Impact of noisy annotators' reliability in a crowdsourcing system performance

Margarita Cabrera-Bean, Carles Díaz-Vilor, Josep Vidal
Dept. of Signal Theory and Communications,Universitat Politecnica de Catalunya (UPC)
Barcelona, Spain
{marga.cabrera,josep.vidal}@upc.edu, carles.diaz.vilor@alu-etsetb.upc.edu

***Abstract* — Crowdsourcing is a powerful tool to harness citizen assessments in some complex decision tasks. When multiple annotators provide their individual labels a more reliable collective decision is obtained if the individual reliability parameters are incorporated in the decision making procedure. The well-known Maximum A Posteriori (MAP) rule weights the individual labels in proportion to the annotators' reliability. In this work we analyze how the crowdsourcing system performance is degraded with the use of noisy annotators' reliability parameters and we derive an alternative MAP based rule to be applied when these parameters are neither known nor even estimated by the decision system. We also derive analytical expected error rates and their upper bounds obtained by each rule as a useful tool to estimate the number of necessary annotators in the collective decision system depending on the level of noise present in the estimated reliability parameters.**

***Index Terms*—Crowdsource, Expected error rate bound, Specificity, Sensitivity.**

I. Introduction

The diagnose, detection or process of labeling an object by soliciting contributions from a large group of people and especially from the online community has become an extended practice in a great variety of applications. Crowdsourcing (so it is called) is nowadays an effective tool to perform human aided tasks where computerized machines fail or whenever there is lack of expert people to analyze a set of data and extract concluding decisions. A crowdsourcing labelling process can be briefly described in the following terms: each person from a non-expert group of people (annotators) has to assign a label to a common sample, accessible from the Internet and a central server has to process the labels from all the annotators in order to infer the true label of the item under quality criteria.

In a binary crowdsourcing system, the reliability of each annotator is measured as a weighted sum of both, sensitivity (proportion of positives that are correctly identified) and specificity (proportion of negatives that are correctly identified).

Crowdsourcing initiatives span over different fields as for instance astronomy [1], biology [2], social support for disaster relief [3] and medicine applications just to cite some of the available initiatives. One of the most important applications of crowdsourcing is telediagnosis in some diseases when the presence of medical experts in the local area is highly sparse. Malaria diagnosis is a significant binary decision example where the available test results exhibit much poorer sensitivity than specificity. Although there are many proposals i.e [12], to solve the problem applying automated algorithms, crowdsourcing can improve the malaria diagnosis accuracy in some critical contexts. So far two malaria-diagnosis gaming platforms have been developed and are on-line available to be used by volunteers. The MalariaSpot.org platform [13] provides a system to identify malaria parasites in Red Blood Cells (RBC) shown in images of thick blood films and the BioGames platform [14], [15] is a hybrid (automated and crowdsourced) system to diagnose images of RBC as infected (Positive) or uninfected (Negative).

Despite the annotators may be non-qualified experts on the task they are assigned, the combination of decisions frequently provides reliable answers and highly accurate tests. Even a simple algorithm such as Majority Voting (MV) provides good enough performance if the number of annotators is sufficiently high [4]. Hence it is crucial to decide how many are needed for a certain task and given system accuracy.

Different strategies [5][6][7][8] have been designed and tested to online estimate the accuracy of a crowdsourced decision and to define termination strategies to decide if a new annotator is needed to improve the accuracy of a question or the current accuracy is good enough. In [5] a crowdsourcing engine is presented where the fusion center employs a prediction model to estimate how many annotators are required to achieve a specific accuracy among the total number of cases examined.

A similar procedure to on-line estimate the accuracy in the answer to a query is developed in [6], where authors design and implement a cost sensitive method for crowdsourcing. They define and online estimate the profit of the crowdsourcing job as a function of the value of the question, the risk of getting incorrect answers, and the cost of workers in the crowdsourcing system. A question is terminated in real time if the marginal expected profit of obtaining more answers is not positive.

