The Thirty-Second International Florida
Artificial Intelligence Research Society Conference (FLAIRS-32)

An Empirical Evaluation of the Effect of Adversarial Labels on Classifier Accuracy Estimation *

Alexandra Clifford, Cassian Corey, John T. Holodnak
alexandra.clifford@ll.mit.edu, cassian.corey@ll.mit.edu, john.holodnak@ll.mit.edu
MIT Lincoln Laboratory

Abstract

This paper examines the effect of providing adversarial labels to several algorithms that use noisy labels from multiple experts to estimate classifier accuracy, referred to hereafter as "estimators." We propose four adversary labeling strategies and use experiments on synthetic data to gauge their impact on the estimators. Our results show that even a single adversary can considerably impact the effectiveness of an estimator. In addition, we find that estimators that weight the input of all experts equally tend to be much more affected by the inclusion of adversaries than those that can separately model each expert and that the impact of adversaries is lessened when the experts have higher accuracy.

1 Introduction

In machine learning, classifier performance is typically assessed by comparing the labels output by the classifier to the known true labels. However, as machine learning algorithms are increasingly deployed in real-world settings to tackle difficult problems, it has become common for the true labels to be unknown and difficult even for experts to determine. Recently, several algorithms, referred to throughout this work as "estimators," have been proposed to estimate classifier accuracy using noisy labels from multiple experts, see for example (Whitehill et al. 2009; Donmez, Lebanon, and Balasubramanian 2010; Li and Yu 2014; Jaffe, Nadler, and Kluger 2015; Lehner 2015).

Given that crowdsourcing platforms are commonly used to gather labels and that it is well known that crowdsourced workers exhibit large variation in expertise and subsequently in label accuracy (Whitehill et al. 2009) and furthermore may lazily or intentionally mislabel data (Whitehill et al. 2009; Difallah, Demartini, and Cudré-Mauroux 2012), it seems natural to consider what effect adversarial labels have on estimators. In addition, the problem is not necessarily limited to just datasets labeled by crowdsourced workers. Even when labels are collected in a less haphazard fashion, such as by querying a variety of subject matter experts, there is the potential for insider threats to maliciously mislabel data, perhaps with knowledge of the labels provided by other experts. As a consequence of these observations, there is a need to understand how adversarial labels affect estimators.

In this work, we study whether the error of several estimators in predicting classifier accuracy is affected by the presence of adversaries, if particular adversarial strategies that we introduce have more impact than others, and if any of the estimators appear robust to adversaries.

1.1 Related Work and Contributions

Recently, there has been considerable work in the area of adversarial machine learning, particularly with regard to image recognition problems and deep learning in general. For an overview, see the reviews by Akhtar and Mian (2018) and Yuan et al. (2017). Adversarial learning has also been studied for other problems including malware detection (Hu and Tan 2018) and voice recognition (Carlini et al. 2016).

To the best of our knowledge, however, there is no published work specifically analyzing the effectiveness of classifier accuracy estimation algorithms in the presence of an adversary. The closest parallels appear in the works of Parisi et al. (2014) and Whitehill et al. (2009). Parisi et al. (2014) investigated the effect of adversaries (referred to as a cartel) on their classifier ranking algorithm, but only through the lens of how well that ranking could be leveraged to predict the true label. Whitehill et al. (2009) devised a model for inferring the true label using noisy labels from multiple experts, which can also be used to estimate classifier accuracy, and investigated the effect of adversaries on the model's ability to infer the true label (but not on its ability to estimate the accuracy of the individual experts).

In addition to not directly addressing our question of interest, the works mentioned above considered only one or two models of adversary behavior. To be specific, Parisi et al. (2014) considered an adversary model in which a cartel of experts chooses labels with respect to a specific target, which may not be the true label, and Whitehill et al. (2009) considered adversaries that either randomly assign labels or always submit the wrong label.

