Blind Multi-class Ensemble Learning with Dependent Classifiers

Panagiotis A. Traganitis, and Georgios B. Giannakis
Dept. of ECE & Digital Technology Center, Univ. of Minnesota, USA.

***Abstract*—In recent years, advances in pattern recognition and data analytics have spurred the development of a plethora of machine learning algorithms and tools. However, as each algorithm exhibits different behavior for different types of data, 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 such a high-performance meta-learner, by combining the outputs from multiple algorithms. The present work introduces a simple blind scheme for learning from ensembles of classifiers. Blind refers to the combiner who has no knowledge of the ground-truth labels that each classifier has been trained on. While most current works presume that all classifiers are independent, this work introduces a scheme that can handle dependencies between classifiers. Preliminary tests on synthetic data showcase the potential of the proposed approach.**

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

I. Introduction

The vast amounts of data that are generated daily [5] have resulted in the pressing need to efficiently extract information from them. To this end, a large number of algorithms have been developed by the machine learning, data mining, and signal processing communities [1], [10]. However, no one algorithm is suited for all tasks, as each relies on different assumptions and thus behaves differently on different datasets. *Ensemble learning* refers to the task of designing a 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 [20], economics [23], and distributed detection [26], and has recently gained attention with the advent of crowdsourcing [2], [11] as well as services such as Amazon's Mechanical Turk, CrowdFlower, and Clickworker, to name a few.

Multiple approaches have been developed for supervised ensemble learning [7], the most popular ones being random forests [3] and boosting [8]. These methods use labels to learn the optimal combination of annotator responses. In many cases however, labeled data are not available to train the combining meta-classifier, 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 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 [6]. As the EM algorithm does not guarantee convergence to the ML solution, recent works pursue alternative estimation methods. Recently, [14] advocated a spectral decomposition technique of the second-order statistics of annotator responses for binary classification, that yields the reliability parameters of annotators, when class probabilities are unknown. In the multi-class setting, [16] employs an iterative method that solves multiple binary classification problems. In addition, [15] and [29] utilize third-order moments and orthogonal tensor decomposition to estimate the unknown reliability parameters and then initialize the EM algorithm of [6], while [25] and [24] use joint matrix and joint tensor factorizations respectively.

All aforementioned approaches assume that annotators are conditionally independent. However, this assumption may not always hold. For example, dependencies between annotators may arise if they are trained on very similar datasets. The method of [14] for binary classification was extended in [13] to handle dependencies between annotators, while [21] introduced a deep learning framework for binary ensemble classification when annotators are dependent.

The present work puts forth a novel scheme for *multi-class blind ensemble learning with dependent classifiers*, built upon simple concepts from probability, linear algebra and optimization theory. The proposed scheme enables the assessment of annotator reliability and judiciously fuses their responses, and the presence of annotator dependencies markedly extends the scope of our previous work in [24].

Notation: Unless otherwise noted, lowercase bold letters, $\boldsymbol{x}$, denote vectors, uppercase bold letters, $\mathbf{X}$, represent matrices, and calligraphic uppercase letters, $\mathcal{X}$, stand for sets. The (i,j)th entry of matrix $\mathbf{X}$ is denoted by $[\mathbf{X}]_{ij}$; and its rank by $\text{rank}(\mathbf{X})$; $\mathbf{X}^\top$ denotes the transpose of matrix $\mathbf{X}$; $\mathbb{R}^D$ stands for the D-dimensional real Euclidean space; $\|\cdot\|$ denotes the

Work in this paper was supported by NSF grants 1500713 and 1514056. Emails: {traga003,georgios}@umn.edu

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

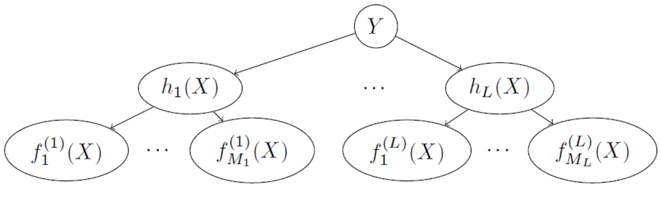

Fig. 1. Graphical representations of the probabilistic models employed for blind ensemble learning with dependent classifiers.