To manage the problem of characterizing the annotators with individual reliability parameters, a confusion matrix was associated to each annotator in [9]. In this matrix each row represents a discrete conditional distribution of the decision of

This work is supported by the "Ministerio de Economía y Competitividad" of the Spanish Government and ERDF funds (TEC2013-41315-R DISNET), the Catalan Government (2014 SGR 60 AGAUR) and Eusipco'11 organization.

978-0-9928-6265-7/16/$31.00 ©2016 IEEE

an annotator given ground truth. The model is usually called the Dawid-Skene model by the crowdsourcing community. In [9] the expectation-maximization (EM) algorithm is proposed to jointly infer the confusion matrices, the prior probabilities and the unobserved true labels from the annotators' answers.

In [10] the problem was generalized assuming that the ground truth is generated by a logistic regression model and the annotators' reliability parameters follow a beta distribution. The proposed EM algorithm iteratively estimates the labels and measures the annotators' parameters conditioned to the ground truth. In [11] the concept of a general decomposable aggregation rule is defined and different criteria as MV and MAP are presented as particular cases of a general aggregation decision rule under the general Dawid-Skene crowdsourcing model. After this definition, a set of upper bounds on the error rate is developed.

In this work we focus on analyzing a crowdsourced binary decision system under realistic conditions as a function of the annotators' reliability parameters (sensitivity and specificity). We stress the influence of these parameters on the mean error rate of the crowdsourcing decision system when the optimal decision criterion MAP is applied. Taking into account the statistical nature of the annotators' reliability parameters we distinguish two different approaches. A first approach is obtained when these parameters are known (MAP rule) or are estimated by the system (Noisy MAP, or NMAP procedure). Consequently, a second approach consists in modeling the reliability parameters as random variables following a predetermined distribution, (Averaged MAP, or AMAP). Hence, our main contributions are:

1. The analytic derivation of an upper bound for the expected error rate (EER) when the NMAP procedure is applied.
2. The analytic development of the AMAP criteria and the corresponding analytical upper bound for the mean error rate.
3. A comparison of the NMAP, MAP and AMAP performances.

Thus, from the results of this work we can obtain the minimum number of users required to ensure a maximum EER, thereby providing a way to decide, depending on the uncertainty level of the noisy annotators' reliability, in which conditions the optimal MAP rule outperforms the suboptimal AMAP. As the evaluation becomes prohibitive for a relatively low number of users, i.e. more than fifteen, the upper bounds make the evaluation viable as the number of annotators increases.

The rest of this document is organized as follows. In section III, III.A is dedicated to give the expected error rate and an upper bound for it when the MAP rule is applied. In III. B and III.C we develop the same indicators (expected error rate and upper bound) for both the NMAP and the AMAP rules, respectively. Section IV is devoted to present computerized graphical results and conclusions are summarized in section V.

II. Mathematical framework

A binary decision crowdsourcing system has to process the decision variables from N annotators and decide if an event has occurred, ω=1, or otherwise ω=0. In this process the annotator n delivers its decision in the binary variable $x(n)$ and is characterized by means of his reliability parameters

$$r_n^{(k)} = \Pr\{x(n)=k|\omega=k\} \quad n\in\{1,..,N\} \quad k\in\{0,1\}\,. \tag{1}$$

Thus the random variable $x(n)$ is Bernoulli distributed, i.e,

$$\Pr\{x(n)|k\} = \left(r_n^{(k)}\right)^{I(x(n),k)}\left(1-r_n^{(k)}\right)^{\left(1-I(x(n),k)\right)}, \tag{2}$$

where the Boolean function is defined as $I(a,b)=1$ if $a=b$ and 0 otherwise.

For the rest of this document the observed vector contains the binary answers delivered by the N annotators, i.e. $\mathbf{x} = [x(1), x(2), .. x(N)\ ]^T$ where as usual superindex T denotes transpose vector. The parameter P_k denotes the prior probability of label k, i.e. $P_k = \Pr\{\omega=k\},\ k\in\{0,1\}$.

