AGGREGATION OF CROWDSOURCED LABELS BASED ON WORKER HISTORY

Mihai Georgescu, Xiaofei Zhu

L3S Research Center
Leibniz Universität Hannover

Introduction

- Introduce of a novel yet simple method for aggregation of different crowdsourced labels, taking into account the worker expertise (confidence)
- Assess different ways of computing the worker confidence, as well as various ways of incorporating it in the computation of the aggregated label
- Evaluation on different datasets and comparison with other state-of-the art methods

Crowdsourcing

- Crowdsourcing is the act of taking a job traditionally performed by a designated agent (usually an employee) and outsourcing it to an undefined, generally large group of people in the form of an open call.[1]

- The crowd workers are motivated by a small financial incentive
- Usually done via microtask platforms such as Amazon’s Mechanical Turk or Crowdflower
- Requester posts HITs that are solved by workers for a financial reward
- Unknown workers with various expertise can replace domain experts
- Advantages: cost effective, workers availability and diversity
- Disadvantages: questionable quality of work

1) crowdsourcing.typepad.com

Crowdsourcing & Machine Learning

- Crowdsourcing is widely used for label acquisition in supervised machine learning, alleviating the need of hiring experts sometimes
- The quality of crowdsourced work is questionable
- Redundancy often employed, requiring multiple labels
- Need to aggregate multiple noisy labels to create reliable labeled data

- Commonly used aggregation methods:
  - Majority voting
  - EM based algorithms that provide the hidden labels and evaluate the workers simultaneously

Problem statement

- Objective: Infer labels from multiple and possibly noisy labels (acquired via crowdsourcing) assuming no authoritative ground truth is available

- Solution: An improved EM method with a flexible mutually reinforced integration of the worker confidence in the aggregated label
  - E Step: compute the aggregated crowd label of instances
  - M Step: update the worker confidence

Crowd Aggregated Label

- Aggregation of the labels from all workers $L_w^i \in \{Yes, No\}$
- Each worker's contribution is weighted based on his expertise
- Crowd Soft Label ϵ[0,1] (positive or negative)indicate how reliable the aggregated label is.
- Crowd Hard Label ϵ{Yes,No} final label

positive soft label

negative soft label

$$L_{crowd}^i = \begin{cases} Yes, & \boxed{l_i^+} - \boxed{l_i^-} \geq 0 \\ No, & l_i^+ - l_i^- < 0 \end{cases}$$

- Variations:
  - Boosting of worker confidence in the aggregated label
  - Involvement of self-reported expertise assessment

Worker confidence

- Accuracy of the individual worker labels when compared to Crowd Labels
- Variations:
  - Discrimination between positive/ negative label quality
    - No discrimination

$$C_w^* = \frac{tp_w + tn_w}{tp_w + tn_w + fp_w + fn_w}$$

    - Discrimination

$$C_w^+ = \frac{tp_w}{tp_w + fp_w} \qquad C_w^- = \frac{tn_w}{tn_w + fn_w}$$

  - Hard or soft evaluation depending on type of Crowd Label used

Aggregated Crowd Label Computation (E Step)

- No discrimination between positive and negative label quality

$$C_w^* = \frac{tp_w + tn_w}{tp_w + tn_w + fp_w + fn_w}$$

$$l_i^+ = \frac{\sum_w \boxed{C_w^*} \cdot I(L_w^i = Yes)}{\sum_w C_w^* \cdot I(L_w^i = Yes) + \sum_w C_w^* \cdot I(L_w^i = No)}$$

- Discrimination between positive and negative label quality

$$l_i^+ = \frac{\sum_w C_w^+ \cdot I(L_w^i = Yes)}{\sum_w \boxed{C_w^+} \cdot I(L_w^i = Yes) + \sum_w \boxed{C_w^-} \cdot I(L_w^i = No)}$$

$$C_w^+ = \frac{tp_w}{tp_w + fp_w} \qquad C_w^- = \frac{tn_w}{tn_w + fn_w}$$

$$I(x) = \begin{cases} 0, & x = false \\ 1, & x = true \end{cases}$$

Boosting: $\hat{C}_w = boost(C_w) \quad e^x \text{ or } x^p; p \in \mathbb{R}$

Worker confidence computation (M Step)

