LEARNING FROM UNEQUALLY RELIABLE BLIND ENSEMBLES OF CLASSIFIERS

Panagiotis A. Traganitis[†], Alba Pagès-Zamora[⋆], and Georgios B. Giannakis[†]
[†] Dept. of ECE and Digital Technology Center, University of Minnesota, USA
[⋆]SPCOM Group, Universitat Politècnica de Catalunya BarcelonaTech, Spain

Abstract—**The rising interest in pattern recognition and data analytics has spurred the development of a plethora of machine learning algorithms and tools. However, as each algorithm has its strengths and weaknesses, one is motivated to judiciously fuse multiple algorithms in order to find the "best" performing one, for a given dataset. Ensemble learning aims to create a high-performance meta-algorithm, by combining the outputs from multiple algorithms. The present work introduces a simple blind scheme for learning from ensembles of classifiers, using joint matrix factorization. Blind refers to the combiner who has no knowledge of the ground-truth labels that each classifier has been trained on. Performance is evaluated on synthetic and real datasets.**

Index Terms—**Ensemble learning, multi-class classification, unsupervised**

I. Introduction

Nowadays, as vast amounts of data are constantly generated [8], there is a need to efficiently extract information from them. To this end, a number of algorithms have been designed by the machine learning, data mining, and signal processing communities [2], [15]. However, no one algorithm is suited for all tasks, as each relies on different assumptions and exhibits different strengths and weaknesses. *Ensemble learning* refers to the task of designing a skillful meta-learner by combining the results provided by multiple different learners or annotators.[1] In particular, ensemble classification refers to fusing the results provided by different classifiers. Such a setup emerges in diverse disciplines including medicine [28], biology [24], and economics [27], and has recently gained attention with the advent of crowdsourcing [4], [16] as well as services such as Amazon's Mechanical Turk [20] and Clickworker, to name a few.

Multiple approaches have been developed for supervised ensemble learning [10], the most popular ones being random forests [6], boosting [12], [13], and bagging [5]. These methods use labels to learn the optimal combination of algorithm responses. In many cases however, labeled data are not available to train the combining meta-classifier, or, the individual classifiers cannot be retrained, justifying the need for *unsupervised* (or *blind*) ensemble methods. One such paradigm is provided by crowdsourcing, where people are tasked with providing classification labels. Probably the simplest scheme for blind ensemble classification is majority voting, where the estimated label of a datum is the one that most annotators agree upon. This scheme, while relatively easy to implement, implicitly presumes that all annotators are equally "reliable," which is a typically unrealistic assumption, both in crowdsourcing as well as in ensemble learning setups. Other blind ensemble methods aim to estimate the parameters that characterize the annotators' performance, namely the sensitivity and specificity in binary classification problems, or the entries of the so-called confusion matrix [26] in multi-class settings. A joint maximum likelihood (ML) estimator of the unknown labels and the confusion matrices has been reported using the expectation-maximization (EM) algorithm [9]. As the EM algorithm does not guarantee convergence to the ML solution, recent works pursue alternative estimation methods. For instance, [17] advocates a spectral decomposition technique for binary classification, that yields the sensitivity and specificity of annotators, assuming class probabilities are a priori unknown. In the multi-class setting, [18] and [29] introduce tensor-based methods to estimate the unknown parameters and then initialize the EM algorithm of [9].

The present work puts forth a novel scheme for *multi-class blind ensemble classification*, built upon simple concepts from probability, linear algebra and optimization theory that enable assessing the reliability of multiple annotators and combining their answers.

Notation: Unless otherwise noted, boldface capital letters $\mathbf{X}$ denote matrices, boldface lowercase letters $\boldsymbol{x}$ denote vectors, and brackets indicate the entry of a vector or matrix; $\mathbb{R}^D$ stands for the D-dimensional real Euclidean space, diag$(\boldsymbol{x})$ for the diagonal matrix with $\boldsymbol{x}$ in its diagonal, $\mathbf{1}$ for the all ones vector, $\Pr$ for probability, or the probability mass function; $\sim$ denotes "distributed as," and $\mathbb{E}[\cdot]$ denotes expectation.