*This research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Air Force Life Cycle Management Center Contract FA8702-15-D-0001. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government.
Copyright © 2019, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

In this paper, we propose four new adversary labeling strategies and evaluate their impact on several representative estimators. To briefly state our major contributions, we find that:

1. The addition of adversarial labels impacts all of the estimators we considered;
2. Estimators that equally weight experts tend to be much more affected by the inclusion of even a single adversary;
3. Increasing the accuracy of the experts (i.e. the non-malicious label-providing entities) tends to mitigate the impact of the adversaries.

1.2 Outline

We now provide an overview of the remainder of the paper. In Section 2, we provide a brief summary of the estimators considered in this paper. In Section 3, we define our notation and describe how the labels of the classifier, experts, and adversaries are generated. The results of our experiments are shown and discussed in Section 4. Finally, we conclude and offer recommendations for future work in Section 5.

2 Estimator Strategies

In this section, we describe the specific estimators considered in this paper and briefly mention several other estimators that appear in the literature.

Each estimator takes as input a set of classifier labels, that is the classifier's predicted labels for a given dataset, and a set of expert labels that are the experts' labels for the same dataset. The "classifier" is expected to be either an unsupervised or previously trained algorithm and the "experts" may be actual human experts, crowdsourced workers, or additional unsupervised or previously trained algorithms. Note that in our work, some of the experts may be adversaries. To be clear, some of the estimators output an estimate to the accuracy of each set of provided labels, but we consider only the problem of estimating the accuracy of the classifier.

Majority Vote (MV) Majority Vote uses the expert labels to determine a "best-guess" for each data instance (i.e., the label chosen by the most experts). Classifier accuracy is estimated by comparing the classifier labels to the best-guess labels.

Iterative Weighted Majority Vote (IWMV) Iterative Weighted Majority Vote (Li and Yu 2014) uses the classifier and expert labels to determine a best guess for each data instance and then estimates accuracy by comparing the labels of the classifier and experts to this best guess. In its first iteration, the best guess is assigned via unweighted majority voting. In each subsequent iteration, the classifier and expert accuracy estimates are used to inform a weighted vote.

Agreement (AGR) The estimator proposed by Lehner (2015) treats the classifier and experts separately. It first derives an approximation to expert accuracy via pair-wise agreement between experts on the dataset. It uses this estimate to in turn estimate the probability of each potential label, for each instance in the dataset. Finally, the amount of agreement between the most probable labels and the classifier labels is used (indirectly) to estimate classifier accuracy.

Covariance (COV) The approach of Jaffe, Nadler, and Kluger (2015) is an extension of the ranking algorithm by Parisi et al. (2014), which exploits structure in the classifier/expert covariance matrix to rank the classifier/experts according to their balanced accuracies. Jaffe, Nadler, and Kluger (2015) extended the method to estimate the true positive and true negative rates, and also the class prior.

Maximum Likelihood (MLE) Several authors have investigated approximating the maximum likelihood estimate of classifier accuracy via the Expectation-Maximization algorithm, see for example (Dawid and Skene 1979; Donmez, Lebanon, and Balasubramanian 2010; Sinha, Rao, and Balasubramanian 2018). Donmez, Lebanon, and Balasubramanian (2010) and Sinha, Rao, and Balasubramanian (2018) perform inference in a model with one parameter for each classifier/expert, while Dawid and Skene (1979) model the entire confusion matrix. In our implementation, we use the one-parameter model.

Other Estimators A number of other estimators have been proposed but are not examined in this paper. For example, Whitehill et al. (2009) model both the accuracy of experts and the difficulty of data instances. Platanios, Blum, and Mitchell (2014) derive an optimization-based estimator that makes extensive use of constraints based on relationships between agreement and accuracy. This estimator is superficially similar to AGR. Platanios, Dubey, and Mitchell (2016) also propose several graphical models that encode relationships between experts.

For brevity, we include only the estimators described in detail above. We feel that they cover enough of the spectrum of estimation approaches to show the importance of considering adversarial labels and understand their general effects.

3 Experimental Design

In this section, we describe the specifics of our problem setting as well as how we generate the labels of the classifier, experts, and adversaries.