ℓ_2-norm; and the vector outer product is denoted by $\circ$. Pr denotes probability, or the probability mass function; $\sim$ denotes "distributed as," and $\mathbb{E}[\cdot]$ denotes expectation. Underlined capital letters $\underline{X}$ denote tensors; while $[[\mathbf{A}, \mathbf{B}, \mathbf{C}]]_K$ is used to denote compactly a K-factor PARAFAC tensor [9], [22] with factor matrices $\mathbf{A} = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_K], \mathbf{B} = [\boldsymbol{b}_1, \ldots, \boldsymbol{b}_K], \mathbf{C} = [\boldsymbol{c}_1, \ldots, \boldsymbol{c}_K]$, that is $[[\mathbf{A}, \mathbf{B}, \mathbf{C}]]_K = \sum_{k=1}^{K} \boldsymbol{a}_k \circ \boldsymbol{b}_k \circ \boldsymbol{c}_k$.

II. Problem Statement and Preliminaries

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 labels. Let $f_m(x_n) \in \{1, \ldots, K\}$ 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 $\{f_m(x_n), m = 1, \ldots, M\}_{n=1}^{N}$, to estimate the ground-truth labels of the data $\{y_n\}$.

Most prior works tackle the unsupervised ensemble classification task using the Dawid and Skene model [6]. Under this model, responses of different annotators per datum, are conditionally independent, given the ground-truth label Y of the same datum X; that is

$$\begin{aligned}&\Pr\left(f_1(X) = k_1, \ldots, f_M(X) = k_M | Y = k\right)\\ &= \prod_{m=1}^{M} \Pr\left(f_m(X) = k_m | Y = k\right).\end{aligned}$$

This model, while convenient, fails to account for dependencies between annotators. To circumvent this issue, the present work puts forth a more general model that allows for dependencies between annotators. In this model we consider L groups of dependent annotators. Suppose that group ℓ has M_ℓ annotators ($\sum_{\ell=1}^{L} M_\ell = M$), and denote the annotator responses of group ℓ to datum X as $\{f_m^{(\ell)}(X)\}_{m=1}^{M_\ell}$. The dependencies within each group ℓ are captured using a hidden variable $h_\ell(X) \in \{1, \ldots, K\}$, conditioned on which the responses of annotators within the group become independent, that is

$$\begin{aligned}&\Pr\left(f_1^{(\ell)}(X) = k_1, \ldots, f_{M_\ell}^{(\ell)}(X) = k_{M_\ell} | h_\ell(X) = k\right)\\ &= \prod_{m=1}^{M_\ell} \Pr\left(f_m^{(\ell)}(X) = k_m | h_\ell(X) = k\right) \qquad \forall \ell. \qquad (1)\end{aligned}$$

This suggests that, while dependent on the same hidden variable, within each group annotators make independent decisions. The hidden variables $\{h_\ell(X)\}_{\ell=1}^{L}$ are also assumed conditionally independent given the ground-truth label of the datum Y, that is

$$\begin{aligned}&\Pr\left(h_1(X) = k_1, \ldots, h_L(X) = k_L | Y = k\right)\\ &= \prod_{\ell=1}^{L} \Pr\left(h_\ell(X) = k_\ell | Y = k\right). \qquad (2)\end{aligned}$$

This model has been used successfully in [13] as well as in [4] in the case of distributed detection. A graphical representation of the proposed model is shown in Fig. 1.

Per group ℓ, each annotator $f_m^{(\ell)}$ can be characterized by the so called *confusion* matrix $\mathbf{\Gamma}_m^{(\ell)}$, whose (k, k')-th entry is

$$[\mathbf{\Gamma}_m^{(\ell)}]_{kk'} := \Gamma_m^{(\ell)}(k, k') = \Pr\left(f_m^{(\ell)}(X) = k | h_\ell(X) = k'\right).$$