II. Preliminaries and Problem Formulation

Consider a dataset consisting of N data (possibly vectors) $\{x_n\}_{n=1}^N$ each belonging to one of K possible classes with corresponding labels $\{y_n\}_{n=1}^N$, e.g. $y_n = k$ if x_n belongs to class k. The pairs $\{(x_n,y_n)\}$ are drawn independently from an unknown joint distribution $\mathcal{P}$, and X and Y denote random variables such that $(X,Y) \sim \mathcal{P}$. Consider now M annotators that observe $\{x_n\}_{n=1}^N$, and provide estimates of

Work supported by NSF grants 1500713 and 1514056, as well as by "Ministerio de Economía y Competitividad" of the Spanish Government, ERDF funds (TEC2013-41315-R,TEC2015-69648-REDC,TEC2016-75067-C4-2-R) Emails: {traga003@umn.edu, alba.pages@upc.edu, georgios@umn.edu}

[1] The terms annotator, learner, and classifier will be used interchangeably.

labels. Let $f_m(x_n) \in \{1,\dots,K\}$ denote the label assigned to datum x_n by the m-th annotator. The task of *unsupervised ensemble classification* is, given only the annotator responses $\{f_m(x_n), m = 1,\dots,M\}_{n=1}^{N}$, to estimate the ground-truth labels of the data $\{y_n\}$. Before proceeding, we adopt the following assumptions.

As1. Class prior probabilities $\pi_k := \Pr(Y = k)$ are known and $\boldsymbol{\pi} := [\pi_1,\dots,\pi_K]^\top$

As2. Responses of different annotators for a datum, are conditionally independent, given the ground-truth label of the same datum Y; that is, for $m \neq m'$, it holds that $\Pr(f_m(X), f_{m'}(X)|Y) = \Pr(f_m(X)|Y)\Pr(f_{m'}(X)|Y)$

As3. The majority of annotators are better than random.

As1 is used to simplify the proposed algorithm, while As2 suggests that annotators make decisions independently of each other, which is rather a standard assumption in most prior works [9], [17], [25], [29]. As3 ensures convergence of the iterative algorithm in Sec. III.

Reliability per annotator f_m can be quantified by the so called *confusion* matrix $\boldsymbol{\Gamma}_m$, whose (k,k')-th entry is

$$[\boldsymbol{\Gamma}_m]_{kk'} := \Gamma_m(k,k') = \Pr(f_m(X) = k|Y = k'). \tag{1}$$

The $K \times K$ matrix $\boldsymbol{\Gamma}_m$ has non-negative entries that obey the simplex constraint, $\sum_{k=1}^{K} \Pr(f_m(X) = k|Y = k') = 1$, for $k' = 1,\dots,K$, hence columns of $\boldsymbol{\Gamma}_m$ sum up to 1, $\boldsymbol{\Gamma}_m^\top \mathbf{1} = \mathbf{1}$ and $\boldsymbol{\Gamma}_m \geq \mathbf{0}$. Each column of $\boldsymbol{\Gamma}_m$ showcases the average behavior of annotator m, and its probability of providing the correct answer, when presented with a datum from each class. For annotators that are better than random, the largest elements of each column of their confusion matrix will be those on the diagonal of $\boldsymbol{\Gamma}_m$; that is $[\boldsymbol{\Gamma}_m]_{kk} \geq [\boldsymbol{\Gamma}_m]_{k'k}$, for $k', k = 1,\dots,K$.

A. Maximum a posteriori label estimation

Given only annotator responses for all data, a straightforward approach to estimating their ground-truth labels is through maximum a posteriori (MAP) [19] estimation. In particular, for datum x the MAP estimate of y is

$$\hat{y}_{\text{map}}(x) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \log(\mathcal{L}(k,x)\Pr(Y = k)) \tag{2}$$

where $\mathcal{L}(k,x) := \Pr(f_1(x),\dots,f_M(x)|Y = k)$ denotes the likelihood of x. Since annotators make independent decisions it holds that $\mathcal{L}(k,x) = \prod_{m=1}^{M} \Pr(f_m(x)|Y = k)$ and thus the MAP estimator for y can be rewritten as

$$\hat{y}_{\text{map}}(x) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \log\pi_k + \sum_{m=1}^{M} \log(\Gamma_m(f_m(x),k)) \tag{3}$$

If all classes are equiprobable, then (3) yields the ML estimator of y. In order to obtain the MAP or ML estimate of the label, $\{\boldsymbol{\Gamma}_m\}_{m=1}^{M}$ must be available. Interestingly, the next section will show that these matrices can be recovered by the statistics of the annotator responses.

