Detecting adversaries in Crowdsourcing

Panagiotis A. Traganitis and Georgios B. Giannakis
Dept. of Electrical and Computer Engineering, University of Minnesota, MN, USA
emails: {traga003,georgios}@umn.edu

2021 IEEE International Conference on Data Mining (ICDM) | 978-1-6654-2398-4/21/$31.00 ©2021 IEEE | DOI: 10.1109/ICDM51629.2021.00174

***Abstract*—Despite its successes in various machine learning and data science tasks, crowdsourcing can be susceptible to attacks from dedicated adversaries. This work investigates the effects of adversaries on crowdsourced classification, under the popular Dawid and Skene model. The adversaries are allowed to deviate arbitrarily from the considered crowdsourcing model, and may potentially cooperate. To address this scenario, we develop an approach that leverages the structure of second-order moments of annotator responses, to identify large numbers of adversaries, and mitigate their impact on the crowdsourcing task. The potential of the proposed approach is empirically demonstrated on synthetic and real crowdsourcing datasets.**

***Index Terms*—Crowdsourcing, Classification, Adversaries, Ensemble learning.**

I. Introduction

Crowdsourcing has emerged as a powerful paradigm for tackling various machine learning, data mining, and data science tasks. Crowdsourcing, via services such as Amazon's Mechanical Turk [8] enlists inexpensive crowds of human workers, or annotators, to accomplish any given task. The focus of much research on crowdsourcing is centered on properly aggregating the noisy annotator labels, to obtain results as close to the ground-truth as possible. This is challenging due to the sparsity of annotator responses and the variability in annotator ability and effort. To add insult to injury, crowdsourcing is vulnerable to attacks by determined and coordinated adversaries, which provide erroneous responses aiming to reduce the performance of the overall system, or cause misclassification of specific data.

This paper puts forth a novel method for detecting arbitrary adversaries in crowdsourced classification. As a first step, we first analyze the structure of the correlation matrix of annotator responses, under the popular Dawid and Skene model. Afterwards, a subspace clustering-based approach is developed to split annotators into two groups. Honest and adversarial annotator groups are then distinguished by utilizing some additional side information. In this work, two types of side information are considered: a.) Knowledge of one trusted annotator, or b.) the assumption that the majority ($> 50\%$) of annotators are honest. Finally, a heuristic approach to prudently aggregate annotator responses is provided. Compared to other state-of-the-art approaches, the proposed method is based on the more general Dawid and Skene model, and can handle a potentially much larger number of adversaries. Additional details and experiments can be found in the long version of this paper [18][1].

Work in this paper was supported by NSF grants 1901134, 2126052, 2128593 and ARO-STIR grant 00093896.

Notation. Unless otherwise noted, lowercase bold letters, $\boldsymbol{x}$, denote column 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 $[\mathbf{X}]_{ij}$; vec$(\mathbf{X})$ denotes a vector consisting of the stacked columns of $\mathbf{X}$. The Frobenius and nuclear norms of a matrix $\mathbf{X}$ are denoted by $\|\mathbf{X}\|_F$ and $\|\mathbf{X}\|_*$ respectively. The rank of a matrix $\mathbf{X}$ is denoted by rank$(\mathbf{X})$ and diag$(\boldsymbol{x})$ denotes a diagonal matrix with the vector $\boldsymbol{x}$ on its diagonal. tr$(\mathbf{X})$ denotes the trace of matrix $\mathbf{X}$, that is the sum of the values on its diagonal. Pr denotes probability, or the probability mass function; $\sim$ denotes "distributed as;" $^\top$ represents transpose; card$(\mathcal{A})$ denotes the cardinality, i.e. the number of elements, of set $\mathcal{A}$; $\mathbb{E}[\cdot]$ denotes expectation, and $\mathbb{1}(\mathcal{A})$ is the indicator function for the event $\mathcal{A}$, that takes value 1 when $\mathcal{A}$ occurs, and 0 otherwise.

II. Problem statement and preliminaries

Consider a dataset consisting of N independent and identically distributed (i.i.d.) data $\{x_n\}_{n=1}^N$ 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. Class prior probabilities are collected in $\boldsymbol{\pi} := [\pi_1, \ldots, \pi_K]^\top = [\Pr(y_n = 1), \ldots, \Pr(y_n = K)]^\top$. An ensemble of M annotators or workers observe $\{x_n\}_{n=1}^N$, and provide noisy estimates of labels, with $g_m(x_n) \in \{1, \ldots, K\}$ denoting the label assigned to the n-th datum by annotator m. When an annotator does not provide a response for a datum x_n, we encode this by $g_m(x_n) = 0$. Given only the annotator responses $\{g_m(x_n), m = 1, \ldots, M\}_{n=1}^N$, *crowdsourced classification* seeks to properly aggregate information contained in annotator responses and estimate the ground-truth labels of the data $\mathbf{y} := [y_1, \ldots, y_N]^\top$.

The Dawid and Skene (DS) model [2] asserts that annotators have constant behavior and are conditionally independent given the (unknown) true label of a datum y_n, that is their errors are independent. Note that, this is reminiscent of a Naive Bayes model, conditioned on the true labels y_n. An annotators response $g_m(x_n)$ for a datum x_n, depends only on that datum and only through its label y_n. Further, under the Dawid and Skene model annotators are characterized by a so-called *confusion matrix*, that captures the statistical behavior of an annotator when presented with a datum from each class.