III. MAP based rules

The well-known MAP rule by applying Bayes outcomes

$$\begin{aligned}\hat{\omega}_{MAP} &= \arg\max_{k\in\{0,1\}} \Pr(\omega=k|\mathbf{x}) = \arg\max_{k\in\{0,1\}} s_k \\ s_k(\mathbf{x}) &\doteq \log\Pr\{\mathbf{x}|\omega=k\} + \log P_k\end{aligned} \tag{3}$$

where $s_k(\mathbf{x})$ represents the score function for the potential class k. Assuming independent annotators, the score function is

$$s_k(\mathbf{x}) = \sum_{n=1}^{N} I(x(n),k)\log r_n^{(k)} + \left(1-I(x(n),k)\right)\log\left(1-r_n^{(k)}\right) + \log P_k$$

i.e. the decision vector $\mathbf{x}$ is processed to label $\hat{\omega}_{MAP} = k$ if $s_k(\mathbf{x}) > s_{\bar{k}}(\mathbf{x})$, where $\bar{k}=0$ if $k=1$ and vice versa.

A. MAP rule and Expected Error Rate Upper Bound

The EER under the MAP rule is given in (4) where the condition k is used to denote $\omega=k$, simplifying thereby the nomenclature:

$$P_{e-MAP} = \sum_{k=0}^{1} P_k \Pr\{\hat{\omega}_{MAP} = \bar{k}|k\} \tag{4}$$

The decision region $\Re_k$ is the set of observations $\{\mathbf{x}|s_k(\mathbf{x}) > s_{\bar{k}}(\mathbf{x})\}$ that determines the conditioned error probability in (4) for $k\in\{0,1\}$ as

$$\Pr\{\hat{\omega}_{MAP} = \bar{k}|k\} = \sum_{\mathbf{x}\in\Re_{\bar{k}}} \prod_{n=1}^{N} \left(r_n^{(k)}\right)^{I(x(n),k)}\left(1-r_n^{(k)}\right)^{\left(1-I(x(n),k)\right)} \tag{5}$$

Consequently the EER can be computed as a function of the annotators' reliability parameters $r_n^{(k)},\ k\in\{0,1\}\ n\in\{1,..N\}$, applying (4) and (5). For large N an upper bound is commonly used to give an at-worst value of the EER when applying the MAP criteria. The Hoeffding inequality given in the appendix is next obtained for the MAP case. Let us write the EER as

$$P_{e-MAP} = \sum_{k=0}^{1} P_k \Pr\left\{ s_{\bar{k}}(\mathbf{x}) > s_k(\mathbf{x}) \middle| k \right\} \tag{6}$$

and check that the conditioned error probability $\Pr\{s_{\bar{k}}(\mathbf{x}) > s_k(\mathbf{x})|k\}$ is directly related to the probability that the sum of N independent random variables be highly concentrated to its expected value. To show this property let's define the N random variables $\sigma_{\bar{k}k}(x(n))$ and their sum $\Sigma_{\bar{k}k}(\mathbf{x})$ as

$$\begin{aligned} \sigma_{\bar{k}k}(x(n)) &\doteq I(x_n,k)\log\left(\tfrac{1-r_n^{(\bar{k})}}{r_n^{(k)}}\right) + \left(1-I(x_n,k)\right)\log\left(\tfrac{r_n^{(\bar{k})}}{1-r_n^{(k)}}\right), \\ \Sigma_{\bar{k}k}(\mathbf{x}) &\doteq \sum_{n=1}^{N} \sigma_{\bar{k}k}(x(n)), \end{aligned} \tag{7}$$

and the parameter $M_{\bar{k}k}$ as the conditioned expected value

$$M_{\bar{k}k} \doteq \mathbb{E}\left\{ \Sigma_{\bar{k}k}(\mathbf{x}) \middle| k \right\} = -\sum_{n=1}^{N} \mathrm{D}\left(r_n^{(k)} \middle\| 1 - r_n^{(\bar{k})} \right), \tag{8}$$

written in terms of $\mathrm{D}(A\|B)$, the Kullback-Leibler divergence of two Bernoulli probability distributions of parameters A and B respectively. At that point the conditioned probability in (6) can be described as

$$\Pr\left\{ s_{\bar{k}}(\mathbf{x}) > s_k(\mathbf{x}) \middle| k \right\} = \Pr\left\{ \Sigma_{\bar{k}k}(\mathbf{x}) - M_{\bar{k}k} > \log\frac{P_k}{P_{\bar{k}}} - M_{\bar{k}k} \right\} \tag{9}$$