B. Statistics of annotator responses

Consider each label represented by the annotators using the canonical $K \times 1$ vector $\boldsymbol{e}_k$, meaning the k-th column of the $K \times K$ identity matrix $\mathbf{I}$. Let $\mathbf{f}_m(X)$ denote the m-th annotator's response in vector format. Since $\mathbf{f}_m(X)$ is just a vector representation of $f_m(X)$, we can write $\Pr(f_m(X) = k|Y = k') \equiv \Pr(\mathbf{f}_m(X) = \boldsymbol{e}_k|Y = k')$. With $\boldsymbol{\gamma}_{m,k}$ denoting the k-th column of $\boldsymbol{\Gamma}_m$, it thus holds that

$$\mathbb{E}[\mathbf{f}_m(X)|Y = k] = \sum_{k'=1}^{K} \boldsymbol{e}_{k'} \Pr\big(f_m(X) = k'|Y = k\big) = \boldsymbol{\gamma}_{m,k} \tag{4}$$

where the first equality comes from the definition of conditional expectation, and the second one holds because $\boldsymbol{e}_k$'s are columns of $\mathbf{I}$. Using (4) and the law of total probability, the mean vector of responses from annotator m, is hence given by

$$\mathbb{E}[\mathbf{f}_m(X)] = \sum_{k=1}^{K} \mathbb{E}[\mathbf{f}_m(X)|Y = k]\Pr(Y = k) = \boldsymbol{\Gamma}_m\boldsymbol{\pi} \tag{5}$$

The $K \times K$ cross-correlation matrix between the responses of annotators m and $m' \neq m$, namely $\mathbf{R}_{mm'} := \mathbb{E}[\mathbf{f}_m(X)\mathbf{f}_{m'}^\top(X)]$, can be expressed as

$$\begin{aligned}\mathbf{R}_{mm'} &= \sum_{k=1}^{K} \mathbb{E}[\mathbf{f}_m(X)|Y = k]\mathbb{E}[\mathbf{f}_{m'}^\top(X)|Y = k]\Pr(Y = k)\\ &= \boldsymbol{\Gamma}_m \text{diag}(\boldsymbol{\pi})\boldsymbol{\Gamma}_{m'}^\top \end{aligned}\tag{6}$$

where we successively relied on the law of total probability, As2, and (4). Accordingly, and upon defining $\boldsymbol{\Pi} := \text{diag}(\boldsymbol{\pi}) - \boldsymbol{\pi}\boldsymbol{\pi}^\top$, the cross-covariance matrix $\mathbf{C}_{mm'} := \mathbb{E}\left[(\mathbf{f}_m(X) - \mathbb{E}[\mathbf{f}_m(X)])(\mathbf{f}_{m'}(X) - \mathbb{E}[\mathbf{f}_{m'}(X)])^\top\right]$ is given by

$$\mathbf{C}_{mm'} = \boldsymbol{\Gamma}_m \boldsymbol{\Pi} \boldsymbol{\Gamma}_{m'}^\top. \tag{7}$$

With $\mathbf{F}_m := [\mathbf{f}_m(x_1), \mathbf{f}_m(x_2), \dots, \mathbf{f}_m(x_N)]$ the sample mean of the m-th annotator responses can be readily obtained as

$$\boldsymbol{\mu}_m = \frac{1}{N}\sum_{n=1}^{N} \mathbf{f}_m(x_n) = \frac{1}{N}\mathbf{F}_m\mathbf{1}. \tag{8}$$

Accordingly, the sample cross-covariance $\mathbf{S}_{mm'}$ matrices between the responses of annotators m and $m' \neq m$, are given by

$$\mathbf{S}_{mm'} = \frac{1}{N-1}(\mathbf{F}_m - \boldsymbol{\mu}_m\mathbf{1}^\top)(\mathbf{F}_{m'} - \boldsymbol{\mu}_{m'}\mathbf{1}^\top)^\top. \tag{9}$$

Clearly, $\mathbf{S}_{mm'} = \mathbf{S}_{m'm}^\top$, and as N increases, $\{\boldsymbol{\mu}_m\}$ and $\{\mathbf{S}_{mm'}\}$ approach their ensemble counterparts in (5) and (7).