[1] http://arxiv.org/abs/2110.04117

978-1-6654-2398-4/21/$31.00 ©2021 IEEE
DOI 10.1109/ICDM51629.2021.00174

For an annotator m the $K \times K$ confusion matrix is denoted by $\mathbf{H}_m$, and has entries $h_{m,k,c} := [\mathbf{H}_m]_{k,c} = \Pr(g_m(x_n) = k|y_n = c)$. Clearly, entries of $\mathbf{H}_m$ are non-negative and its columns sum up to 1. Since responses of different annotators per datum n are presumed conditionally independent, given the ground-truth label y_n, the joint pmf of annotator responses for datum x_n is $\Pr(g_1(x_n) = k_1, \ldots, g_M(x_n) = k_M|y_n = c) = \prod_{m=1}^{M} \Pr(g_M(x_n) = k_m|y_n = c) = \prod_{m=1}^{M} h_{m,k_m,c}$.

If annotator confusion matrices and class priors are known, the label of datum n can be estimated using a maximum a posteriori (MAP) classifier, as $\hat{y}_n = \underset{c\in\{1,\ldots,K\}}{\arg\max} \log \pi_c + \sum_{m=1}^{M} \log(h_{m,g_m(x_n),c})$,where we used the conditional independence of the annotators, and the monotonicity of the logarithm. In realistic crowdsourcing scenaria, however, confusion matrices $\{\mathbf{H}_m\}_{m=1}^{M}$ and class priors $\boldsymbol{\pi}$ are unknown and have to be estimated.

A. Crowdsourcing with adversaries

Suppose now, that a subset $\mathcal{H} \subseteq \mathcal{M} = \{1, \ldots, M\}$, of $M_{\mathcal{H}} := \text{card}(\mathcal{H})$ annotators are honest, and a subset $\mathcal{A} \subset \mathcal{M}$, of $M_{\mathcal{A}} := \text{card}(\mathcal{A})$ are adversaries. The adversaries in $\mathcal{A}$ also observe the data $\{x_n\}_{n=1}^{N}$ and seek to undermine the crowdsourcing task. In order to ensure robust estimation of the ground-truth labels, one has to *detect the presence of these adversaries and take mitigating steps*. The following assumptions hold throughout the rest of the paper:

- **Assumption 1.** Honest annotators adhere to the Dawid and Skene model; that is, given the ground-truth label y_n of a datum x_n, the responses of annotators are conditionally independent. Additionally, honest annotators are better than random.
- **Assumption 2.** The number of honest annotators $\text{card}(\mathcal{H})$ is strictly greater than K^2, and they are distinct.
- **Assumption 3.** Adversarial annotators observe $\{x_n\}_{n=1}^{N}$ and deviate arbitrarily from the Dawid and Skene model.

Under the aforementioned assumptions, in this work we seek to *identify adversarial annotators* and if possible *mitigate* their impact on the crowdsourcing classification task, using only the available annotator responses.

Assumption 1 is fairly standard in crowdsourcing and enables estimating annotator parameters and label aggregation. As will be shown later, Assumption 2 is necessary in this context for distinguishing honest annotators from adversaries. Further, this assumption also indicates that the proposed approach can, in principle, tolerate up to $M_{\mathcal{A}} = M - K^2$ adversaries, which depending on M and K, may be much larger than the $M/2$ number of adversaries that is allowed by competing alternatives. Nevertheless, the number of tolerated adversaries will depend on the additional information employed in Sec. IV. Assumption 3 implies that adversaries take into consideration only the data, and not the responses of honest annotators. It does not place any further restriction on the behavior of the adversaries, and suggests that their behavior is captured by an *unknown* conditional pmf $p_{\mathcal{A}} := \prod_{n=1}^{N} \Pr\left(\{g_m(x_n) = k_{m,n}\}_{m\in\mathcal{A}} \middle| \{y_{n'} = k_{n'}\}_{n'=1}^{N}\right)$. Note here that adversaries are not necessarily conditionally independent with each other, and their responses may depend on all observed data. However, since they only observe the available data, they are considered conditionally independent from the honest workers.

B. Prior art

The simplest method for aggregating crowdsourced labels is majority voting, where the estimated label for a specific data point is the one most annotators agree on. This however, assumes that all annotators are of equal ability, which is, in many cases, unrealistic. The seminal paper of Dawid and Skene [2] proposed the aforementioned model and introduced an expectation maximization (EM) algorithm for estimating annotator confusion matrices and class priors, that is guaranteed to converge to a local optimum. Recent spectral methods use second- and third-order moments of annotator responses to infer confusion matrices and are often used to initialize the EM algorithm [6], [17], [23]. Other works, advocate simpler, but less expressive models, such as the "one-coin" model, where each annotator is characterized by a single parameter [4]. The work of [11] considered crowdsourced classification under the one-coin model as a rank-one matrix completion problem.