Hard Evaluation

- Examine all items for which the worker provided a label and assess if it coincides with the crowd aggregated hard label depending on its type

Soft Evaluation

- Use the crowd soft labels coupled with the answers provided by the worker, when assessing the workers confidence over all the items he provided labels for

$$tp_w = \sum_i I(L_w^i = Yes) \cdot I(L_{crowd}^i = Yes)$$

$$tn_w = \sum_i I(L_w^i = No) \cdot I(L_{crowd}^i = No)$$

$$fp_w = \sum_i I(L_w^i = Yes) \cdot I(L_{crowd}^i = No)$$

$$fn_w = \sum_i I(L_w^i = No) \cdot I(L_{crowd}^i = Yes)$$

$$tp_w = \sum_i I(L_w^i = Yes) \cdot l_i^+$$

$$tn_w = \sum_i I(L_w^i = No) \cdot l_i^-$$

$$fp_w = \sum_i I(L_w^i = Yes) \cdot l_i^-$$

$$fn_w = \sum_i I(L_w^i = No) \cdot l_i^+$$

Method settings

- Type of boosting function applied
- Discrimination between quality of positive and negative labels
- Soft of hard evaluation of the worker confidence

Evaluation

- Datasets
- Settings vs. Performance
- Comparison to Majority Voting
- Involvement of Self-Reported familiarity
- Comparison to other state-of-the art aggregation methods

Datasets

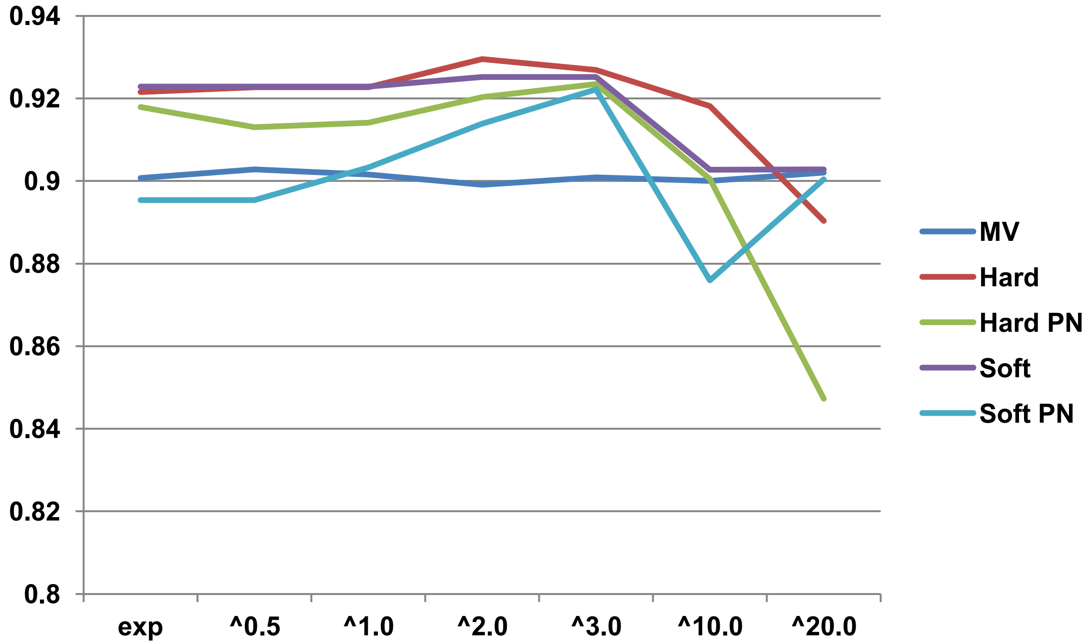
- HCB
  - Conflated relevance judgements
- WB
  - Images contain ducks or not
- WVSCM
  - Images contain enjoyment or social smiles
- RTE
  - Textual entailment judgements
- MEval(MMsys)
  - Images from the fashion domain
  - Label1
    - Is the image related to fashion
  - Label2
    - Is a certain category present in the image
    - A familiarity with the category is requested