Moreover, we assume the realistic conditions of reliable enough annotators, i.e., $r_n^{(k)} > 1 - r_n^{(\bar{k})}$ and $r_n^{(\bar{k})} > 1 - r_n^{(k)}$, and we define the dispersion vector $\boldsymbol{\gamma}$ as $\boldsymbol{\gamma} \doteq [\gamma_1, \gamma_2 \ldots. \gamma_N]^T$ where

$$\gamma_n \doteq \max\left(\sigma_{\bar{k}k}(x(n))\right) - \min\left(\sigma_{\bar{k}k}(x(n))\right) = \log\left(\tfrac{r_n^{(\bar{k})}}{1-r_n^{(k)}} \tfrac{r_n^{(k)}}{1-r_n^{(\bar{k})}} \right). \tag{10}$$

Note that the better the annotator reliability the higher γ_n is. By applying the Hoeffding inequality to (9) and provided that $\log(P_k/P_{\bar{k}}) - M_{\bar{k}k} > 0$ which is frequently true in realistic conditions, we obtain that the expected error rate given in (6) is upper bounded as

$$\begin{aligned} &P_{e-MAP} \leq B_{MAP} \\ &B_{MAP} = \sum_{k=0}^{1} P_k \exp\left(-\frac{2}{|\boldsymbol{\gamma}|^2} \left(\log\frac{P_k}{P_{\bar{k}}} + \sum_{n=1}^{N} \mathrm{D}\left(r_n^{(k)} \middle\| 1 - r_n^{(\bar{k})} \right) \right)^2 \right) \end{aligned} \tag{11}$$

Thus, we have two options to evaluate the MAP rule performance: to compute the EER given in (4) and (5) or to calculate the bound in (11). The former option has two main drawbacks. On the one hand the number of involved flops increases exponentially with the number of annotators N since the EER depends on the entire space of the processed vector $\mathbf{x}$ whose size is 2^N. Furthermore, the products of N factors involved in (5) requires a prohibitive storage memory. On the other hand there is no an easy and efficient way to recursively compute the EER for N+1 annotators departing from the EER obtained for N annotators. These issues are avoided when the upper bound is computed: the number of required flops and local storage memory increases linearly with the number of annotators, and both the numerator and the denominator of the exponents in (11) can be easily converted to a recursive function with the number of annotators.

Analogous remarks could be stated from the NMAP and the AMAP rules in the following subsections, however we will not mention them to avoid repetition.

So, in view of these properties, the results given in section IV are based on the exact EER when the number of annotators N is lower than 15. As it is shown in table I, with 15 annotators the number of flops is maintained under e+6. Otherwise we have opted for computing the EER obtained by simulations and the upper bound.

TABLE I. MAP EER AND UB, NUMBER OF FLOPS

Flops	Number of users		
	N	*15*	*16*
EER (4)	$2^{N+1}.N$	983040	2097152
UP (11)	6N	90	96

B. Noisy MAP rule and Expected Error Rate Upper Bound

Let's assume in this section than the MAP rule is applied after a training process where the reliability of annotators is estimated. This is a common practice: when an annotator first accesses to a crowdsourcing system, he initially labels a set of objects whose ground truth labels are available at the system. This set of objects is used to determine an estimated sensitivity $\hat{r}_n^{(1)}$ and an estimated specificity $\hat{r}_n^{(0)}$ to each new annotator. Therefore, the MAP rule is realistically applied using $\hat{r}_n^{(0)}, \hat{r}_n^{(1)}$, $n \in \{1, .. N\}$ instead of the actual values for the reliability parameters.