Regarding adversarial attacks in crowdsourcing, [13] modified the EM algorithm of [2] to detect and eliminate spammers during the label aggregation phase, whereas [16] proposed a spectral algorithm for detecting spammers before the aggregation phase. In the binary classification setting, [7] proposed a penalty based algorithm for detecting adversaries, and [9] considers arbitrary adversaries under the one-coin model. Recently, [10] introduced a rank-1 matrix completion algorithm for aggregating labels in the presence of adversaries, under the one-coin model. However, the three aforementioned works assumed that most annotators ($> 50\%$) are honest. Compared to current adversarial crowdsourcing approaches, this work introduces an algorithm that is based on the general Dawid and Skene model, and can potentially detect a large number of adversaries.

III. Annotator Correlation

Based on the Dawid and Skene model of Sec. II, we will first examine the structure of the second-order moments of honest and adversarial annotator responses. As annotator responses are categorical variables, the measure of correlation considered here is the probability of agreement, or agreement rate between two annotators $\sigma_{m,m'} := \Pr(g_m(x_n) = g_{m'}(x_n)), m, m' \in \mathcal{M}$. For the remainder of this section, we will also assume without loss of generality, that the first $M_{\mathcal{H}}$ annotators are honest, and the remaining are adversarial.

A. Correlation of honest annotators

Let $\mathbf{g}_m(x_n)$ denote the response of annotator m when observing datum x_n, in "one-hot" format, that is, if $g_m(x_n) = k$

then $\mathbf{g}_m(x_n) = \mathbf{e}_k$, where $\mathbf{e}_k$ denotes the canonical $K \times 1$ vector that has a one in its' k-th entry and zeroes elsewhere.

Invoking the law of total probability, the assumed conditional independence of annotators, and the definitions of Sec. II, the $K \times K$ co-occurrence matrix between annotators $m, m' \in \mathcal{H}$ is [17]

$$\mathbf{R}_{m,m'} := \mathbb{E}[\mathbf{g}_m(x_n)\mathbf{g}_{m'}(x_n)] = \mathbf{H}_m \text{diag}(\boldsymbol{\pi})\mathbf{H}_{m'}^\top. \quad (1)$$

From $\mathbf{R}_{m,m'}$, the probability of agreement $\sigma_{m,m'}$ between annotators $m, m' \in \mathcal{H}$ is

$$\begin{aligned}\sigma_{m,m'} &= \text{tr}\left(\mathbf{R}_{m,m'}\right) = \text{tr}\left(\mathbf{H}_m \text{diag}(\boldsymbol{\pi})\mathbf{H}_{m'}^\top\right) \quad (2)\\ &= \text{vec}(\text{diag}(\boldsymbol{\pi})^{1/2}\mathbf{H}_m^\top)^\top \text{vec}(\text{diag}(\boldsymbol{\pi})^{1/2}\mathbf{H}_{m'}^\top) = \boldsymbol{v}_m^\top \boldsymbol{v}_{m'}\end{aligned}$$

where we have used the properties of the trace and $\boldsymbol{v}_m := \text{vec}(\text{diag}(\boldsymbol{\pi})^{1/2}\mathbf{H}_m^\top)$ is a $K^2 \times 1$ vector. Eq. 2 in turn implies that the $M_\mathcal{H} \times M_\mathcal{H}$ agreement matrix between honest annotators $\boldsymbol{\Sigma}_\mathcal{H}$, with entries $[\boldsymbol{\Sigma}_\mathcal{H}]_{m,m'} = \sigma_{m,m'}, m, m' \in \mathcal{H}$ has a low-rank plus diagonal form $\boldsymbol{\Sigma}_\mathcal{H} = \mathbf{C}_\mathcal{H} + \mathbf{I}_\mathcal{H} = \mathbf{V}^\top\mathbf{V} + \mathbf{I}_\mathcal{H}$, where $\mathbf{I}_\mathcal{H}$ denotes the identity matrix of appropriate dimension, $\mathbf{C}_\mathcal{H} := \mathbf{V}^\top\mathbf{V}$, and $\mathbf{V} := [\boldsymbol{v}_1, \ldots, \boldsymbol{v}_{M_\mathcal{H}}]$ is a $K^2 \times M$ matrix. Finally, Assumption 2 asserts that $\text{rank}(\mathbf{C}_\mathcal{H}) = K^2$.

B. Correlation between honest and adversarial annotators

As mentioned in Sec. II-A the behavior of adversarial annotators is captured by an *unknown* joint pmf $p_\mathcal{A}$. Despite $p_\mathcal{A}$ being unknown, the conditional independence between the group of adversaries and honest annotators enables characterization of their cross-moments. Based on this conditional independence [cf. Sec.II-A], the co-occurrence matrix between an honest annotator $m \in \mathcal{H}$ and an adversary $m' \in \mathcal{A}$ is

$$\mathbf{R}_{m,m'} = \mathbb{E}[\mathbf{g}_m(x_n)\mathbf{g}_{m'}(x_n)] = \mathbf{H}_m \text{diag}(\boldsymbol{\pi})\tilde{\mathbf{H}}_{m'}^\top. \quad (3)$$

where we have used the law of total probability, the fact that data are i.i.d., and defined $\tilde{h}_{m,k,n} := [\tilde{\mathbf{H}}_m]_{k,n} = \sum_{\mathbf{c}_{-n}} \Pr(g_m(x_n) = k | \mathbf{y} = \mathbf{c}) \prod_{j \neq n} \Pr(y_j = c_j)$. In addition, $\mathbf{c}_{-n}$ is an $N - 1 \times 1$ vector containing $\{c_j\}_{j=1}^N$ except c_n. It is worth noting that, for the purposes of this work, we are not interested in estimating $\tilde{\mathbf{H}}_m$, but are merely employing them to discover the properties of the annotator agreement matrix.