3.1 Problem Setting

Let $\mathcal{D} \subset \mathcal{X}$ be a dataset of N instances x_i, $1 \leq i \leq N$, with unknown true labels $y_i \in \mathcal{Y}$, $1 \leq i \leq N$. In this paper, we assume for simplicity a binary classification problem and as a result $\mathcal{Y} = \{0, 1\}$.

Let $E = \{e_1, \ldots, e_{N_E}\}$ be the set of experts and let $A = \{a_1, \ldots, a_{N_A}\}$ be the set of adversaries. We denote the classifier and expert labels for instance x_i as $\hat{y}_i^C$ and $\hat{y}_i^{e_j}$, $1 \leq j \leq N_E$, respectively. We denote the adversary labels as $\hat{y}_i^{a_j}$, $1 \leq j \leq N_A$. We assume for added simplicity that each instance has been labeled by the classifier we wish to evaluate, by each expert, and by each adversary.

In our simulations, the classifier and expert labels are sampled from a simple Bayesian Network in which the classifiers and experts are conditionally independent, given

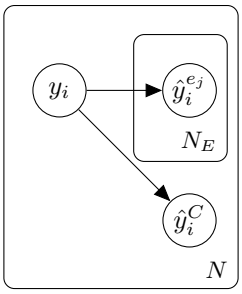

Figure 1: Generative model for the classifier and experts.

the true label. The Bayesian Network is depicted in Figure 1. Since we are considering the binary case, the classifier and the experts each have two associated parameters, $P(\hat{y} = 1 \,|\, y = 1)$ and $P(\hat{y} = 0 \,|\, y = 0)$, which we refer to as the class-conditional accuracies. After the classifier and expert labels have been generated, we apply one of the adversary strategies (described next) to generate the labels of the adversaries.

The goal of the estimators is to estimate the accuracy α^C of the classifier on $\mathcal{D}$. The accuracy estimators discussed in Section 2 take as input the labels predicted by the classifier and the experts, that is

$$\{(\hat{y}_i^C, \hat{y}_i^{e_1}, ..., \hat{y}_i^{e_{N_E}}, \hat{y}_i^{a_1}, ..., \hat{y}_i^{a_{N_A}})\}_{i=1}^N,$$

and output an approximation $\hat{\alpha}^C$ to the classifier accuracy.

3.2 Adversary Strategies

We now define four adversary labeling strategies.

Noisily Wrong In the noisily wrong strategy, the adversary attempts to pick the wrong label for each instance. We model this as an expert with class-conditional accuracies that are worse than random, that is, we set $P(\hat{y}^{a_j} = 1 \,|\, y = 1), 1 \leq j \leq N_A$ and $P(\hat{y}^{a_j} = 0 \,|\, y = 0), 1 \leq j \leq N_A$, so that they are less than 0.5. For each data instance, given the true label, the adversary predictions are drawn from a Bernoulli distribution parameterized by $P(\hat{y}_j^a = 1 \,|\, y = 1)$ if $y = 1$ and $P(\hat{y}_j^a = 0 \,|\, y = 0)$ if $y = 0$. Note that the noisily wrong adversaries are conditionally independent of the classifier/experts. A version of this strategy is mentioned by Parisi et al. (2014).

Random Target In this strategy, adversaries each randomly choose a target expert with which to intentionally disagree. The target expert is chosen uniformly at random from the set of experts. Then, labels are provided to perfectly disagree with those provided by the target expert. That is, for adversary j, an expert e is randomly chosen, and then, for each data instance, we set $y_i^{a_j} := 1 - y_i^e$.

Strategic Target Similar to the random target strategy, the adversaries apply labels opposite to one of the experts' labels. However, instead of randomly choosing an expert with which to disagree, the adversaries target the best expert.

Stealth The stealth strategy models an adversary seeking to avoid detection. The adversary first attempts to generate a correct label, and then, after observing the predicted labels of the classifier, experts, and previous adversaries, will change the label if they can cause or break a tie. As such, they have class-conditional accuracy parameters $P(\hat{y}^{a_j} = 1 \,|\, y = 1), 1 \leq j \leq N_A$ and $P(\hat{y}^{a_j} = 0 \,|\, y = 0), 1 \leq j \leq N_A$.