Let us name $\hat{s}_k(\mathbf{x})$ to the new score functions obtained if the MAP rule is applied using the estimated reliability parameters instead of the true ones. Let us call NMAP (Noisy MAP) to the resultant decision rule in this case. Analogously let's define $\hat{\Sigma}_{\bar{k}k}(\mathbf{x})$, $\hat{\sigma}_{\bar{k}k}(x(n))$, $\hat{\gamma}_n$ and $\hat{\boldsymbol{\gamma}}$ as in (7) but depending on the estimated $\hat{r}_n^{(k)}$ instead of the actual values $r_n^{(k)}$. Thus, the EER in the case of NMAP is found as

$$P_{e-NMAP} = \sum_{k=0}^{1} P_k \Pr\left\{ \hat{s}_{\bar{k}}(\mathbf{x}) > \hat{s}_k(\mathbf{x}) \middle| k \right\}, \tag{12}$$

and its upper bound derived using the expected value

$$\begin{aligned} \hat{M}_{\bar{k}k} &\doteq \mathbb{E}\left\{ \hat{\Sigma}_{\bar{k}k}(\mathbf{x}) \middle| k \right\} = \\ &= \sum_{n=1}^{N} \left(r_n^{(k)} \log\left(\tfrac{1-\hat{r}_n^{(\bar{k})}}{\hat{r}_n^{(k)}} \right) + \left(1 - r_n^{(k)}\right) \log\left(\tfrac{\hat{r}_n^{(\bar{k})}}{1-\hat{r}_n^{(k)}} \right) \right) \end{aligned}$$

Finally provided that $\log(P_k/P_{\bar{k}}) - \widehat{M}_{\bar{k}k} > 0$ and using the

Hoeffding inequality:

$$P_{e-NMAP} \le B_{NMAP},$$
$$B_{NMAP} = \sum_{k=0}^{1} P_k \exp\left(-\frac{2}{|\boldsymbol{\gamma}|^2} \left(\log \frac{P_k}{P_{\bar{k}}} - \hat{M}_{\bar{k}k} \right)^2 \right) \tag{13}$$

The upper bound in (13) depends on both, the actual reliability parameters $r_n^{(k)}$ and the correspondent estimated $\hat{r}_n^{(k)}$or noisy reliability parameters.

C. Averaged MAP rule and EER Upper Bound

In this section we propose a MAP rule to be applied when at the crowd-computing centre the annotators' reliabilities are neither available nor even estimated from their answers. We adopt a Bayesian strategy by assuming that these parameters are independent random variables equally beta distributed for all the annotators. Let's assume that the beta distribution means $r^{(k)}$, $k\epsilon\{0,1\}$ are the only available parameters. The MAP rule can then be applied by averaging the random Maximum a Posteriori distribution which results in a different criteria from the MAP one given in (3). We call averaged MAP or directly AMAP to the obtained rule.

$$\hat{\omega}_{AMAP} = \arg\max_{k \in \{0,1\}} \mathbb{E}_{r_n^{(k)}} \left\{ \Pr\left\{ \mathbf{x} \middle| k, r_n^{(k)}{}_{n \in \{1,..,N\}} \right\} P_k \right\} \tag{14}$$

The expectation operator in (14) is obtained with respect to the random variables set $r_n^{(k)}$ $k \in \{0,1\}$ $n \in \{1,..N\}$, which, considering the annotators' independency, generates the score functions

$$\tilde{s}_k(\mathbf{x}) = \prod_{n=1}^{N} \mathbb{E}_{r_n^{(k)}} \left\{ \Pr\left\{ x_n \middle| k, r_n^{(k)} \right\} \right\} P_k \tag{15}$$

Each averaged k-conditioned distribution in (15) is also a Bernoulli distribution parameterized by the respective mean $r^{(k)}$, given that the beta distribution is conjugate prior for the Bernoulli distribution.

$$\mathbb{E}_{r_n^{(k)}} \left\{ \Pr\left\{ x_n \middle| k, r_n^{(k)} \right\} \right\} = \left(r^{(k)} \right)^{I(x_n,k)} \left(1 - r^{(k)} \right)^{1-I(x_n,k)}, \tag{16}$$