Then, (3) yields $\sigma_{m,m'}$ between an honest annotator $m \in \mathcal{H}$ and an adversarial one $m' \in \mathcal{A}$ as $\sigma_{m,m'} = \text{tr}\left(\mathbf{H}_m \text{diag}(\boldsymbol{\pi})\tilde{\mathbf{H}}_{m'}^\top\right) = \boldsymbol{v}_m^\top \tilde{\boldsymbol{u}}_{m'}$, with $\boldsymbol{v}_m$ as defined in Sec. III-A and $\tilde{\boldsymbol{u}}_{m'} := \text{vec}\left(\text{diag}(\boldsymbol{\pi})^{1/2}\tilde{\mathbf{H}}_{m'}^\top\right)$. The agreement rate between all honest and adversarial annotators is then captured in the $M_\mathcal{H} \times M_\mathcal{A}$ matrix $\mathbf{C}_{\mathcal{H},\mathcal{A}} = \mathbf{C}_{\mathcal{A},\mathcal{H}}^\top$, with entries $[\mathbf{C}_{\mathcal{H},\mathcal{A}}]_{m,m'} = \boldsymbol{v}_m^\top \tilde{\boldsymbol{u}}_{m'}$ for $m \in \mathcal{H}, m' \in \mathcal{A}$. Thus, $\mathbf{C}_{\mathcal{H},\mathcal{A}} = \mathbf{V}^\top\tilde{\mathbf{U}}$, where $\tilde{\mathbf{U}} := [\tilde{\boldsymbol{u}}_1, \ldots, \tilde{\boldsymbol{u}}_{M_\mathcal{A}}]$, and $\text{rank}(\mathbf{C}_{\mathcal{H},\mathcal{A}}) \leq K^2$.

Bringing it all together, the $M \times M$ agreement matrix between all annotators, honest and adversarial, $\boldsymbol{\Sigma}$ has the following block form

$$\boldsymbol{\Sigma} = \mathbf{C} + \mathbf{I} = \left[\begin{array}{c|c} \mathbf{C}_\mathcal{H} & \mathbf{C}_{\mathcal{H},\mathcal{A}} \\ \hline \mathbf{C}_{\mathcal{A},\mathcal{H}} & \mathbf{C}_\mathcal{A} \end{array}\right] + \left[\begin{array}{cc} \mathbf{I}_\mathcal{H} & \\ & \mathbf{I}_\mathcal{A} \end{array}\right] \quad (4)$$

where the $M_\mathcal{A} \times M_\mathcal{A}$ matrix $\mathbf{C}_\mathcal{A}$ denotes the correlation between adversaries, $\mathbf{I}_\mathcal{A}$ is a $M_\mathcal{A} \times M_\mathcal{A}$ identity matrix, and $\mathbf{I}$ is a $M \times M$ identity matrix. Note that $[\mathbf{C}_\mathcal{H}, \mathbf{C}_{\mathcal{H},\mathcal{A}}]^\top = [\mathbf{V}^\top\mathbf{V}, \mathbf{V}^\top\tilde{\mathbf{U}}]^\top = \left(\mathbf{V}^\top[\mathbf{V}, \tilde{\mathbf{U}}]\right)^\top$. Thus, the $M_\mathcal{H}$ columns of $\mathbf{C}$ corresponding to honest workers will be of rank K^2, as long as $\text{rank}([\mathbf{V}, \tilde{\mathbf{U}}]) = K^2$. Finally, since $\text{rank}\left([\mathbf{C}_\mathcal{H}, \mathbf{C}_{\mathcal{H},\mathcal{A}}]^\top\right) \leq K^2$, $\mathbf{C}$ is a rank deficient matrix.

IV. Identifying adversaries

In this section, we will take advantage of the structure of the annotator agreement matrix, specifically $\mathbf{C}$ [cf. (4)], in order to develop a method to distinguish honest workers from adversaries in crowdsourcing. In a nutshell, the proposed method seeks annotators whose agreement matrix fits the low-rank model discussed in the previous section. These annotators are deemed honest, while the rest are considered adversaries.

A. Estimating $\mathbf{C}$

Given the $M \times M$ empirical agreement matrix $\hat{\boldsymbol{\Sigma}}$, we can decouple the diagonal matrix $\mathbf{I}$ and the rank deficient matrix $\mathbf{C}$, using robust principal component analysis (RPCA [1]) or robust matrix completion (RMC [14]) methods, that is

$$\begin{aligned}\{\hat{\mathbf{C}}, \hat{\mathbf{S}}\} = &\arg\min_{\mathbf{C},\mathbf{S}} \|\mathbf{C}\|_* + \lambda\|\text{vec}(\mathbf{S})\|_1 \quad (5)\\ &\text{subject to } \boldsymbol{\Omega} \circ \hat{\boldsymbol{\Sigma}} = \boldsymbol{\Omega} \circ (\mathbf{C} + \mathbf{S})\end{aligned}$$