The $K \times K$ matrix $\mathbf{\Gamma}_m^{(\ell)}$ has non-negative entries that obey the simplex constraint, $\sum_{k=1}^{K} \Pr\left(f_m^{(\ell)}(X) = k | h_\ell(X) = k'\right) = 1$, for $k' = 1, \ldots, K$, hence columns of $\mathbf{\Gamma}_m^{(\ell)}$ sum up to 1, $\mathbf{\Gamma}_m^{(\ell)\top} \mathbf{1} = \mathbf{1}$ and $\mathbf{\Gamma}_m^{(\ell)} \geq \mathbf{0}$. Each column of $\mathbf{\Gamma}_m^{(\ell)}$ showcases the average behavior of annotator m, and its probability of following the value of the hidden variable $h_\ell(X)$, when presented with a datum from each class. Collect the set of constraints per matrix in 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, and let $\mathcal{C}_p := \{\boldsymbol{u} \in \mathbb{R}^K : \boldsymbol{u} \geq \mathbf{0}, \boldsymbol{u}^\top \mathbf{1} = 1\}$ denote the simplex constraint set for a vector. Accordingly, group ℓ can be characterized by a $K \times K$ confusion matrix $\mathbf{\Phi}_\ell \in \mathcal{C}$ for the hidden variable h_ℓ. For this matrix its (k, k')-th entry is $[\mathbf{\Phi}_\ell]_{kk'} := \Phi_\ell(k, k') = \Pr\left(h_\ell(X) = k | Y = k'\right).$ Before proceeding we make the following assumptions.

As1. The groups of dependent annotators are known.
As2. Within each group all annotators are better than random.
As3. The majority of hidden variables are better than random.

As1 is used to simplify the proposed algorithm, while As2 and As3 alleviate the permutation ambiguity introduced by the iterative algorithm in Sec. III. For annotators that are better than random, the largest elements of each column of their confusion matrix will be those on the diagonal of $\mathbf{\Gamma}_m^{(\ell)}$; that is $[\mathbf{\Gamma}_m^{(\ell)}]_{kk} \geq [\mathbf{\Gamma}_m^{(\ell)}]_{k'k}$, for $k', k = 1, \ldots, K$.

III. Blind Ensemble Learning with dependent classifiers

Building on the proposed model of the previous section, this section puts forth a novel approach for blind ensemble learning using dependent classifiers. Our method exploits the hierarchical structure of the joint pmf [cf. Fig. 1] in a two step approach: First, estimates of the hidden variables for all groups and all data are obtained $\{\hat{h}_\ell(x_n)\}_{\ell=1,n=1}^{L,N}$; then, the estimates of the hidden variables are used to obtain an estimate of data labels $\{\hat{y}_n\}_{n=1}^{N}$.

A. Maximum a posteriori hidden variable estimation

Given only annotator responses for all data, an approach to estimating hidden variable values for each group, that minimizes the probability of error, is maximum a posteriori (MAP) [18] detection. In particular, for datum x the MAP estimate of $h_\ell(x)$ is

$$\hat{h}_\ell(x) = \underset{k\in\{1,\dots,K\}}{\operatorname{argmax}} \log\left(\mathcal{L}_\ell(k,x)\Pr(h_\ell(x)=k)\right) \quad (3)$$

where

$$\mathcal{L}_\ell(k,x) := \Pr\left(f_1^{(\ell)}(x)=k_1,\dots,f_{M_\ell}^{(\ell)}(x)=k_{M_\ell}|h_\ell(x)=k\right)$$

denotes the likelihood of x for group ℓ. From (1) it holds that $\mathcal{L}_\ell(k,x) = \prod_{m=1}^{M_\ell}\Pr\left(f_m^{(\ell)}(x)=k_m|h_\ell(x)=k\right)$ and thus the MAP estimator for $h_\ell(x)$ can be rewritten as

$$\hat{h}_\ell(x) = \underset{k\in\{1,\dots,K\}}{\arg\max}\log p_k^{(\ell)} + \sum_{m=1}^{M_\ell}\log(\Gamma_m^{(\ell)}(f_m^\ell(x),k)) \quad (4)$$

where $p_k^{(\ell)} = \Pr(h_\ell(X)=k)$. If all classes are considered equiprobable, then (4) yields the ML estimator of $h_\ell(x)$. In order to obtain the MAP or ML estimate of the hidden variable, $\{\mathbf{\Gamma}_m^{(\ell)}\}_{m=1}^{M_\ell}$ and $\boldsymbol{p}^{(\ell)} = [p_1^{(\ell)},\dots,p_K^{(\ell)}]^\top$ must be available. Interestingly, the next section will show that $\{\mathbf{\Gamma}_m^{(\ell)}\}_{m=1}^{M_\ell}$ and $\boldsymbol{p}^{(\ell)}$ can be recovered by the statistics of the responses from annotators in group ℓ.

