to jointly estimate worker accuracies and labels. GLAD [8] implements unsupervised EM to jointly estimate labels and worker expertise while modeling example difficulty. Dawid-Skene (DS) [1] and Raykar et al. (RY) [4] implement unsupervised EM to jointly estimate labels and worker confusion matrices, with RY differing from DS in modeling individual worker priors. Naive Bayes (NB) [6] implements a fully supervised estimation of worker confusion. CUBAM [7]'s unsupervised MAP estimation jointly models example difficulty and annotator specific measurements of noise, expertise and bias.

2. DATA AND EXPERIMENTAL SETUP

Data. Consensus algorithms are evaluated on the MMSys 2013 and the test fraction (20%) of Fashion 10000. Both datasets elicit binary judgments on the same set of tasks from Amazon Mechanical Turk (AMT) workers. Task 1 asks workers to identify an image as being fashion-related or not, and Task 2 asks workers to indicate whether or not a desired fashion object is present. For further details of tasks, see [3].

[1] ir.ischool.utexas.edu/square
[2] www.multimediaeval.org/mediaeval2013/crowd2013

MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain

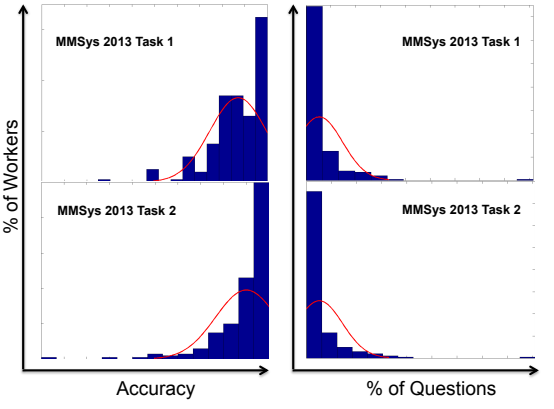

Figure 1: **Left histogram**: distribution of worker accuracies. **Right histogram**: # of examples labeled per worker.

"Gold" data is defined for these tasks by majority voting of trusted annotators over examples with a clear majority label.

Experiment. Following the same benchmarking procedure as in SQUARE [5], we consider 5 degrees of supervision: 10%, 20%, 50%, 80%, and 90%. In each case, we use cross-fold validation, i.e. for the 10% supervision setting, estimation uses 10% train data and is evaluated on the remaining 90%, this procedure is repeated across the other nine folds, finally, average performance across the folds is reported. We report unsupervised performance on 100% of data, with evaluation limited to examples with gold labels.

In the unsupervised setting, uninformed, task-independent hyper-parameters and class priors are unlikely to be optimal. While one might optimize these parameters by maximizing likelihood over random restarts or grid search, we do *not* attempt to do so. Instead, with *light-supervision*, we assume no examples are labeled, but informative priors are provided (matching the training set empirical distribution). *Full-supervision* assumes gold-labeled examples are provided. To evaluate ZC, RY, DS and GLAD methods under full-supervision, labels are predicted for all examples (without supervision) but replaced by gold on training data.

3. RESULTS AND DISCUSSION

MMSys 2013. Average performance over Tasks 1 and 2 are reported in Table 1. We highlight best scoring methods but note these are not necessarily statistically significant.

Results show close to constant performance (roughly 91-

Method	Metric	No Supervision	Light-Supervision 10%	20%	50%	80%	90%	Full-Supevision 10%	20%	50%	80%	90%	Count
MV	*Acc*	91.10	91.05	91.05	91.10	91.00	91.10	91.05	91.05	91.10	91.00	91.10	0
	F_1	91.25	90.95	90.95	91.00	90.90	91.05	90.95	90.95	91.00	90.90	91.05	1
ZC	*Acc*	91.35	**91.35**	**91.35**	**91.35**	91.20	**91.25**	**91.40**	91.40	91.50	91.40	91.40	5
	F_1	91.35	91.20	91.25	91.25	91.05	91.05	91.20	91.25	91.35	91.25	91.25	5
GLAD	*Acc*	91.25	91.25	91.25	**91.35**	**91.30**	91.10	91.35	91.30	91.45	91.45	91.40	2
	F_1	91.35	91.10	91.05	91.20	91.10	90.90	91.10	91.10	91.25	91.25	91.20	1
NB	*Acc*	-	-	-	-	-	-	90.95	91.15	91.40	91.30	91.40	0
	F_1	-	-	-	-	-	-	90.80	91.00	91.25	91.15	91.25	0
DS	*Acc*	91.10	90.95	90.95	90.85	90.55	89.90	91.05	91.10	91.45	**91.85**	**91.90**	2
	F_1	90.95	90.75	90.70	90.60	90.25	89.60	90.85	90.90	91.20	91.65	91.70	2
RY	*Acc*	**91.55**	91.05	91.05	91.20	91.15	91.15	91.25	**91.50**	**91.60**	91.75	91.75	3
	F_1	91.45	90.85	90.90	91.10	91.00	91.05	91.05	91.35	91.40	91.60	91.65	4
CUBAM	*Acc*	91.10	-	-	-	-	-	-	-	-	-	-	0
	F_1	91.35	-	-	-	-	-	-	-	-	-	-	0

