L3S at MediaEval 2013 Crowdsourcing for Social Multimedia Task

Mihai Georgescu, Xiaofei Zhu
L3S Research Center, Leibniz Universität Hannover
Appelstr. 9a
30167 Hanover, Germany
{georgescu,zhu}@l3s.de

ABSTRACT

In this paper we present results of our initial research on aggregating noisy crowdsourced labels, by using a modified version of the EM algorithm introduced in [1]. We propose different methods of estimating the worker confidence, a measure that indicates how well the worker is performing the task, and of integrating it in the computation of the aggregated label. Furthermore, we introduce a novel method of computing the worker confidence by using the soft aggregated labels. In order to assess the effectiveness of our proposed methods, we experiment on the MediaEval 2013 Crowdsourcing for Social Multimedia Task dataset.

1. INTRODUCTION

In this paper we detail the methods proposed for the MediaEval 2013 Crowdsourcing for Social Multimedia Task [5]. The methods in this paper apply the EM method from [1] to infer labels from multiple and possibly noisy labels, assuming that no authoritative ground truth is available, and estimate both the accuracy of the workers and the actual labels using the crowdsourced assessments.

A similar approach was used for building probablistic models [4] to label images using crowdsourcing, for identifying systematic errors done by crowd workers [3], or for crowdsourcing document relevance judgements [2].

In our methods the error-rate is replaced by the worker confidence, used as the weight of a worker contribution in the aggregated label computation. We attempt to improve the standard EM method by using different ways to boost the worker confidence, as well as proposing a novel method for computing it. We introduce the soft evaluation of the worker confidence, where the soft aggregated crowd decision is taken into account instead of the hard aggregated label.

2. APPROACH

In this section we detail the computation of the aggregated decision of a crowd for the label of an instance i, L^i_{crowd} (i.e. *Yes* or *No*), and of the worker confidence. We distinguish between two types of worker confidence depending on whether we make a discrimination between the quality of the positive and negative answers or not. In the case of such a discrimination each worker is characterized by a positive confidence C^+_u and a negative confidence C_{u^-}, otherwise we use a single value for the worker confidence, C^*_u. Majority voting means $C_u = 1$. L^i_{crowd} is computed by aggregating the individual worker labels $L^i_u \in \{Yes, No\}$, ignoring *Not Sure* labels.

In the E step we compute the aggregated crowd labels using Eq. 2 when discriminating between positive and negative and Eq. 1 otherwise, and in the M step we update the worker confidences as defined in Eq. 3 or Eq. 4.

2.1 Aggregated Crowd Labels

In case we do not discriminate between positive and negative answer quality, the probability of an instance being labeled as positive is:

$$p^+_i = \frac{\sum_u C^*_u \cdot I(L^i_u = Yes)}{\sum_u C^*_u \cdot I(L^i_u = Yes) + \sum_u C^*_u \cdot I(L^i_u = No)} \quad (1)$$

In case we differentiate between the positive and negative answer quality this becomes:

$$p^+_i = \frac{\sum_u C^+_u \cdot I(L^i_u = Yes)}{\sum_u C^+_u \cdot I(L^i_u = Yes) + \sum_u C^-_u \cdot I(L^i_u = No)} \quad (2)$$

The probability of an instance being labeled as negative is obviously $p^-_i = 1 - p^+_i$. We will refer to the p^+_i and p^-_i as computed by using either method as **aggregated soft labels**. Moreover, the final **aggregated hard label** assigned by the crowd is given by comparing the difference between the positive probability and the negative one:

$$L^i_{crowd} = \begin{cases} Yes, & p^+_i - p^-_i \geq 0 \\ No, & p^+_i - p^-_i < 0 \end{cases}$$

2.2 Worker Confidence Computation

The undiscriminative confidence in a worker is defined as:

$$C^*_u = \frac{tp_u + tn_u}{tp_u + tn_u + fp_u + fn_u} \quad (3)$$

In case we discriminate between the quality of positive and negative answers we use two types of confidence:

$$C^+_u = \frac{tp_u}{tp_u + fp_u}; C^-_u = \frac{tn_u}{tn_u + fn_u} \quad (4)$$

We distinguish between two types of evaluation of the worker confidence: hard evaluation, where we use only the final, *aggregated hard labels*, and a soft evaluation, where we use the *aggregated soft labels*.