B. Statistics of annotator responses

For the remainder of this subsection we focus on group ℓ. 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 $\mathbf{f}_m^{(\ell)}(X)$ denote the response of the m-th annotator in vector format. Since $\mathbf{f}_m^{(\ell)}(X)$ is just a vector representation of $f_m^{(\ell)}(X)$, we can write $\Pr\left(f_m^{(\ell)}(X)=k|h_\ell(X)=k'\right) \equiv \Pr\left(\mathbf{f}_m^{(\ell)}(X)=\boldsymbol{e}_k|h_\ell(X)=k'\right)$. With $\boldsymbol{\gamma}_{m,k}^{(\ell)}$ denoting the k-th column of $\mathbf{\Gamma}_m^{(\ell)}$, it thus holds that

$$\begin{aligned}\mathbb{E}[\mathbf{f}_m^{(\ell)}(X)|h_\ell(X)=k] &= \sum_{k'=1}^{K}\boldsymbol{e}_{k'}\Pr\left(f_m^{(\ell)}(X)=k'|h_\ell(X)=k\right)\\ &= \boldsymbol{\gamma}_{m,k}^{(\ell)} \qquad (5)\end{aligned}$$

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 (5) and the law of total probability, the mean vector response from annotator m, is hence given by

$$\mathbb{E}[\mathbf{f}_m^{(\ell)}(X)] = \sum_{k=1}^{K}\mathbb{E}[\mathbf{f}_m^{(\ell)}(X)|h_\ell(X)=k]p_k^{(\ell)} = \mathbf{\Gamma}_m^{(\ell)}\boldsymbol{p}^{(\ell)}. \quad (6)$$

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

$$\begin{aligned}\mathbf{R}_{mm'}^{(\ell)} &= \sum_{k=1}^{K}\boldsymbol{\gamma}_{m,k}^{(\ell)}\boldsymbol{\gamma}_{m',k}^{(\ell)\top}p_k^{(\ell)} = \mathbf{\Gamma}_m^{(\ell)}\operatorname{diag}(\boldsymbol{p}^{(\ell)})\mathbf{\Gamma}_{m'}^{(\ell)\top}\\ &= \mathbf{\Gamma}_m^{(\ell)}\mathbf{P}^{(\ell)}\mathbf{\Gamma}_{m'}^{(\ell)\top} \qquad (7)\end{aligned}$$

where $\mathbf{P}^{(\ell)} := \operatorname{diag}(\boldsymbol{p}^{(\ell)})$ and we successively relied on the law of total probability, (1), and (5). Accordingly, it can be shown that the cross-correlation between annotators m, $m'\neq m$ and $m''\neq m,m'$, $\underline{\Psi}_{mm'm''}^{(\ell)} := \mathbb{E}\left[\mathbf{f}_m^{(\ell)}(X)\circ\mathbf{f}_{m'}^{(\ell)}(X)\circ\mathbf{f}_{m''}^{(\ell)}\right]$ forms a K-factor PARAFAC tensor [9]

$$\begin{aligned}\underline{\Psi}_{mm'm''}^{(\ell)} &= \sum_{k=1}^{K}p_k^{(\ell)}\boldsymbol{\gamma}_{m,k}^{(\ell)}\circ\boldsymbol{\gamma}_{m',k}^{(\ell)}\circ\boldsymbol{\gamma}_{m'',k}^{(\ell)} \qquad (8)\\ &= [[\mathbf{\Gamma}_m^{(\ell)}\mathbf{P}^{(\ell)},\mathbf{\Gamma}_{m'}^{(\ell)},\mathbf{\Gamma}_{m''}^{(\ell)}]]_K.\end{aligned}$$

Note here that the diagonal matrix $\mathbf{P}^{(\ell)}$ can multiply any of the factor matrices $\mathbf{\Gamma}_m^{(\ell)}, \mathbf{\Gamma}_{m'}^{(\ell)}$, or, $\mathbf{\Gamma}_{m''}^{(\ell)}$.