Table 1: Accuracy and F_1 results when averaged over both tasks on MMSys 2013 for varying supervision *type* (none, light, and full) and *amount* (10%, 20%, 50%, 80%, and 90%). Maximum values for each metric across methods in each column are bolded (Accuracy) and underlined (F_1). As a simple summary measure, the final column counts the number of result columns (out of 11) in which a given method achieves the maximum value for each metric.

	$T2-Yes$	$T2-No$
$T1-Yes$	**36.86%**	0.34%
$T1-No$	15.43%	**47.37%**

Table 2: Confusion matrix over gold labels assigned to Task 1 (T1) and Task 2 (T2) of MYSys 2013.

92%) across methods, metrics, and varying supervision. Average accuracy of 91.17% across methods, with minimal variance, is perhaps not surprising given the abundance of high quality workers for these tasks (see Figure 1 for details). We hypothesize three key factors. **1.** the response redundancy is limited at most three, with many instances recording only two responses (since workers could select 'not sure' rather than provide a binary judgment). **2.** a high percentage of the work was completed by relatively small percentage of the workers (Figure 1). **3.** Consistently high worker accuracies limit the value of weighted worker voting vs. simple MV.

We further observe task 1 and 2 to be highly correlated. Table 2 shows the confusion matrix of gold labels across the two tasks. Evaluation on Task 2 using Task 1 gold labels for supervision achieved 82.89% average accuracy and 84.98% average F_1 across methods. This presents a possibility of joint estimation we do not investigate.

Fashion 10000. Table 3 reports F_1 scores on the test fraction of Fashion 10000; since gold data was not available for the blind shared task, we only present results estimated with no supervision. MV is seen to score best on Task 1, and CUBAM on Task 2. This contrasts findings on MMSys 2013, though we still observe only a 1% threshold across methods on Task 1. However, Task 2 shows markedly superior performance of CUBAM, which remains to be investigated.

Acknowledgments. This work is supported by National Science Foundation grant IIS-1253413 and by a Temple Fellowship. Opinions expressed in this work are those of the authors and do not reflect views of the sponsors.

Method	Task 1	Task 2
MV	**73.25**	75.43
ZC	72.67	74.88
GLAD	72.89	75.67
DS	72.82	75.20
RY	72.87	75.01
CUBAM	73.03	**77.46**

Table 3: F_1 scores on the test fraction of Fashion 10000.

4. REFERENCES

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

[2] G. Demartini, D. E. Difallah, and P. Cudré-Mauroux. Zencrowd: leveraging probabilistic reasoning and crowdsourcing techniques for large-scale entity linking. In *Proc. WWW*, pages 469–478, 2012.

[3] B. Loni, M. Larson, A. Bozzon, and L. Gottlieb. Crowdsourcing for Multimedia at MediaEval 2013: Challenges, datasets, and evaluation. In *MediaEval 2013 Workshop*, Barcelona, Spain, October 18-19 2013.

[4] V. C. Raykar, S. Yu, L. H. Zhao, and G. H. Valadez. Learning from crowds. In *Journal of Machine Learning Research 11 (2010) 1297-1322*, MIT Press, 2010.

[5] A. Sheshadri and M. Lease. SQUARE: A Benchmark for Research on Computing Crowd Consensus. In *Proceedings of the 1st AAAI Conference on Human Computation (HCOMP)*, 2013.

[6] 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, 2008.

[7] P. Welinder, S. Branson, S. Belongie, and P. Perona. The multidimensional wisdom of crowds. In *NIPS*, pages 2424–2432, 2010.

[8] J. Whitehill, P. Ruvolo, T. Wu, J. Bergsma, and J. Movellan. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *NIPS*, pages 2035–2043, 2009.