III. Confusion Matrix Estimation

Having available first- and second-order statistics of annotator responses $\{\boldsymbol{\mu}_m\}_{m=1}^{M}$ and $\{\mathbf{S}_{mm'}\}_{m,m'=1}^{M}$, estimates of the confusion matrices can be readily extracted from them [cf.(7)]. This procedure can be cast as the following constrained optimization problem, which requires joint factorization of the matrices $\{\mathbf{S}_{mm'}\}$. Specifically, consider

$$\begin{aligned}&\min_{\{\boldsymbol{\Gamma}_m\}_{m=1}^{M}} \; h(\{\boldsymbol{\Gamma}_m\}_{m=1}^{M}) \\ &\text{s.to} \quad \boldsymbol{\Gamma}_m \geq \mathbf{0}, \quad \boldsymbol{\Gamma}_m^\top\mathbf{1} = \mathbf{1}, \quad m = 1,\dots,M\end{aligned}\tag{10}$$

where

$$h(\{\boldsymbol{\Gamma}_m\}) := \frac{1}{2}\sum_{m=1}^{M} \|\boldsymbol{\mu}_m - \boldsymbol{\Gamma}_m\boldsymbol{\pi}\|_2^2 + \frac{1}{2}\sum_{\substack{m=1\\ m'>m}}^{M} \|\mathbf{S}_{mm'} - \boldsymbol{\Gamma}_m\boldsymbol{\Pi}\boldsymbol{\Gamma}_{m'}^\top\|_F^2.$$

Algorithm 1 Blind Multi-class Ensemble Classifier

```
Input: Annotator responses $\{\mathbf{F}_m\}_{m=1}^M$; priors $\boldsymbol{\pi}$; $\lambda > 0$
Output: Estimates of data labels $\{\hat{y}_n\}_{n=1}^N$; Estimates of annotator confusion matrices $\{\hat{\mathbf{\Gamma}}_m\}_{m=1}^M$
1: Compute $\{\boldsymbol{\mu}_m\}$,$\{\mathbf{S}_{mm'}\}$ using (8), and (9).
2: Initialize $\{\mathbf{\Gamma}_m\}$ and $\{\mathbf{\Phi}_m\}$ according to Sec. III-B.
3: do
4:     Update $\{\mathbf{\Gamma}_m\}_{m=1}^M$ using (15).
5:     Update $\{\mathbf{\Phi}_m\}_{m=1}^M$ using (16).
6:     Update $\{\mathbf{\Delta}_m\}_{m=1}^M$ using (17).
7: while not converged
8: for $n = 1,\dots,N$ do
9:     Estimate label $y_n$ using (11).
10: end for
```

Collect the set of constraints per matrix to the convex set $\mathcal{C} := \{\mathbf{\Gamma} \in \mathbb{R}^{K\times K} : \mathbf{\Gamma} \geq \mathbf{0}, \mathbf{\Gamma}^\top \mathbf{1} = \mathbf{1}\}$, where essentially each column lies on a probability simplex. After obtaining estimates $\{\hat{\mathbf{\Gamma}}_m\}_{m=1}^M$, estimates of the labels $\{\hat{y}_n\}_{n=1}^N$ can be obtained using the ML/MAP estimator described in Section II-A; that is for $n = 1,\dots,N$,

$$\hat{y}_{\text{map}}(x_n) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \log\pi_k + \sum_{m=1}^M \log\hat{\Gamma}_m(f_m(x_n),k) \quad (11)$$

where $\hat{\Gamma}_m(k',k) = [\hat{\mathbf{\Gamma}}_m]_{k'k}$. The ensuing section provides an iterative algorithm for solving (10).

A. ADMM algorithm for estimating confusion matrices

In this section, the alternating direction method of multipliers (ADMM) is employed to solve the constrained optimization problem (10); see e.g. [3] and [14]. The ADMM allows for decoupling the constraints across annotators, resulting in a simple and efficient iterative algorithm.