To describe the strategy specifically for the jth adversary, let $V_i^e(z)$ be the number of experts such that $\hat{y}_i^e = z$ and let $V_i^C(z)$ be 1 if $\hat{y}_i^C = z$, where $z \in \{0, 1\}$. In addition, let $V_i^{a_1:a_{j-1}}(z)$ be the number of adversaries (among the first $j-1$) such that $\hat{y}_i^a = z$. The adversary label for each data instance is then chosen in two steps. First, the adversary makes an initial prediction $\hat{y}_i^*$ by drawing from their own class-conditional distribution, that is, a Bernoulli distribution parameterized by $P(\hat{y}_i^{a_j} = 1 \,|\, y_i = 1)$ if $y_i = 1$ and $P(\hat{y}_i^{a_j} = 0 \,|\, y_i = 0)$ if $y_i = 0$. The adversary then observes the classifier, experts, and adversaries $1, \ldots, j-1$ and adjusts its label if it can cause or break a tie. Specifically,

$$\hat{y}_i^{a_j} = \begin{cases} 1 & \text{if } V_i^C(1) + V_i^E(1) + V_i^{a_1:a_{j-1}}(1) \\ & \quad = V_i^C(0) + V_i^E(0) + V_i^{a_1:a_{j-1}}(0) \\ 1 & \text{if } V_i^C(1) + V_i^E(1) + V_i^{a_1:a_{j-1}}(1) \\ & \quad = V_i^C(0) + V_i^E(0) + V_i^{a_1:a_{j-1}}(0) - 1 \\ 0 & \text{if } V_i^C(1) + V_i^E(1) + V_i^{a_1:a_{j-1}}(1) - 1 \\ & \quad = V_i^C(0) + V_i^E(0) + V_i^{a_1:a_{j-1}}(0) \\ y_i^* & \text{otherwise.} \end{cases}$$

3.3 Simulation Parameters

We now describe some of the specific parameters that we use to generate our simulated datasets. Every time a new dataset is generated, we first set the class prior distribution. Specifically, $P(y = 1) \sim U(0, 1)$ and $P(y = 0) = 1 - P(y = 1)$. In all simulations, we set the number of data instances at $N = 500$. The number of labelers is fixed at seven. This number includes the classifier, the experts, and the adversaries (if any are present).

Independently for each class and expert, we choose each class-conditional accuracy, $P(\hat{y} = 1 \,|\, y = 1)$ and $P(\hat{y} = 0 \,|\, y = 0)$, uniformly at random from a small interval, $[\theta - 0.10, \theta + 0.10]$. We vary the value of θ, which we refer to as the base accuracy, for the experts in the simulations. In the noisily wrong adversary model, we use $\theta = 0.20$ as the base accuracy. In the stealth model, we use the same value of θ as the experts. For the classifier, the class-conditional accuracies are set to 0.90.

4 Results and Discussion

In this section we generate simulated datasets as described in Section 3 and apply the estimators described in Section 2.

4.1 Effect of Adversary Strategies

In our first set of experiments, we compare the effect of the different adversary strategies on the estimators. We use $\theta = 0.80$ to set the base accuracy of the experts. We vary the number of adversaries from zero to five. To be clear,

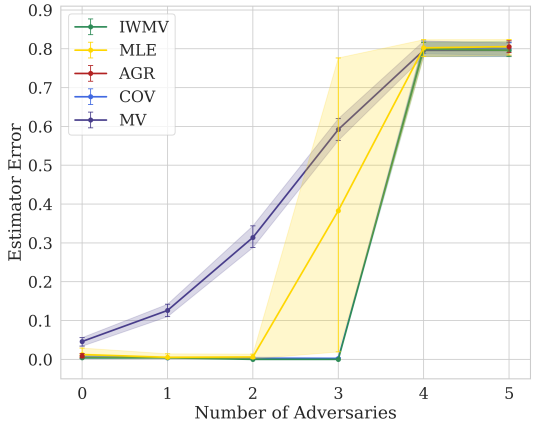

Figure 2: Noisily Wrong: Error $|\hat{\alpha}^C - \alpha^C|$ of the estimators against the number of adversaries when the adversaries apply the noisily wrong strategy. The plotted points represent the median absolute error over 100 trials while the error bands represent the 25th and 75th percentiles.