Let $\boldsymbol{\mu}_m^{(\ell)}, \mathbf{S}_{mm'}^{(\ell)}$ and $\underline{T}_{mm'm''}^{(\ell)}$ denote the sample counterparts of (6), (7) and (8) respectively. The law of large numbers dictates that as $N\to\infty$ these sample statistics converge to their true values. The following subsection will introduce an algorithm for recovering $\{\mathbf{\Gamma}_m^{(\ell)}\}_{m=1}^{M_\ell}$ and $\boldsymbol{p}^{(\ell)}$, $\ell=1,\dots,L$ from the statistics of annotator responses.

C. Confusion matrix estimation algorithm

Having available first-, second-, and third-order statistics of annotator responses for group ℓ, $\{\boldsymbol{\mu}_m^{(\ell)}\}_{m=1}^{M_\ell}$, $\{\mathbf{S}_{mm'}^{(\ell)}\}_{m,m'=1}^{M_\ell}$, and $\{\underline{T}_{mm'm''}^{(\ell)}\}_{m,m',m''=1}^{M_\ell}$, estimates of the confusion matrices can be readily extracted from them [cf. (7),(8)]. This procedure can be cast as the following constrained optimization problem

$$\min_{\substack{\{\mathbf{\Gamma}_m^{(\ell)}\in\mathcal{C}\}_{m=1}^{M_\ell}\\ \boldsymbol{p}^{(\ell)}\in\mathcal{C}_p}} g_\ell(\{\mathbf{\Gamma}_m^{(\ell)}\}_{m=1}^{M_\ell},\boldsymbol{p}^{(\ell)}) \quad (9)$$

where

$$\begin{aligned}g_\ell(\{\mathbf{\Gamma}_m^{(\ell)}\},\boldsymbol{p}^{(\ell)}) &:= \sum_{m=1}^{M_\ell}\|\boldsymbol{\mu}_m^{(\ell)}-\mathbf{\Gamma}_m^{(\ell)}\boldsymbol{p}^{(\ell)}\|_2^2\\ &+ \sum_{\substack{m=1\\ m'>m}}^{M_\ell}\|\mathbf{S}_{mm'}^{(\ell)}-\mathbf{\Gamma}_m^{(\ell)}\mathbf{P}^{(\ell)}\mathbf{\Gamma}_{m'}^{(\ell)\top}\|_F^2\\ &+ \sum_{\substack{m=1\\ m'>m,m''>m'}}^{M_\ell}\|\underline{T}_{mm'm''}^{(\ell)}-[[\mathbf{\Gamma}_m^{(\ell)}\mathbf{P}^{(\ell)},\mathbf{\Gamma}_{m'}^{(\ell)},\mathbf{\Gamma}_{m''}^{(\ell)}]]_K\|_F^2.\end{aligned}$$

We will solve the non-convex optimization in (9) using the alternating optimization method described in [24], which is guaranteed to converge to a stationary point of g_ℓ [12]. As2 is used here to address the permutation ambiguity that is induced by the tensor decomposition of (9). Interested readers are referred to [24] for algorithm and implementation details. Note that this approach is reminiscent of the method of moment estimators [17]. After obtaining estimates $\{\hat{\mathbf{\Gamma}}_m^{(\ell)}\}_{m=1}^{M_\ell}$, and $\hat{\boldsymbol{p}}^{(\ell)}$, estimates of the hidden variable $\{\hat{h}_\ell(x_n)\}_{n=1}^N$ can be obtained using the ML/MAP estimator described in (4); that is for $n = 1, \ldots, N$,

$$\hat{h}_\ell(x) = \operatorname*{arg\,max}_{k\in\{1,\ldots,K\}} \log \hat{p}_k^{(\ell)} + \sum_{m=1}^{M_\ell} \log(\hat{\Gamma}_m^{(\ell)}(f_m^\ell(x), k)). \quad (10)$$

Upon obtaining hidden variable estimates for all groups and data $\{\hat{h}_\ell(x_n)\}_{\ell=1,n=1}^{L,N}$, label estimates for all data $\{\hat{y}_n\}_{n=1}^N$ can be obtained. Similar to Sec. III-B, estimates of the hidden variables are represented in vector form as $\{\hat{\mathbf{h}}_\ell(x_n)\}_{\ell=1,n=1}^{L,N}$. Then, it can be shown that the mean hidden variable of group ℓ, the $K \times K$ cross-correlation between hidden variables ℓ and $\ell' \neq \ell$, $\mathbf{R}_{\ell\ell'} := \mathbb{E}[\mathbf{h}_\ell(X)\mathbf{h}_{\ell'}^\top(X)]$, and the third-order $K\times K\times K$ cross-correlation between hidden variables ℓ, $\ell' \neq \ell$ and $\ell'' \neq \ell', \ell$, $\underline{\Psi}_{\ell\ell'\ell''} := \mathbb{E}\left[\mathbf{h}_\ell(X) \circ \mathbf{h}_{\ell'}(X) \circ \mathbf{h}_{\ell''}(X)\right]$ are