Consider the following optimization problem that is equivalent to (10),

$$\begin{aligned} \min_{\{\mathbf{\Gamma}_m\}_{m=1}^M, \{\mathbf{\Phi}_m\}_{m=1}^M} \quad & \bar{h}(\{\mathbf{\Gamma}_m, \mathbf{\Phi}_m\}_{m=1}^M) \\ \text{s.to} \quad & \mathbf{\Gamma}_m = \mathbf{\Phi}_m, \quad m = 1,\dots,M \end{aligned} \quad (12)$$

with $\{\mathbf{\Phi}_m\}$ being auxiliary variables,

$$\bar{h}(\{\mathbf{\Gamma}_m, \mathbf{\Phi}_m\}_{m=1}^M) = h(\{\mathbf{\Gamma}_m\}_{m=1}^M) + \sum_{m=1}^M \rho_{\mathcal{C}}(\mathbf{\Phi}_m)$$

and $\rho_{\mathcal{C}}$ is an indicator function for the constraints of (10), namely

$$\rho_{\mathcal{C}}(\mathbf{A}) := \begin{cases} 0 & \text{if } \mathbf{A} \in \mathcal{C} \\ \infty & \text{otherwise.} \end{cases} \quad (13)$$

The augmented Lagrangian of (12) is then

$$g = \bar{h}(\{\mathbf{\Gamma}_m, \mathbf{\Phi}_m\}_{m=1}^M) + \frac{\lambda}{2} \sum_{m=1}^M \|\mathbf{\Gamma}_m - \mathbf{\Phi}_m + \mathbf{\Delta}_m\|_F^2 \quad (14)$$

where the $K \times K$ matrices $\{\mathbf{\Delta}_m\}_{m=1}^M$ contain the scaled Lagrange multipliers, and λ is a positive scalar.

Per ADMM iteration, (14) is minimized in an alternating fashion, with respect to (w.r.t.) $\{\mathbf{\Gamma}_m\}$ and $\{\mathbf{\Phi}_m\}$ before performing a gradient ascent step for $\{\mathbf{\Delta}_m\}$. Specifically, the update for $\mathbf{\Gamma}_m$ at iteration $\ell + 1$ is obtained by setting the gradient of g w.r.t. $\mathbf{\Gamma}_m$ to $\mathbf{0}$, and solving for $\mathbf{\Gamma}_m$. Since $\mathbf{S}_{m'm} = \mathbf{S}_{mm'}^\top$ and $\mathbf{\Pi} = \mathbf{\Pi}^\top$, it is easy to see that the update w.r.t. $\mathbf{\Gamma}_m$ can be expressed as

$$\begin{aligned} &\mathbf{\Gamma}_m^{(\ell+1)} \left(\lambda \mathbf{I} + \boldsymbol{\pi}\boldsymbol{\pi}^\top + \sum_{m' \neq m}^M \mathbf{\Pi}^\top \mathbf{\Gamma}_{m'}^{(\ell)\top} \mathbf{\Gamma}_{m'}^{(\ell)} \mathbf{\Pi} \right) \\ &= \boldsymbol{\mu}_m \boldsymbol{\pi}^\top + \sum_{m' \neq m}^M \mathbf{S}_{m'm}^\top \mathbf{\Gamma}_{m'}^{(\ell)} \mathbf{\Pi} + \mu \mathbf{\Phi}_m^{(\ell)} - \mu \mathbf{\Delta}_m^{(\ell)}. \end{aligned} \quad (15)$$

Here superscripts denote iteration indices. Accordingly, the update for $\mathbf{\Phi}_m$ is given by

$$\mathbf{\Phi}_m^{(\ell+1)} = P_{\mathcal{C}}\left(\mathbf{\Gamma}_m^{(\ell+1)} + \mathbf{\Delta}_m^{(\ell)}\right) \quad (16)$$

where $P_{\mathcal{C}}$ is the projection operator onto the convex set $\mathcal{C}$ with each column of $\mathbf{\Gamma}_m^{(\ell+1)} + \mathbf{\Delta}_m^{(\ell)}$ projected onto the probability simplex. This projection can be performed using efficient methods [11]. Finally, a gradient ascent step is performed per $\mathbf{\Delta}_m$, as follows

$$\mathbf{\Delta}_m^{(\ell+1)} = \mathbf{\Delta}_m^{(\ell)} + \mathbf{\Gamma}_m^{(\ell+1)} - \mathbf{\Phi}_m^{(\ell+1)}. \quad (17)$$

The entire ensemble classification procedure is tabulated in Alg. 1 and the ADMM algorithm is listed in steps 2-7. Note that the computational complexity per ADMM iteration is dominated by (15). Thus, the ADMM algorithm incurs computational complexity of approximately $\mathcal{O}(IM^2K^3)$, where I is the number of required iterations until convergence.