where $\mathbf{S}$ is a sparse $M \times M$ matrix, $\lambda > 0$, $\boldsymbol{\Omega}$ is a binary $M \times M$ matrix, whose entries are equal to 1 if the corresponding entry of $\hat{\boldsymbol{\Sigma}}$ is observed and 0 otherwise, and $\circ$ denotes the Hadamard (element-wise) matrix product. The parameter λ trades off the low rank of $\mathbf{C}$ and the sparsity of $\mathbf{S}$, and here it is set to $\lambda = 1/\sqrt{\alpha M}$, where α is the percentage of observed entries in $\hat{\boldsymbol{\Sigma}}$ [1]. As the identity matrix $\mathbf{I}$ [cf. (4)] generally does not adhere to the low rank structure of $\mathbf{C}$, we expect it to be captured in $\mathbf{S}$. Additionally, $\mathbf{S}$ may capture any spurious correlations between annotators. The optimization problem in (5) is a convex problem that can be solved using off-the-shelf solvers, such as CVX [5].

B. Clustering annotators

With $\hat{\mathbf{C}}$ at hand, we now turn our attention to the task of detecting adversarial annotators. This task is equivalent to detecting the honest annotators, by identifying the columns of $\mathbf{C}$ corresponding to honest workers. Recall that, under Assumption 2 $[\mathbf{C}_\mathcal{H}, \mathbf{C}_{\mathcal{H},\mathcal{A}}]^\top$, will form a low dimensional subspace of dimension at most K^2. This prompts us to look into subspace clustering approaches, which are designed to group data drawn from a union of subspaces [19], to segment the annotators into two groups. In this work, we opt for the Elastic Net subspace clustering algorithm [22], which solves the following optimization problem

$$\min_\mathbf{Z} \|\hat{\mathbf{C}} - \hat{\mathbf{C}}\mathbf{Z}\|_F^2 + \rho\left(\rho_2\|\text{vec}(\mathbf{Z})\|_1 + \frac{1-\rho_2}{2}\|\mathbf{Z}\|_F^2\right), \quad (6)$$

where $\mathbf{Z}$ is a $M \times M$ coefficient matrix that captures the subspace structure of the columns of $\hat{\mathbf{C}}$, and $\rho, \rho_2 > 0$. After obtaining $\mathbf{Z}$ from (6), spectral clustering [20] is performed on $|\mathbf{Z}| + |\mathbf{Z}^\top|$, with $|\cdot|$ denoting element-wise absolute value, to

Algorithm 1 Crowdsourcing with adversaries

1: **Input:** Annotator correlation matrix $\hat{\mathbf{\Sigma}}$, number of classes K, ρ.
2: **Output:** $\hat{\mathbf{y}}$, estimated annotator groups $\hat{\mathcal{H}}, \hat{\mathcal{A}}$
3: Extract $\hat{\mathbf{C}}$ from $\hat{\mathbf{\Sigma}}$ using (5).
4: Cluster annotators by solving (6), obtain two groups of annotator indices $\mathcal{C}_1, \mathcal{C}_2$.
5: Determine sets of honest $\hat{\mathcal{H}}$ and adversarial workers $\hat{\mathcal{A}} = \mathcal{M} \setminus \hat{\mathcal{H}}$ using side information [cf. Sec. IV-B].
6: Aggregate labels of adversaries $\{g_m(x_n)\}_{n=1,m\in\hat{\mathcal{A}}}^{N}$. Denote fused labels as $\{\hat{t_n}\}_{n=1}^{N}$.
7: Aggregate labels of honest workers and fused adversary labels $\{g_m(x_n)\}_{n=1,m\in\hat{\mathcal{H}}}^{N} \cup \{\hat{t}_n\}_{n=1}^{N}$ to form the final estimated data labels $\{\hat{y}_n\}_{n=1}^{N}$.

obtain cluster assignments. Here, $\rho_2 = 0.95$ and ρ is tuned to find the cluster closest to a rank K^2 subspace.

Let $\mathcal{C}_1$ and $\mathcal{C}_2$ denote the annotator indices corresponding to the two clusters that resulted from subspace clustering of $\hat{\mathbf{C}}$. In order to categorize the two formed annotator groups into honest $\hat{\mathcal{H}}$ and adversarial $\hat{\mathcal{A}}$, some additional information is required. This information may be similar to what most prior works consider, that is, most annotators ($> 50\%$) are honest [7], [9], [10]. In such a case, the annotators deemed as honest are the ones forming the largest group, and are collected in $\hat{\mathcal{H}}$. The annotators deemed as adversaries are collected in $\hat{\mathcal{A}} = \mathcal{M} \setminus \hat{\mathcal{H}}$. Such an approach can tolerate up to $0.5M$ adversaries. Another type of side information, may be knowledge that one (or more) specific annotator m_H is trusted, or honest. In this case, the set of annotators deemed as honest is the one that contains the index m_H. This type of side information has the potential to allow for greater numbers of adversaries if annotators are grouped correctly. Both types of side information mentioned here will be tested in Sec. V.

C. Aggregating labels

Upon grouping annotators into honest and adversarial the final step involves aggregating the noisy labels. To extract any label information that may be present in the adversarial responses, a two-step heuristic label aggregation approach is outlined below.