Dataset	Items	Workers	Labels	GT Items
HCB	19033	762	88385	2275
WB	240	53	9600	240
WVSCM	2134	64	17729	159
RTE_RTE	800	164	8000	800
RTE_TEMP	462	76	4620	462
MEval-Label1	31076	1429	89449	5750
MEval-Label2	31039	1426	87840	5986
MMSys-Label1	4711	202	13727	13727
MMSys-Label2	4710	208	13474	13474

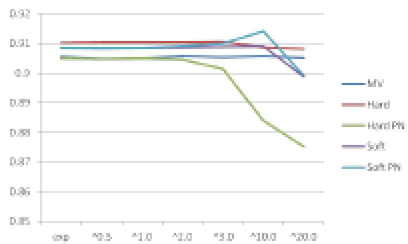
Settings vs. Performance

Plotted performance in terms of F1 measure of all settings and compared to MV across all datasets.

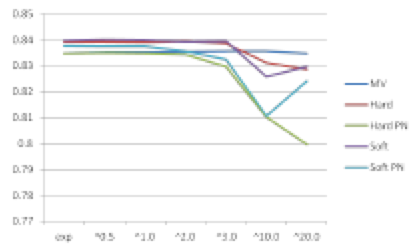
Example: RTE_RTE



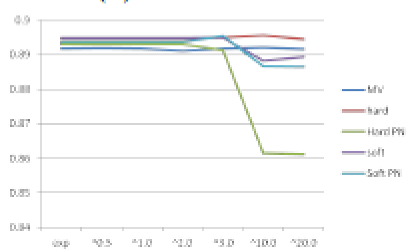
Settings vs. performance (F1)



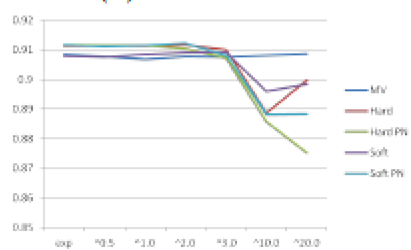
(a) MEval-Label1



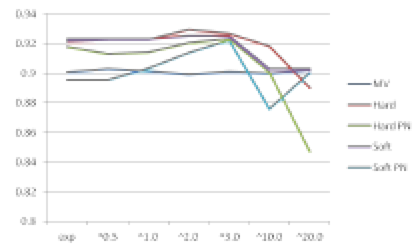
(b) MEval-Label2



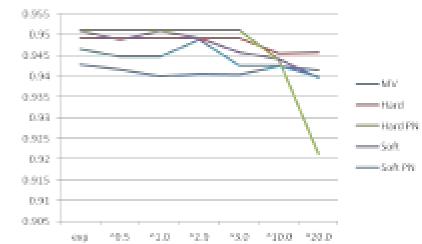
(c) MMSys-Label1



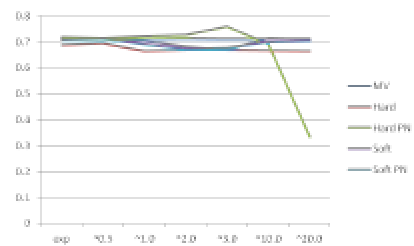
(d) MMsys-Label2



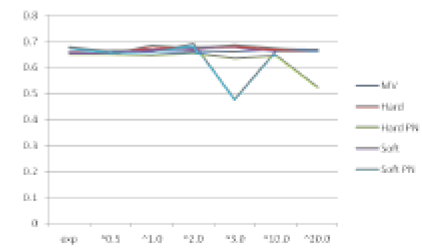
(e) RTE-RTE



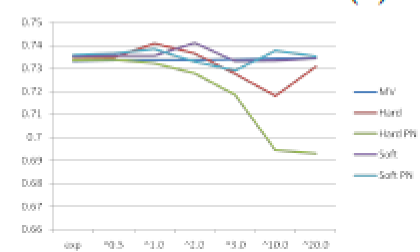
(f) RTE-TEMP



(g) WB



(h) WVSCM



(i) HCB

Majority Voting vs Best Setting

Dataset	Eval	PN	Boost	F1	MV - F1	Improvement
HCB	soft	no	x^2	0.7410	0.735717	0.0053
WB	hard	yes	x^3	0.7577	0.709924	0.0478
WVSCM	hard	no	x^3	0.6857	0.666667	0.0190
RTE_RTE	hard	no	x^2	0.9295	0.893112	0.0364
RTE_TEMP	hard	yes	x^1	0.9511	0.948617	0.0025
MEval-Label1	soft	yes	x^10	0.9142	0.906695	0.0075
MEval-Label2	soft	no	x^0.5	0.8400	0.836652	0.0033
MMSys-Label1	soft	yes	x^3	0.8950	0.890581	0.0044
MMSys-Label2	soft	yes	x^2	0.9336	0.905926	0.0277