So, the AMAP rule is equivalent to apply the MAP criteria using $r^{(k)}$, $k \in \{0,1\}$ as the reliability parameters for all the annotators. If $P_k = 0.5$ the AMAP rule becomes the well-known majority voting rule. As in previous sections the expected error rate is derived from the new score functions as

$$P_{e-AMAP} = \sum_{k=0}^{1} P_k \Pr\left\{ \tilde{s}_{\bar{k}}(\mathbf{x}) > \tilde{s}_k(\mathbf{x}) \middle| k \right\}, \tag{17}$$

and we adopt the notation items $\tilde{\Sigma}_{\bar{k}k}(\mathbf{x})$, $\tilde{\sigma}_{\bar{k}k}(x(n))$, $\tilde{\gamma}$ and $\tilde{\boldsymbol{\gamma}}$ defined as in (7) as functions of $r^{(k)}$ instead of the actual reliability values $r_n^{(k)}$. The expected value for the random sum is computed as

$$\tilde{M}_{\bar{k}k} \doteq \mathbb{E}\left\{ \tilde{\Sigma}_{\bar{k}k}(x(n)) \middle| k \right\} = -N\mathrm{D}\left(r^{(\bar{k})} \middle\| 1 - r^{(\bar{k})} \right) \tag{18}$$

Finally, by applying the Hoeffding inequality and assuming $r^{(k)} > 1 - r^{(k)}$ for $k \in \{0,1\}$ and $\log(P_k/P_{\bar{k}}) - \tilde{M}_{\bar{k}k} > 0$ the upper bound is obtained as a function of the reliability means $r^{(k)}$, $k \in \{0,1\}$.

$$P_{e-AMAP} \le B_{AMAP}$$
$$B_{AMAP} = \sum_{k=0}^{1} P_k \exp\left(-\frac{2}{N\tilde{\gamma}^2} \left(\log \frac{P_k}{P_{\bar{k}}} - \tilde{M}_{\bar{k}k} \right)^2 \right) \tag{19}$$

IV. PERFORMANCE EVALUATION

In this section, numerical results obtained by applying Montecarlo simulations are presented to compare the performance of the proposed detection rules (MAP, NMAP and AMAP). The NMAP rule is applied for different levels of annotators' reliability uncertainty and its performance is measured in terms of the EER as a function of N, i.e. the number of annotators labeling an object.

We generate the parameters $r_n^{(k)}$ for $n\epsilon\{1,..,N\}$, sampling a beta distribution determined by the mean m_k and the variance s_k for $k\epsilon\{0,1\}$. In order to generate consistent results regarding the random nature of the reliability parameters, all of them are averaged over 400 random trials and have been obtained with a prior probability $P_k = 0.5$ for $k\epsilon\{0,1\}$.

The uncertainty has been modeled through the beta-distributed random variable $e_n^{(k)}$ of variance s_e and biased to have zero mean, i.e. the NMAP rule in (12) and (13) uses the noisy estimates given by

$$\hat{r}_n{}^{(k)} = r_n{}^{(k)} + e_n{}^{(k)}; \quad k \in \{0,1\}$$
$$SNR_k = 10\log_{10}\left(\frac{m_k^2 + s_k}{s_e} \right) \tag{20}$$

Figure 1 shows the analytical EER as a function of the number of annotators and the Signal to Noise Rate SNR measured as power quotient given in (20). We can verify that the MAP rule outperforms the AMAP rule depending on the level of power noise.

In figure 2 we report the upper bounds for the analyzed rules when the number of annotators is higher than 20. We also include the expected error rate obtained by simulation (simulated EER). In this case for each one of the 400 generated sets of reliability parameters we generate 200000 decision random vectors following the Bernoulli distributions, so each point of the simulated EER has been obtained by averaging 8e+7 points. As can be concluded from figure 2, when the number of annotators is higher than 15, similar conclusions can be obtained from the upper bounds and from the simulated EERs. The advantage of using the MAP rule is drastically degraded when the reliability parameters have been noisily estimated. Although the provided upper bounds are not exceptionally tight to the EER, they are very useful to decide (depending on annotators reliability) when it is better to apply the AMAP (equivalent to majority voting rule for the simulated case of prior probability $P_k = 0.5$ for $k\epsilon\{0,1\}$), than the optimum MAP rule.