Remark 1. The total number of unknowns in (10) is $MK(K-1)$, since each column of a confusion matrix must sum up to 1. The total number of equations is then, $M(M-1)K^2/2 + MK$, where the factor $M(M-1)/2$ comes from the fact that $\mathbf{S}_{mm'} = \mathbf{S}_{m'm}^\top$ for $m \neq m'$. This implies that problem (10) is solvable with at least three annotators.

Remark 2. The estimates $\{\hat{y}_n\}$ and $\{\hat{\mathbf{\Gamma}}_m\}$ provided by Alg. 1 can also be employed to initialize the EM algorithm of [9].

Remark 3. With $\{\hat{\mathbf{\Gamma}}_m\}$ available, annotator reliability can be determined by inspecting the columns of the confusion matrices.

B. Algorithm initialization

Problem (10) is non-convex due to the multiplicative coupling between confusion matrices of different annotators. Thus, initialization plays an instrumental role to ensure convergence of $\{\mathbf{\Gamma}_m\}$ to a point that will provide sensible results. Since, by As3, most annotators are better than random, a simple initialization scheme is to generate a random $\mathbf{\Gamma}_m^{(0)}$ such that $\mathbf{\Gamma}_m^{(0)} \in \mathcal{C}$, and $[\mathbf{\Gamma}_m^{(0)}]_{kk} \geq [\mathbf{\Gamma}_m^{(0)}]_{k'k}$, for $k',k = 1,\dots,K$ and $m = 1,2,\dots,M$. As corroborated by our numerical tests, this initialization scheme is very effective in practice.

IV. Numerical Tests

The performance of the proposed algorithm was evaluated using synthetic and real datasets. Using both MAP and ML estimation in step 9, Alg. 1 is compared to majority voting, and in the case of synthetic data, also to "oracle" estimators, that

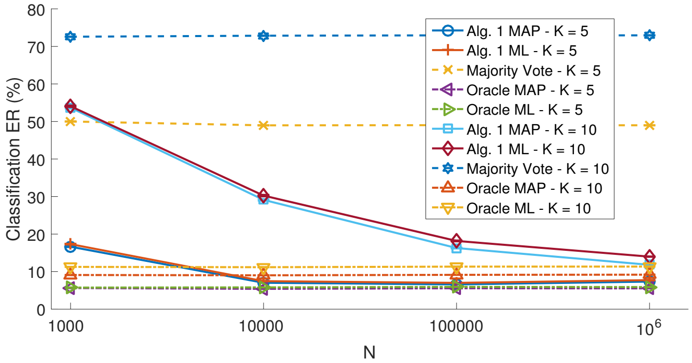

Fig. 1: Classification ER for synthetic datasets with $K=5$, $K=10$, and $M=20$ annotators.

is MAP/ML estimators that know the true confusion matrices of the annotators. The metric utilized in all experiments is the classification error rate (ER), defined as the percentage of misclassified data, where $ER = 100\%$ indicates that all N data have been misclassified, and $ER=0\%$ indicates perfect classification accuracy. All results represent averages over 10 independent Monte Carlo runs, using MATLAB [23]. In all experiments λ is set to 10^{-1}.

A. Synthetic data

For the synthetic data tests, N ground-truth labels $\{y_n\}_{n=1}^{N}$, each corresponding to one out of K possible classes, were generated i.i.d. at random according to $\boldsymbol{\pi}$, that is $y_n \sim \boldsymbol{\pi}$, for $n = 1,\dots,N$. Afterwards, $\{\boldsymbol{\Gamma}_m\}_{m=1}^{M}$ were generated at random, such that $\boldsymbol{\Gamma}_m \in \mathcal{C}$, for all $m = 1,\dots,M$. Then annotators responses are generated as follows: if $y_n = k$, then the response of annotator m will be generated randomly according to $\boldsymbol{\gamma}_{m,k}$, that is $f_m(x_n) \sim \boldsymbol{\gamma}_{m,k}$. This value will then be converted to the appropriate vector format $\mathbf{f}_m(x_n)$, as described in Sec. II-B. In all cases, $\lfloor M/2 \rfloor + 1$ annotators were generated to be better than random, as per As3, and $\lfloor M/2 \rfloor - 1$ were generated completely at random. Fig. 1 shows the classification ER for a synthetic dataset with $M = 20$ annotators for varying N, for two different cases: one with $K = 5$, and one with $K = 10$. For $K = 5$, data were generated with $\boldsymbol{\pi} = [0.1365, 0.3396, 0.1961, 0.0973, 0.2305]^\top$, while for $K = 10$ the prior probabilities were $\boldsymbol{\pi} = [0.1806,\ 0.1991,\ 0.1241,\ 0.1334,\ 0.0425,\ 0.0118,\ 0.0002, 0.1201, 0.1506, 0.0376]^\top$. Clearly, the proposed scheme (denoted as *Alg. 1 MAP* and *Alg. 1 ML*) outperforms majority voting, and as N increases its ER approaches that of the "oracle" ones. This makes sense since as N increases, the sample averages (8) and (9) approach their ensemble counterparts in (5) and (7), enabling more accurate estimation of the confusion matrices. Fig. 2 shows the same experiment, but for fixed $N=10^6$, and varying number of annotators M. Again, Alg. 1 markedly outperforms majority voting, and as M increases it approaches the performance of the "oracle" estimators. This result suggests that more annotators are always preferable, as long as the majority of them are better than random. Furthermore, note that the proposed algorithms perform well, in spite of the presence of possibly unreliable annotators, which speaks for the potential of the novel approach in adversarial learning setups [1], [7].