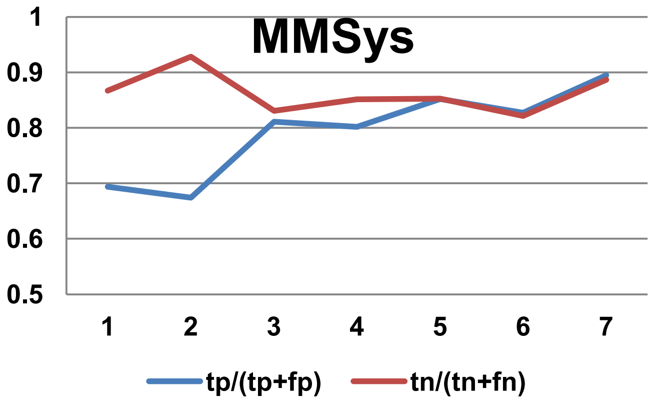
Involvement of familiarity

- For the MMSys and Meval (fashion domain) additional information is requested from the worker
- Self reported familiarity to the category to be recognized as an integer between 1 and 7
- Can be incorporated in the computation of the crowd aggregated label

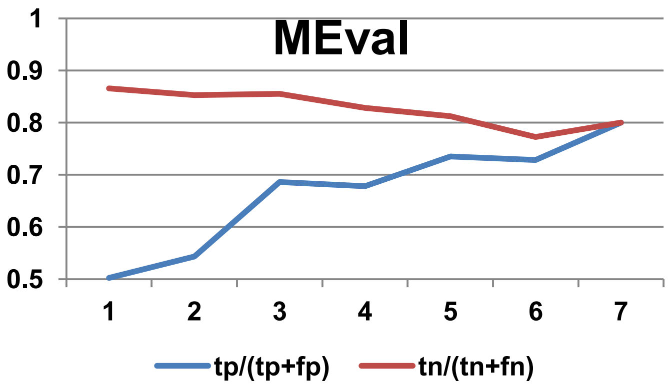
$$norm(x) = (x - 1)/6 \text{ if } x \in \mathbb{N} \text{ and } 0.5$$

$$\check{C}_w = C_w \cdot \breve{n}orm(fam^i_w).$$

Familiarity Correction (FC)



Observation of correlation between the self-reported familiarity to the task and the positive and negative accuracies.



$$\hat{C}_w = \begin{cases} 0.6 & fam_w^i < 3, L_w^i = Yes \\ 0.9 & fam_w^i < 3, L_w^i = No \\ 0.8 & fam_w^i > 3, L_w^i = Yes \\ 0.8 & fam_w^i > 3, L_w^i = No \end{cases}$$

Involvement of Familiarity

- In how many cases does it help when compared to not using it
  - F – just familiarity
  - FC – involving the correction

- Improvement in terms of F1 when compared to the setting without it
  - 7 boosting functions x PN discrimination = total 14 settings

Dataset	Eval	F+	F-	FC+	FC-
MMEval-Label2	hard	6	8	8	6
MMEval-Label3	soft	9	5	10	4
MMSys-Label2	hard	3	11	4	10
MMSys-Label3	soft	9	5	10	4