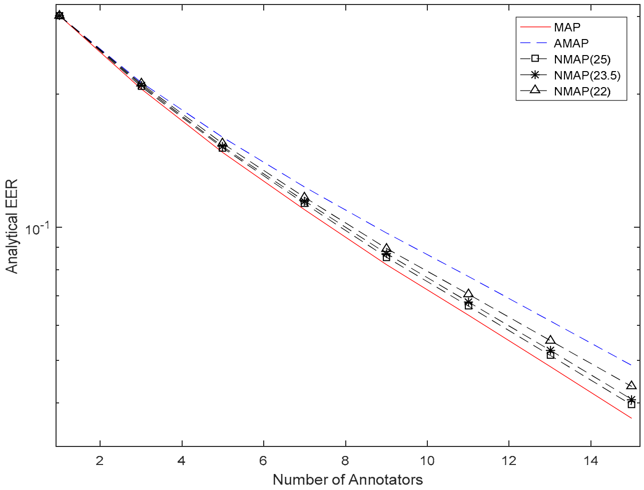

Fig. 1. Analytical EER as a function of the number of annotators *N*. Reliability parameters mean and standard deviation $m_k = 0.7$, s_k=0.01 for $k\epsilon\{0,1\}$. Cases NMAP have been obtained for SNR$_k$=25, 23.5 and 22 dB for $k\epsilon\{0,1\}$.

V. Conclusions

We have derived a MAP based rule (AMAP) to be applied in a crowdsourcing system when the annotators' reliability parameters are neither known nor even estimated. We have compared the AMAP analytical expected error rate and its upper bound to the one provided by the MAP rule when the annotators' reliabilities are noisily estimated (NMAP) and we have numerically analyzed how the lack of accuracy of the annotators' reliability impacts on the performance. The results can be used to decide how many annotators are necessary to obtain a determined quality of the system measured in terms of the expected error rate. The work has been applied to a binary decision system, taking as a reference a medical diagnosis system but it can be easily generalized to a multiple answer crowdsourcing system.

VI. Appendix

The Hoeffding concentration inequality provides an upper bound on the probability that the sum of random variables deviates from its expected value. Let $\sigma_1, \sigma_2, \dots, \sigma_N$ be independent random variables and σ_i strictly bounded by the interval $[a_i, b_i]$. Provided the real parameter $t \geq 0$, the random sum $\Sigma = \sum_{i=1}^{N} \sigma_i$ accomplishes

$$\Pr\{\Sigma - \mathbb{E}\{\Sigma\} \geq t\} \leq \exp\left(-2t^2 \Big/ \sum_{n=1}^{N} (b_i - a_i)^2 \right).$$

References

[1] C. C. M. Kyba, J. M. Wagner, H. U. Kuechly, C. E. Walker, C. D. Elvidge, F. Falchi, T. Ruhtz, J. Fischer, F. Hölker. "Citizen Science Provides Valuable Data for Monitoring Global Night Sky Luminance". Scientific Reports 3, Article number: 1835 (2013) doi:10.1038/srep01835.

[2] J. Deng, J Krause, L. Fei-Fei. "Fine-grained crowdsourcing for fine-grained recognition". In: CVPR (2013)

[3] Z. Matthew; M. Graham; T. Shelton; S. Gorman (2010) "Volunteered Geographic Information and Crowdsourcing Disaster Relief: A Case Study of the Haitian Earthquake," World Medical & Health Policy Vol. 2: Iss. 2, Article 2. DOI: 10.2202/1948-4682.1069