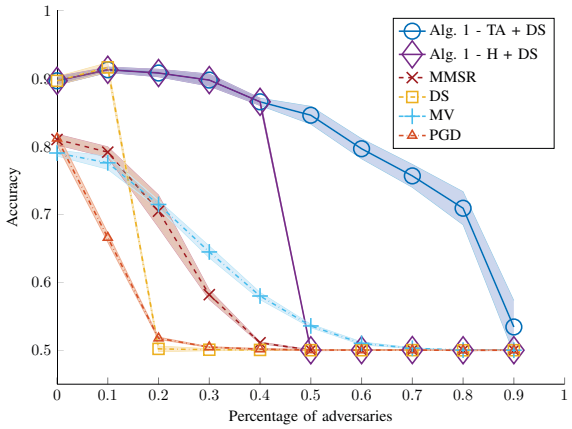
First, responses from annotators deemed adversarial (with indices in $\hat{\mathcal{A}}$) are aggregated using standard crowdsourcing techniques, e.g. the EM algorithm of [2]. This yields the aggregated labels $\{\hat{t}_n\}_{n=1}^{N}$, with $\hat{t}_n \in \{1, \ldots, K\}$. This step approximates the unknown pmf of adversaries as a Dawid and Skene model, and condenses their effect into the estimated $\hat{t}_n$'s. Second, to produce the final aggregated labels the responses of annotators deemed honest (with indices in $\hat{\mathcal{H}}$) alongside $\{\hat{t}_n\}_{n=1}^{N}$, from the first step, are aggregated using standard crowdsourcing algorithms, to produce the final estimated labels $\{\hat{y}_n\}_{n=1}^{N}$. In order to minimize the effect of misclassified annotators from the clustering stage, estimated labels are provided for data that have received a response from at least K different annotators. The entire algorithm for detecting adversaries and aggregating labels is tabulated in Alg. 1.

V. Numerical tests

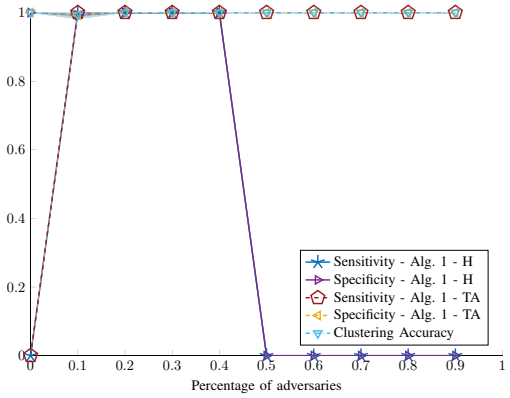
The performance of the proposed algorithm is validated in this section using synthetic and real datasets. Alg. 1 is compared against majority voting, denoted as *MV*, the EM algorithm of [2], denoted as *DS*, the matrix completion method for one-coin models of [11], denoted as *PGD*, and the state-of-the-art method for adversarial crowdsourcing of [10], denoted as *MMSR*. For these numerical tests, Alg. 1 uses *DS* at the aggregation stage. When considering that most annotators are honest, we will denote our approach as *Alg. 1 - H + DS*, whereas when we consider *one* trusted annotator, we will denote our approach as *Alg. 1 - TA + DS*. For *Alg. 1 - TA + DS* one randomly chosen honest worker is deemed trusted. The parameter ρ of Alg. 1 is selected using grid search [cf. Sec. IV-B] from the set $\{1.1, 2, 5, 10, 20, 100, 500, 800, 1000\}$. In all cases, *DS* is initialized using *MV*. All algorithms are compared in terms of classification accuracy, that is, the percentage of correctly classified data: Accuracy $= \frac{1}{N}\sum \mathbb{1}(\hat{y}_n = y_n)$. For the synthetic data tests, Alg. 1 is also evaluated on the performance of detecting adversaries, in terms of sensitivity (a.k.a. true positive rate, or recall) and specificity (true negative rate) [12], as well as clustering accuracy, that is, how accurately the groups of annotators are recovered. All algorithms were evaluated using MATLAB, all results represent the averages of 20 runs, and shaded areas indicate standard deviation around the average. Adversaries adopt strategies similar to the ones used in the numerical tests of [10]. Per run, $M_{\mathcal{A}} = \lfloor M p_{\text{adv}} \rfloor$ annotators are randomly selected to act as adversaries, with p_{adv} denoting the percentage of adversaries. A percentage of p_{corr} data are randomly selected to be corrupted by the adversaries. Adversaries provide the same wrong response for the corrupted data, and for the remaining $1-p_{\text{corr}}$ percentage of data they provide the ground-truth label.

A. Synthetic data

Here, a synthetic dataset with $M = 60$ annotators, $K = 3$ classes and $N = 5,000$ data points was randomly generated; labels y were drawn i.i.d. from $\boldsymbol{\pi} = 1/K\mathbf{1}$, and honest annotator confusion matrices $\{\mathbf{H}_m\}$ were randomly generated to satisfy Assumption 1. Using these confusion matrices honest annotator responses were generated, i.e. if $y_n = k$ $g_m(x_n)$ is drawn according to the k-th column of $\mathbf{H}_m$. All annotators (honest and adversarial) provide responses for data with probability $p_{\text{obs}} = 0.2$. Fig. 1 shows the results for this synthetic dataset as the percentage of adversaries p_{adv} varies. In this figure, the percentage of corrupted data is $p_{\text{corr}} = 0.5$. As the number of adversaries increases, overall classification performance drops. The classification accuracy *DS* and *PGD* drops quickly to 0.5, whereas *MV* and *MMSR* decline more gracefully, and provide better classification performance than *DS* and *PGD* up until 50% of the annotators are adversaries.