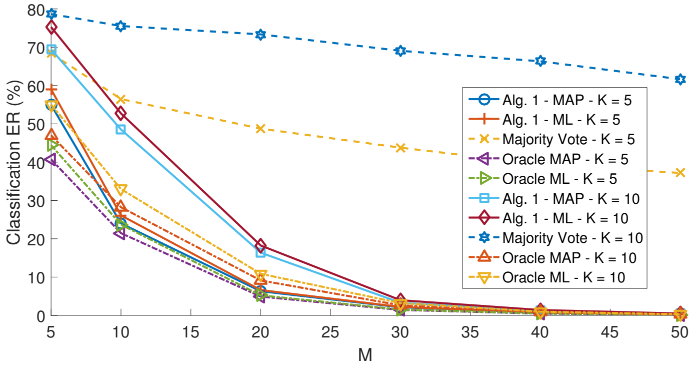

Fig. 2: Classification ER for synthetic datasets with $K=5$, $K=10$, and $N=10^6$ data.

B. Real data

Further tests were conducted using two real datasets. The MNIST [21] dataset, and the Connect-4 [22] dataset. MNIST contains $N = 70,000$ 28×28 images of handwritten digits, each belonging to one of $K = 10$ classes (one per digit). Connect-4 contains $N = 67,557$ vectors of size 42×1, each representing the possible positions in a connect-4 game. These vectors belong to one of $K = 3$ classes, indicating whether the first player won, lost, or, if the game ended as a tie. A collection of $M = 12$ classification algorithms, from MATLAB's machine learning toolbox, were trained on different randomly selected subsets of $1,000$ data instances for MNIST and 300 data instances for Connect-4. Afterwards, the algorithms provided labels for all data in each dataset. Vector $\boldsymbol{\pi}$ was estimated by measuring the frequency of each label from the entire dataset. Table I lists the ER performance of the proposed scheme compared to majority voting (MV) and the annotator with the highest accuracy (Single best) for these two datasets. As with synthetic data, the proposed method outperforms majority voting, as well as the single best classifier.

Dataset	Alg. 1 MAP	Alg. 1 ML	MV	Single best
MNIST	**7.97%**	**7.96%**	9.07%	10.68%
Connect-4	**30.2%**	**31.68%**	45.14%	38.07%

TABLE I: Classification ER for Alg.1 and majority voting for real datasets MNIST and Connect-4.

V. Conclusions

This paper introduced a novel approach to blind ensemble and crowdsourced classification that relies solely on the annotator responses to assess their quality and combine their answers. The novel scheme was implemented using ADMM, and its performance was evaluated on real and synthetic data. Future research will focus on extensive numerical tests with real datasets, as well as algorithms that can infer prior probabilities from annotator responses, along with distributed and online implementations.

References

[1] B. Biggio, G. Fumera, and F. Roli, “Multiple classifier systems for robust classifier design in adversarial environments,” *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 27–41, 2010.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, “Distributed optimization and statistical learning via the alternating direction method of multipliers,” *Foundations and Trends in Machine Learning*, vol. 3, no. 1, pp. 1–122, 2011.

[4] D. C. Brabham, “Crowdsourcing as a model for problem solving: An introduction and cases,” *Convergence*, vol. 14, no. 1, pp. 75–90, 2008.