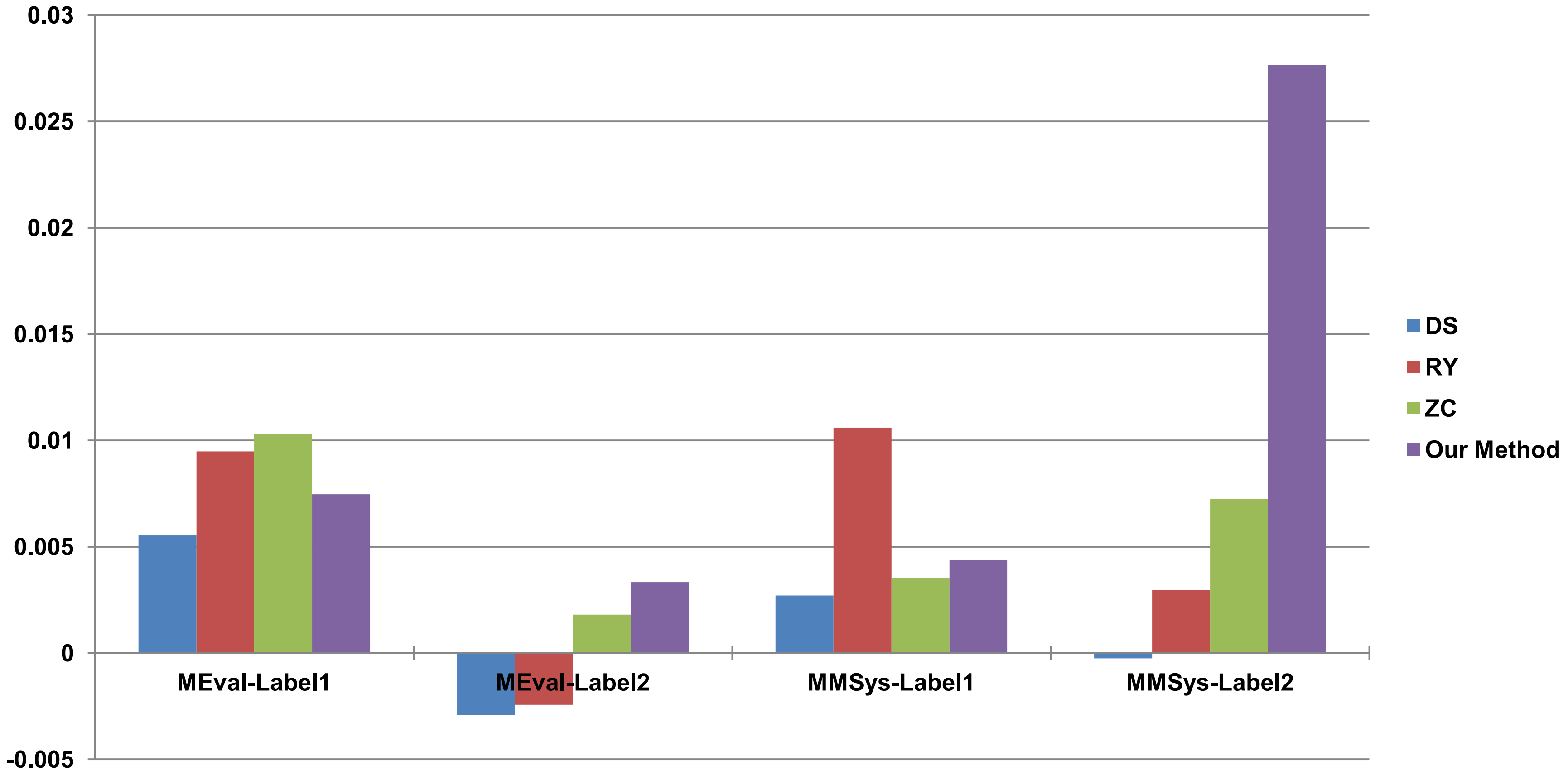
Comparison to other aggregation methods

Compare the performance of our method in terms of F1 improvement when compared to Majority Voting:

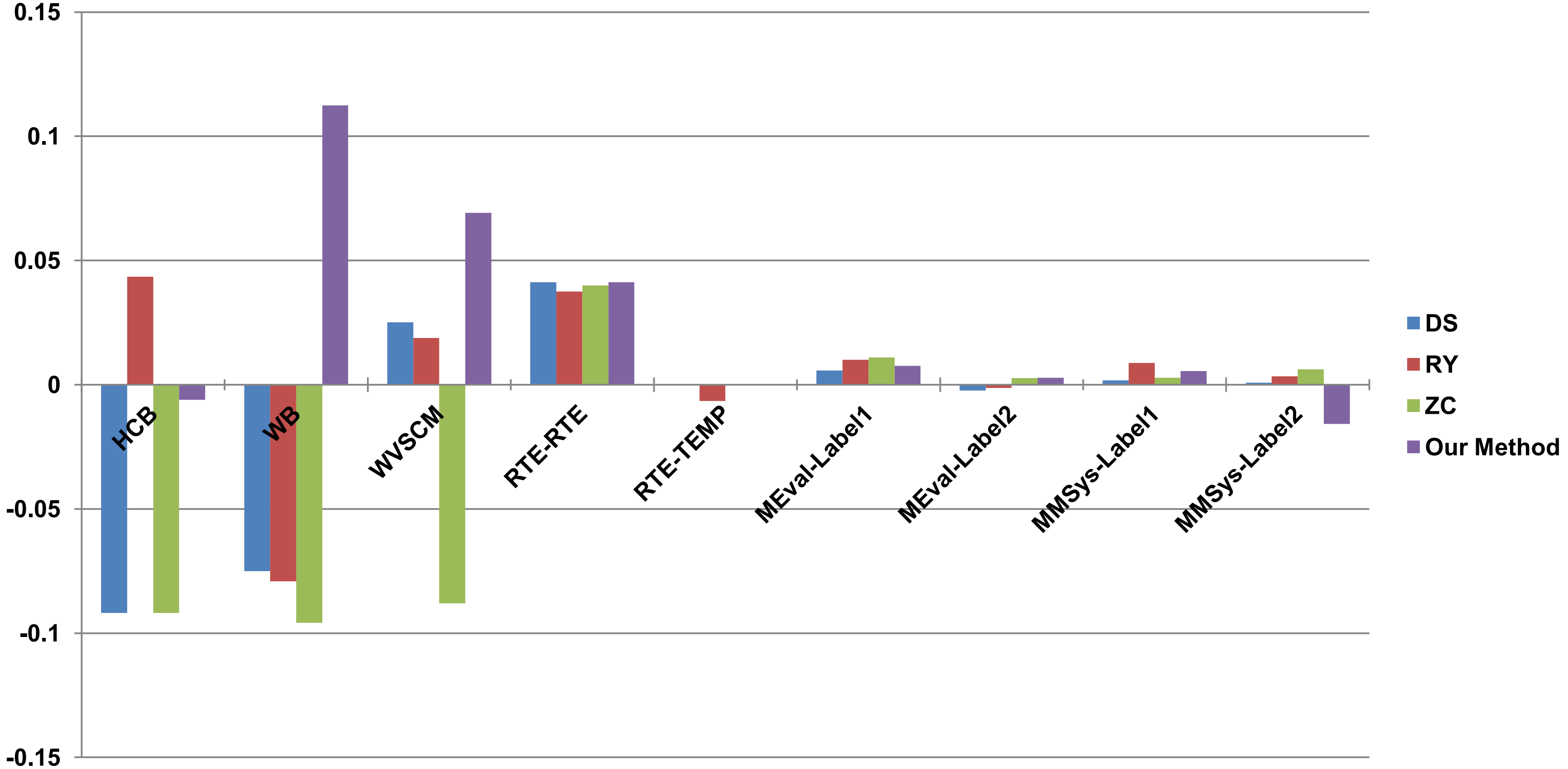
- Dawid-Skene (DS)
  - Probabilistic, confusion matrices and class priors, EM
- Raykar (RY)
  - Bayesian approach and worker priors for each class, bias towards sensitivity or specificity
- ZenCrowd (ZC)
  - Probabistic, workers acting independent of each other and the item’s true class

F1 Measure on general datasets

F1 Measure on Fashion Domain datasets



Accuracy on all datasets



Conclusions & Future Work

- Novel method for the aggregation of crowd labels in order to find the underlying hidden labels, while at the same time estimating the worker quality
- Flexible model based on an EM technique where the computation of the aggregated worker labels is mutually reinforced by the computation of worker confidences
- Extensive experimentation on diverse datasets

- Testing the proposed methods on synthetic data and noise resistance
- Introduce different levels of supervision into the algorithms

THANK YOU!

Q&A