$$\begin{aligned} \mathbb{E}[\mathbf{h}_\ell(X)] &= \mathbf{\Phi}_\ell \boldsymbol{\pi} \\ \mathbf{R}_{\ell\ell'} &= \mathbf{\Phi}_\ell \mathbf{\Pi} \mathbf{\Phi}_{\ell'}^\top \\ \underline{\Psi}_{\ell\ell'\ell''} &= [[\mathbf{\Phi}_\ell\mathbf{\Pi}, \mathbf{\Phi}_{\ell'}, \mathbf{\Phi}_{\ell''}]]_K \end{aligned} \quad (11)$$

respectively, where $\boldsymbol{\pi} := [\Pr(Y=1), \ldots, \Pr(Y=K)]^\top$ denotes the vector of class prior probabilities and $\mathbf{\Pi} = \operatorname{diag}(\boldsymbol{\pi})$. Afterwards, we solve a moment matching problem similar to (9) to compute estimates of the hidden variable confusion matrices and prior probabilities

$$\min_{\substack{\{\mathbf{\Phi}_\ell \in \mathcal{C}\}_{\ell=1}^L \\ \boldsymbol{\pi}\in\mathcal{C}_p}} g(\{\mathbf{\Phi}_\ell\}_{\ell=1}^L, \boldsymbol{\pi}) \quad (12)$$

where

$$g(\{\mathbf{\Phi}_\ell\}, \boldsymbol{\pi}) := \sum_{\ell=1}^L \|\boldsymbol{\mu}_\ell - \mathbf{\Phi}_\ell\boldsymbol{\pi}\|_2^2 + \sum_{\substack{\ell=1 \\ \ell'>\ell}}^L \|\mathbf{S}_{\ell\ell'} - \mathbf{\Phi}_\ell\mathbf{\Pi}\mathbf{\Phi}_{\ell'}^\top\|_F^2$$
$$+ \sum_{\substack{\ell=1 \\ \ell'>\ell,\ell''>\ell'}}^L \|\underline{T}_{\ell\ell'\ell''} - [[\mathbf{\Phi}_\ell\mathbf{\Pi}, \mathbf{\Phi}_{\ell'}, \mathbf{\Phi}_{\ell''}]]_K\|_F^2$$

and μ_ℓ, $\mathbf{S}_{\ell\ell'}$, $\underline{T}_{\ell\ell'\ell''}$ denote the sample counterparts of $\mathbb{E}[\mathbf{h}_\ell(X)], \mathbf{R}_{\ell\ell'}$ and $\mathbf{\Psi}_{\ell\ell'\ell''}$ respectively. As with (9), As3 is used here to address the permutation ambiguity introduced by the tensor decomposition of (12). Finally, with the estimated $\{\hat{\mathbf{\Phi}}_\ell\}$, and $\boldsymbol{\pi}$ at hand, estimates of data labels are obtained through an ML/MAP detector

$$\hat{y}(x) = \operatorname*{arg\,max}_{k\in\{1,\ldots,K\}} \log \hat{\pi}_k + \sum_{\ell=1}^L \log(\hat{\Phi}_\ell(h_\ell(x), k)). \quad (13)$$

The entire ensemble classification process is tabulated in Alg. 1.

Remark 1. The estimates $\{\hat{y}_n\}$ and $\{\hat{\mathbf{\Phi}}_\ell\}$ provided by Alg. 1 can also be employed to initialize the EM algorithm of [6].

Remark 2. Even though for this work, annotator dependency groups are presumed known, they can also be found using clustering techniques [13], [27].