In case of a hard evaluation of the performance of a user we use the following definitions:

$$tp_u = \sum_i I(L^i_u = Yes) \cdot I(L^i_{crowd} = Yes)$$

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

$$tn_u = \sum_i I(L_u^i = No) \cdot I(L_{crowd}^i = No)$$

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

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

In the case of a soft evaluation of the worker confidence we use the following definitions:

$$tp_u = \sum_i I(L_u^i = Yes) \cdot p_i^+; tn_u = \sum_i I(L_u^i = No) \cdot p_i^-$$

$$fp_u = \sum_i I(L_u^i = Yes) \cdot p_i^-; fn_u = \sum_i I(L_u^i = No) \cdot p_i^+$$

2.3 Worker Confidence Correction

Furthermore we can apply the following corrections to the confidence when aggregating the multiple votes: boosting the confidence ($\hat{C} = boost(\check{C}_u)$) or involving the worker self-reported familiarity with the category for which Label 2 is assigned to the image (fam_u^i) in the computation of the confidence ($\check{C}_u = C_u \cdot norm(fam_u^i)$). Based on an observation of a correlation of the familiarity and the type of answers and their accuracy, we can also use a familiarity correction strategy

$$\hat{C}_u = \begin{cases} 0.6 & fam_u^i < 3, I_u = Yes \\ 0.9 & fam_u^i < 3, I_u = No \\ 0.8 & fam_u^i > 3, I_u = Yes \\ 0.8 & fam_u^i > 3, I_u = No \end{cases}$$

The boosting function $boost(x)$ can be e^x or $x^p; p \in \mathbb{R}$.

The transformation of familiarity from an integer within 1 and 7 or missing to a real subunitary positive number, is done by the $norm(x)$ function. $norm(x) = (x-1)/6$ if $x \in \mathbb{N}$ and 0.5 if missing.

2.4 Method Settings

The computation of the labels in the EM algorithm as well as of the final decisions after the iterations are finished depend on the following settings:

- the use of positive/negative answer discrimination
- the evaluation of worker confidences using soft labels
- the boosting type
- the use of familiarity in the computation
- the use of the familiarity correction

For picking candidates for the submitted runs, and finding the best setting, we evaluated the performance of our methods on the MMSys 2013 Dataset. The selected settings that are used for the submitted runs are detailed in Table1.

The first two runs use the discrimination between the positive and negative worker confidence. *Run1* is using the EM algorithm with hard iterations for both labels. *Run2* represents the EM algorithm using the soft iterations for both labels without any special boosting strategy or involving the familiarity.

3. RESULTS

The performance of each submission in terms of the F1-measure is presented in Table 2. As already mentioned in experiments carried out with the MMSys 2013 dataset, we notice a better performance in the case of the second label. We can see that for the first label the best performance is achieved in *Run5*, and for the second label by *Run3*. We notice that in the case of Label 1, discriminating between positive and negative label quality provides a performance increase, while in the case of Label 2 the effect is opposite.

R	L	EM decision					Final decision			F1
		S	PN	B	F	FC	B	F	FC	
1	1	-	✓	x^1	-	-	x^1	-	-	0.895
	2	-	✓	x^1	-	-	x^1	-	-	0.909
2	1	✓	✓	x^1	-	-	x^1	-	-	0.894
	2	✓	✓	x^1	-	-	x^1	-	-	0.911
3	1	-	-	$x^{0.5}$	✓	-	x^{20}	-	-	0.900
	2	-	-	x^2	✓	✓	x^2	-	-	0.913
4	1	✓	✓	x^3	✓	✓	x^1	-	-	0.898
	2	-	-	x^2	✓	-	x^2	-	-	0.913
5	1	✓	✓	e^x	-	-	e^x	-	-	0.894
	2	✓	✓	x^2	✓	✓	x^2	-	-	0.913

Table 1: Setting for each submission run (R) and label(L), depicted in terms of: using soft labels in the worker confidence calculation(S), discrimination between positive and negative answer quality(PN), boosting type (B), using familiarity in the computation (F), familiarity correction (FC), in the computation of the decision during the EM iterations as well as in the final decision, along with the F1 measure achieved on the MMSys 2013 dataset

Submission	Label1	Label2
Run1	0.7328	0.7533
Run2	0.7340	0.7412
Run3	0.7264	**0.7592**
Run4	0.7263	0.7391
Run5	**0.7346**	0.7371

Table 2: Performance of each submission

Acknowledgments

This work was partially funded by the European Commission FP7 under grant agreements No. 287704 for the CUbRIK project.

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] M. Hosseini, I. J. Cox, N. Milić-Frayling, G. Kazai, and V. Vinay. On aggregating labels from multiple crowd workers to infer relevance of documents. In *Advances in Information Retrieval*, pages 182–194. Springer, 2012.

[3] P. G. 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.

[4] G. Kasneci, J. V. Gael, D. Stern, and T. Graepel. CoBayes: bayesian knowledge corroboration with assessors of unknown areas of expertise. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 465–474. ACM, 2011.

[5] B. Loni, M. Larson, A. Bozzon, and L. Gottlieb. Crowdsourcing for social multimedia at MediaEval 2013: Challenges, data set, and evaluation. In *MediaEval 2013 Workshop, October 18-19, 2013, Barcelona, Spain*, 2013.