(a) Classification performance



(b) Detection performance

Fig. 1. Results for a synthetic dataset with varying percentage of adversaries

For up to $p_{\text{adv}} = 0.5$, *Alg. 1 - H* is the more robust of all algorithms. Impressively *Alg. 1 - TA* outperforms other algorithms for most percentages of adversaries, even though only *one* annotator is known a priori to be honest. Fig.2 shows results for the same dataset, but now the percentage of adversaries is fixed $p_{\text{adv}} = 0.3$, and the percentage of corrupted data p_{corr} varies. As the number of corrupted data increases, the performance of *DS* and *MV* decreases, while interestingly the performance of *MMSR* increases after approximately 60% are corrupted. We conjecture that large numbers of corrupted data, under this adversarial model, enable their detection, as their corresponding correlation increases. The performance of *Alg. 1* remains almost constant throughout, with both variants achieving high classification accuracy. In both scenaria, increasing number of adversaries and increasing number of corrupted data, adversaries are almost perfectly detected.

B. Real data

Further tests were conducted on real crowdsourcing datasets, namely the Bluebird [21] ($N = 108, M = 39, K = 2$), RTE [15] ($N = 800, M = 164, K = 2$), and Dog [3] ($N = 807, M = 109, K = 5$) datasets. Following the

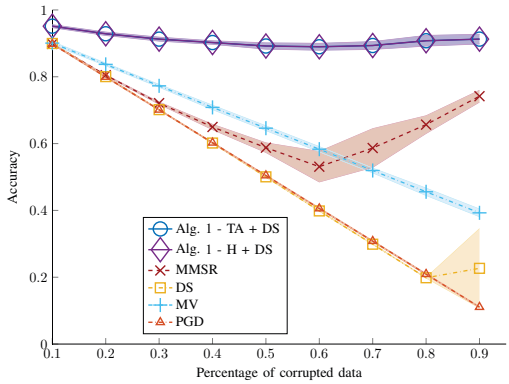


Fig. 2. Results for a synthetic dataset with varying percentage of corrupted data

numerical test strategy of [10], for all datasets, the percentage of corrupted data is fixed to $p_{\text{corr}} = 0.9$, and we vary the number of adversaries. Further, annotators that provided the same response for all data were removed from the datasets. Fig. 3 shows the classification accuracy, as the number of adversaries increases. Adversaries provide responses for data with probability $p_{\text{obs}} = 0.3$. Trends similar to those of the synthetic data tests can be observed. *MV, DS,* and *PGD* tolerate very few adversaries before starting to lose accuracy, in all datasets. *MMSR* outperforms the non-adversarially robust methods for approximately up to $p_{\text{adv}} = 0.5$. Both variants of Alg. 1 achieve high classification accuracy, with *Alg. 1 - TA* outperforming all other algorithms for the range of p_{adv} considered. *Alg. 1 - H* performs similarly to *Alg. 1 - TA* for up to $p_{\text{adv}} = 0.5$, as expected, and exhibits higher classification accuracy than *MMSR* in most cases. This is probably due to the use of the Dawid and Skene model in the derivation of *Alg. 1* instead of the one-coin model used in *MMSR*.

VI. Conclusions

This paper investigated crowdsourcing under adversarial attacks. A subspace clustering based algorithm was developed to detect adversaries and perform label aggregation, and its performance was evaluated on synthetic and real data.

Future research will involve theoretical analysis of the proposed method, alongside algorithms that can handle more advanced adversaries, enhanced label aggregation methods in the presence of adversaries, and online variants of the algorithm to handle streaming annotators and data.

References

[1] E. J. Candès, X. Li, Y. Ma, and J. Wright, "Robust principal component analysis?" *J. ACM*, vol. 58, no. 3, Jun. 2011.

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

[3] J. Deng, W. Dong, R. Socher, L. Li, Kai Li, and Li Fei-Fei, "Imagenet: A large-scale hierarchical image database," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, 2009, pp. 248–255.

[4] A. Ghosh, S. Kale, and P. McAfee, "Who moderates the moderators?: Crowdsourcing abuse detection in user-generated content," in *Proceedings of the 12th ACM Conference on Electronic Commerce*. San Jose, CA: ACM, 2011, pp. 167–176.

[5] M. Grant and S. Boyd, "CVX: Matlab software for disciplined convex programming, version 2.1," http://cvxr.com/cvx.

[6] S. Ibrahim, X. Fu, N. Kargas, and K. Huang, "Crowdsourcing via pairwise co-occurrences: Identifiability and algorithms," in *Advances in Neural Information Processing Systems*, vol. 32. Curran Associates, Inc., 2019, pp. 7847–7857.

[7] S. Jagabathula, L. Subramanian, and A. Venkataraman, "Identifying unreliable and adversarial workers in crowdsourced labeling tasks," *Journal of Machine Learning Research*, vol. 18, no. 93, pp. 1–67, 2017.

[8] 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*. Florence, Italy: ACM, 2008, pp. 453–456.

[9] M. Kleindessner and P. Awasthi, "Crowdsourcing with arbitrary adversaries," ser. Proceedings of Machine Learning Research, vol. 80. Stockholmsmässan, Stockholm Sweden: PMLR, 10–15 Jul 2018, pp. 2708–2717.