Algorithm 1 Blind Multi-class Ensemble Classifier

```
Input: Annotator responses $\{f_m^{(\ell)}(x_n)\}_{m=1,n=1}^{M_\ell,N}, \forall\ell$
Output: Estimates of data labels $\{\hat{y}_n\}_{n=1}^N$;
1: for $\ell = 1, \ldots, L$ do
2:     Estimate $\boldsymbol{p}^{(\ell)}, \{\mathbf{\Gamma}_m^{(\ell)}\}_{m=1}^{M_\ell}$ via (9).
3:     Estimate $\hat{h}_\ell(x_n)$ via (10) for $n = 1, \ldots, N$.
4: end for
5: Estimate $\boldsymbol{\pi}, \{\mathbf{\Phi}_\ell\}_{\ell=1}^L$ via (12).
6: Estimate $\hat{y}_n$ via (13) for $n = 1, \ldots, N$.
```

IV. Numerical Tests

The performance of the proposed algorithm was evaluated using synthetic datasets. Using both MAP and ML detection in step 6, Alg. 1 is compared to majority voting (denoted as *MV*), the method of [16] (denoted as *KOS*), as well as the methods presented in [29] that initialize the EM algorithm (denoted as *EM+MV* for the method that uses majority voting for initialization and *EM+Spectral* for the method that uses tensor decomposition for initialization). 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 [19]. 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 $n = 1, \ldots, N$. Afterwards, annotators were grouped into L groups and $\mathbf{\Phi}_\ell$ and $\{\mathbf{\Gamma}_m^{(\ell)}\}_{m=1}^{M_\ell}$ per group ℓ were generated at random, such that $\mathbf{\Phi}_\ell \in \mathcal{C}, \mathbf{\Gamma}_m^{(\ell)} \in \mathcal{C}$, for all $m = 1, \ldots, M_\ell$. Then annotator responses are generated as follows: if $y_n = k$, then the hidden variable $h_\ell(x_n)$ will be generated randomly according to the k-th column of $\mathbf{\Phi}_\ell$, denoted as $\boldsymbol{\phi}_{\ell,k}$; that is $h_\ell(x_n) \sim \boldsymbol{\phi}_{\ell,k}$. Finally, if $h_\ell(x_n) = k'$ the response of annotator m in group ℓ will be generated randomly according to $\boldsymbol{\gamma}_{m,k'}^{(\ell)}$, that is $f_m^{(\ell)}(x_n) \sim \boldsymbol{\gamma}_{m,k'}^{(\ell)}$.

Fig. 2 shows the classification ER for a synthetic dataset with $K = 4$, $M = 100$ annotators belonging in $L = 10$ groups, with $M_\ell = 10$ for $\ell = 1, \ldots, 10$, for varying N. Here data were generated with $\boldsymbol{\pi} = [0.2951, 0.3281, 0.0460, 0.3308]^\top$. In addition, 7 hidden variable confusion matrices $\{\mathbf{\Phi}_\ell\}$ were generated to be better than random, while 3 were generated with completely random confusion matrices. In all groups, all annotator confusion matrices were generated to be better than random. Clearly, the proposed scheme (denoted as *Alg. 1 MAP* and *Alg. 1 ML*) outperforms majority voting, as well as the remaining methods that are designed to operate under full conditional independence. As N increases the ER of Alg. 1 decreases. This makes sense since as N increases, the sample statistics approach their ensemble counterparts [cf. Sec. III-B], enabling more accurate estimation of the confusion matrices. Fig. 3 shows the same experiment, but for fixed $N = 10^6$, and varying number of annotators M. Again, Alg. 1 markedly outperforms the competing alternatives, and its performance

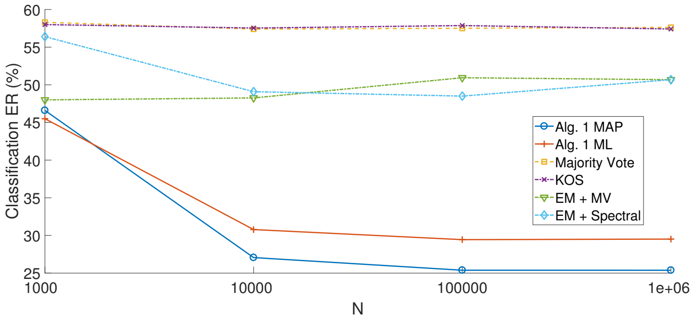

Fig. 2. Classification ER for a synthetic dataset with $K = 4$, $M = 100$, and $L = 10$.