[5] L. Breiman, “Bagging predictors,” *Machine Learning*, vol. 24, no. 2, pp. 123–140, 1996.

[6] ——, “Random forests,” *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[7] D. Chinavle, P. Kolari, T. Oates, and T. Finin, “Ensembles in adversarial classification for spam,” in *Proc. of the ACM Conf. on Information and Knowledge Management*. ACM, 2009, pp. 2015–2018.

[8] K. Cukier, “Data, data everywhere,” *The Economist*, 2010. [Online]. Available: http://www.economist.com/node/15557443

[9] A. P. Dawid and A. M. Skene, “Maximum likelihood estimation of observer error-rates using the EM algorithm,” *Applied Statistics*, pp. 20–28, 1979.

[10] T. G. Dietterich, “Ensemble methods in machine learning,” in *Intl. Workshop on Multiple Classifier Systems*. Springer, 2000, pp. 1–15.

[11] J. Duchi, S. Shalev-Shwartz, Y. Singer, and T. Chandra, “Efficient projections onto the l1-ball for learning in high dimensions,” in *Proc. of the Intl. Conf, on Machine Learning*. ACM, 2008, pp. 272–279.

[12] Y. Freund, “Boosting a weak learning algorithm by majority,” *Information and Computation*, vol. 121, no. 2, pp. 256–285, 1995.

[13] Y. Freund, R. E. Schapire *et al.*, “Experiments with a new boosting algorithm,” in *Proc. of the Intl. Conf. on Machine Learning*, vol. 96, 1996, pp. 148–156.

[14] G. B. Giannakis, Q. Ling, G. Mateos, I. D. Schizas, and H. Zhu, “Decentralized learning for wireless communications and networking,” in *Splitting Methods in Communication and Imaging, Science and Engineering*, R. Glowinski, S. Osher, and W. Yin, Eds. Springer, 2016.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2009.

[16] J. Howe, “The rise of crowdsourcing,” *Wired Magazine*, vol. 14, no. 6, pp. 1–4, 2006.

[17] A. Jaffe, B. Nadler, and Y. Kluger, “Estimating the accuracies of multiple classifiers without labeled data.” in *AISTATS*, vol. 2, 2015, p. 4.

[18] P. Jain and S. Oh, “Learning mixtures of discrete product distributions using spectral decompositions.” *Journal of Machine Learning Research*, 2014.

[19] S. M. Kay, *Fundamentals of Statistical Signal Processing, volume I: Estimation Theory*. Prentice Hall, 1993.

[20] A. Kittur, E. H. Chi, and B. Suh, “Crowdsourcing user studies with mechanical turk,” in *Proc. of SIGCHI Conf. on Human Factors in Computing Systems*. ACM, 2008, pp. 453–456.

[21] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, “Gradient-based learning applied to document recognition,” *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[22] M. Lichman, “UCI machine learning repository,” 2013. [Online]. Available: http://archive.ics.uci.edu/ml

[23] MATLAB, *version 8.6.0 (R2015b)*. Natick, Massachusetts: The MathWorks Inc., 2015.

[24] M. Micsinai, F. Parisi, F. Strino, P. Asp, B. D. Dynlacht, and Y. Kluger, “Picking chip-seq peak detectors for analyzing chromatin modification experiments,” *Nucleic Acids Research*, 2012.

[25] F. Parisi, F. Strino, B. Nadler, and Y. Kluger, “Ranking and combining multiple predictors without labeled data,” *Proc. of the Ntl. Academy of Sciences*, vol. 111, no. 4, pp. 1253–1258, 2014.

[26] D. M. W. Powers, “Evaluation: from precision, recall and f-measure to roc, informedness, markedness and correlation,” *Intl. Journal of Machine Learning Technology*, vol. 2, pp. 37–63, 2011.

[27] A. Timmermann, “Forecast combinations,” *Handbook of Economic Forecasting*, vol. 1, pp. 135–196, 2006.

[28] F. Wright, C. De Vito, B. Langer, A. Hunter *et al.*, “Multidisciplinary cancer conferences: A systematic review and development of practice standards,” *European Journal of Cancer*, vol. 43, pp. 1002–1010, 2007.

[29] Y. Zhang, X. Chen, D. Zhou, and M. I. Jordan, “Spectral methods meet EM: A provably optimal algorithm for crowdsourcing,” in *Advances in Neural Information Processing Systems*, 2014, pp. 1260–1268.