[10] Q. Ma and A. Olshevsky, "Adversarial crowdsourcing through robust rank-one matrix completion," in *Advances in Neural Information Processing Systems*, vol. 33. Curran Associates, Inc., 2020, pp. 21 841–21 852.

[11] Y. Ma, A. Olshevsky, C. Szepesvari, and V. Saligrama, "Gradient descent for sparse rank-one matrix completion for crowd-sourced aggregation of sparsely interacting workers," in *Proceedings of the 35th International Conference on Machine Learning*, ser. Proceedings of Machine Learning Research, vol. 80. PMLR, 10–15 Jul 2018, pp. 3335–3344.

[12] 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.

[13] V. C. Raykar and S. Yu, "Eliminating spammers and ranking annotators for crowdsourced labeling tasks," *Journal of Machine Learning Research*, vol. 13, no. 16, pp. 491–518, 2012.

[14] F. Shang, Y. Liu, J. Cheng, and H. Cheng, "Robust principal component analysis with missing data," in *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*. New York, NY, USA: Association for Computing Machinery, 2014, p. 1149–1158.

[15] R. Snow, B. O'Connor, D. Jurafsky, and A. Y. Ng, "Cheap and fast—but is it good? evaluating non-expert annotations for natural language tasks," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2008, p. 254–263.

[16] P. A. Traganitis and G. B. Giannakis, "Identifying spammers to boost crowdsourced classification," in *46th IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2021.

[17] P. A. Traganitis, A. Pagès-Zamora, and G. B. Giannakis, "Blind multiclass ensemble classification," *IEEE Transactions on Signal Processing*, vol. 66, no. 18, pp. 4737–4752, Sept 2018.

[18] P. A. Traganitis and G. B. Giannakis, "Detecting adversaries in crowdsourcing," *CoRR*, vol. abs/2110.04117, 2021.

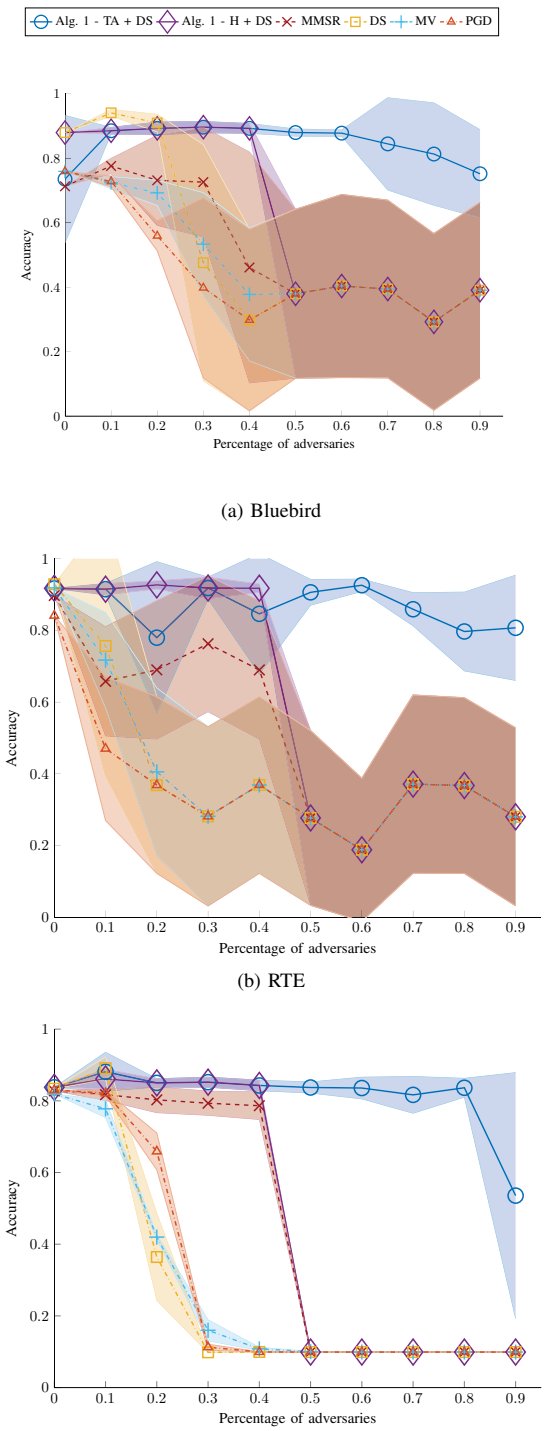
[19] R. Vidal, "A tutorial on subspace clustering," *IEEE Signal Process. Magazine*, vol. 28, no. 2, pp. 52–68, 2010.

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

[21] P. Welinder, S. Branson, P. Perona, and S. J. Belongie, "The multidimensional wisdom of crowds," in *Advances in Neural Information Processing Systems 23*. Curran Associates, Inc., 2010, pp. 2424–2432.

[22] C. You, C.-G. Li, D. P. Robinson, and R. Vidal, "Oracle based active set algorithm for scalable elastic net subspace clustering," in *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Las Vegas, NV, June 2016.

[23] 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.



(a) Bluebird

(b) RTE

(c) Dog

Fig. 3. Classification results for real crowdsourcing datasets with varying percentage of adversaries