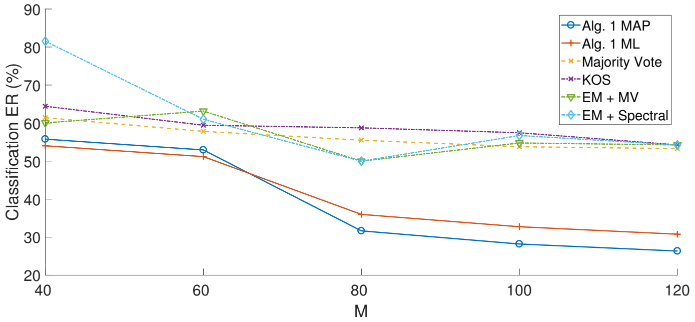

Fig. 3. Classification ER for a synthetic dataset with $K = 4$, $N = 10^6$ and $L = 10$.

increases with M. This result suggests that more annotators are preferable, when their dependencies are taken into account.

V. Conclusions

This paper introduced a novel approach to blind multi-class ensemble and crowdsourced classification that relies solely on the annotator responses to assess their quality and combine their answers, while also taking into account dependencies between them. Future research will focus on extensive numerical tests with real datasets, theoretical analysis of the proposed scheme, as well as algorithms that can infer groups of dependent annotators, along with distributed and online implementations.

References

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

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

[3] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[4] H. Chen, B. Chen, and P. K. Varshney, "A new framework for distributed detection with conditionally dependent observations," *IEEE Transactions on Signal Processing*, vol. 60, no. 3, pp. 1409–1419, March 2012.

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

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

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

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

[9] R. A. Harshman and M. E. Lundy, "PARAFAC: Parallel factor analysis," *Computational Statistics & Data Analysis*, vol. 18, no. 1, pp. 39–72, 1994.

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

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

[12] K. Huang, N. D. Sidiropoulos, and A. P. Liavas, "A flexible and efficient algorithmic framework for constrained matrix and tensor factorization," *IEEE Transactions on Signal Processing*, vol. 64, no. 19, pp. 5052–5065, 2016.

[13] A. Jaffe, E. Fetaya, B. Nadler, T. Jiang, and Y. Kluger, "Unsupervised ensemble learning with dependent classifiers," in *Artificial Intelligence and Statistics*, 2016, pp. 351–360.

[14] A. Jaffe, B. Nadler, and Y. Kluger, "Estimating the accuracies of multiple classifiers without labeled data." in *AISTATS*, vol. 2, San Diego, CA, 2015, p. 4.

[15] P. Jain and S. Oh, "Learning mixtures of discrete product distributions using spectral decompositions." *Journal of Machine Learning Research*, vol. 35, pp. 824–856, 2014.

[16] D. R. Karger, S. Oh, and D. Shah, "Efficient crowdsourcing for multi-class labeling," *ACM SIGMETRICS Performance Evaluation Review*, vol. 41, no. 1, pp. 81–92, 2013.

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

[18] ——, *Fundamentals of Statistical Signal Processing, volume II: Detection Theory*. Prentice Hall, 1998.

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

[20] 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.

[21] U. Shaham, X. Cheng, O. Dror, A. Jaffe, B. Nadler, J. Chang, and Y. Kluger, "A deep learning approach to unsupervised ensemble learning," in *International Conference on Machine Learning*, New York, NY, 2016, pp. 30–39.

[22] N. D. Sidiropoulos, L. De Lathauwer, X. Fu, K. Huang, E. E. Papalexakis, and C. Faloutsos, "Tensor decomposition for signal processing and machine learning," *IEEE Transactions on Signal Processing*, vol. 65, no. 13, pp. 3551–3582, 2017.

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

[24] P. A. Traganitis, A. Pagès-Zamora, and G. B. Giannakis, "Blind multi-class ensemble learning with unequally reliable classifiers," *arXiv preprint arXiv:1712.02903*, 2017.

[25] P. A. Traganitis, A. Pagès-Zamora, and G. B. Giannakis, "Learning from unequally reliable blind ensembles of classifiers," in *Proc. of the 5th IEEE Global Conference on Signal and Information Processing*. Montreal, CA: IEEE, 2017.

[26] P. K. Varshney, *Distributed Detection and Data Fusion*. Springer Science & Business Media, 2012.

[27] U. Von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.

[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.