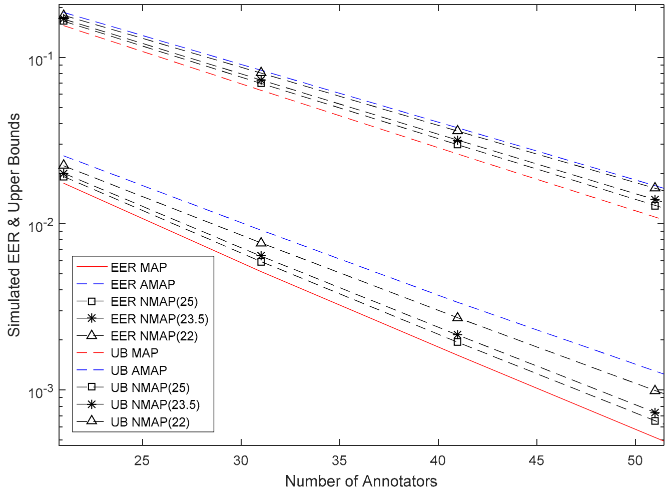

Fig. 2. Upper lines correspond to Upper bound (UB) and lowers lines to simulated expected error rate (EER) as a function of the number of annotators *N*. The reliability parameters have been generated as in figure 1.

[4] R. Snow, B. O'Connor, D. Jurafsky, A. Ng. "Cheap and Fast — But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks". Proceeding of the Conference on Empirical Natural Language Processing, pp 254-263, 2008.

[5] X. Liu, M. Lu, B. C. Ooi, Y. Shen, S. Wu, M. Zhang, "CDAS: A Crowdsourcing Data Analytics System", *Proceedings of the VLDB Endowment, 2012, Vol. 5, No. 10 Copyright 2012 VLDB Endowment* 21508097.

[6] J. Gao, X. Liu, B. C. Ooi, H. Wang, G. Chen , "An Online Cost Sensitive Decision-Making Method in Crowdsourcing Systems", SIGMOD'13, June 22-27, 2013, New York, New York, USA. Copyright 2013 ACM 978-1-4503-2037-5/13/06

[7] Kamar E., Hacker S., Horvitz E. "Combining Human and Machine Intelligence in Large-scale Crowdsourcing". Proceedings of the 11th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2012), 4-8 June 2012, Valencia, Spain.

[8] W. Tang, M. Lease, "Semi-Supervised Consensus Labeling for Crowdsourcing". SIGIR 2011 Workshop on Crowdsourcing for Information Retrieval, July 2011, Beijing, China.

[9] A. P. Dawid, A. M. Skene, "Maximum Likelihood Estimation of Observer Error-Rates Using the EM Algorithm". Journal of the Royal Statistical Society. Series C (Applied Statistics), Vol. 28, No. 1 (1979), pp. 20-28

[10] V. Raykar, S. Yu, L. H. Zhao, C. Florin, L. Bogoni, and L. Moy. "Learning From Crowds". Journal of Machine Learning Research 11, 2010, pp. 1297-1322.

[11] Li, Hongwei, and Bin Yu. "Error Rate Bounds and Iterative Weighted Majority Voting for Crowdsourcing." arXiv preprint arXiv:1411.4086 (2014).

[12] J. Frean. "Reliable enumeration of malaria parasites in thick blood films using digital image analysis". Malaria Journal 2009; http://www.malariajournal.com/content/8/1/218, doi:10.1186/1475-2875-8-218.

[13] M. A. Luengo-Oroz, A. Arranz, J. Frean, "Crowdsourcing Malaria Parasite Quantification: An Online Game for Analyzing Images of Infected Thick Blood Smears", J Med Internet Res 2012;14(6):e167, URL: http://www.jmir.org/2012/6/e167, DOI: 10.2196/jmir.2338, PMID: 23196001, PMCID: PMC3510720.

[14] S. Mavandadi, S. Dimitrov, S. Feng, F. Yu, U. Sikora, O. Yaglidere, S. Padmanabhan, K. Nielsen, A. Ozcan. "Distributed Medical Image Analysis and Diagnosis through Crowd-Sourced Games: A Malaria Case Study". 2012 PLoS ONE 7(5): e37245. doi:10.1371/journal.pone.0037245

[15] S. Mavandadi, S. Feng, F. Yu, S. Dimitrov, R. Yu, A. Ozcan, "BioGames: A Platform for Crowd-Sourced Biomedical Image Analysis and Telediagnosis". Games Health Journal 2012 October; 1(5): 373–376. doi: 10.1089/g4h.2012.0054, PMCID: PMC3665415.