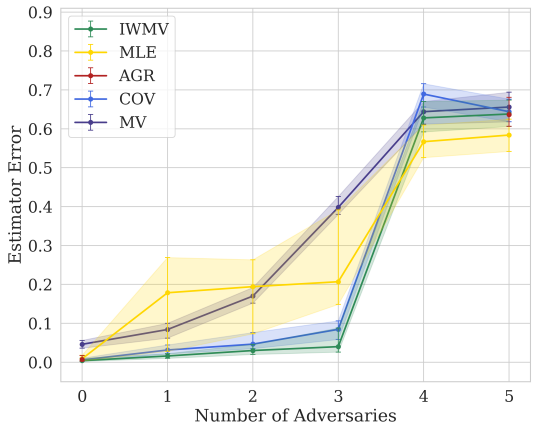

Figure 3: Random Target: Error $|\hat{\alpha}^C - \alpha^C|$ of the estimators against the number of adversaries when the adversaries apply the random target strategy. The plotted points and error bands are as in Figure 2.

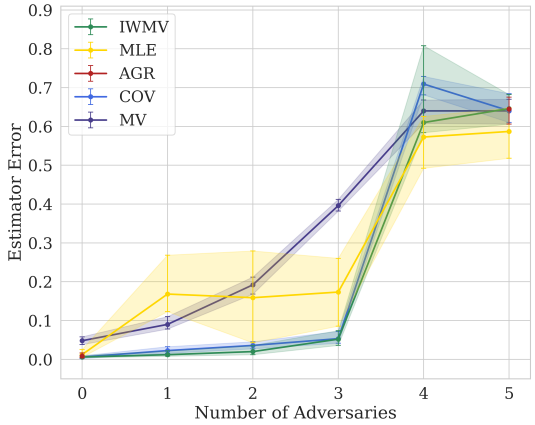

Figure 4: Strategic Target: Error $|\hat{\alpha}^C - \alpha^C|$ of the estimators against the number of adversaries when the adversaries apply the strategic target strategy. The plotted points and error bands are as in Figure 2.

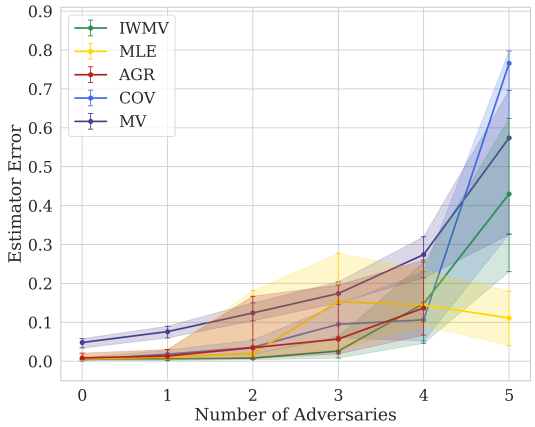

Figure 5: Stealth: Error $|\hat{\alpha}^C - \alpha^C|$ of the estimators against the number of adversaries when the adversaries apply the stealth strategy. The plotted points and error bands are as in Figure 2.

the number of experts decreases as adversaries are added (i.e., when there are zero adversaries, there are six experts and when there is one adversary, there are five experts). For each number of adversaries, we perform 100 separate trials. In many cases when adversaries were added, AGR failed to output an estimate of classifier accuracy, due to violations of its underlying assumptions. As a result, in the following figures, we show the results for AGR only when it succeeded for at least 90 of the 100 trials.

In Figure 2, we show the results for the noisily wrong adversary strategy. It is clear that MV and AGR are most affected, as the error of MV jumps considerably when a single adversary is added and AGR often fails to return an estimate. On the other hand, the errors of IWMV, MLE, and COV remain small until there are either three or four adversaries.

In Figures 3 and 4, we show the results for the random target and strategic target adversary strategies. The results are very similar, most likely because the difference between the strategies is related to which expert the adversaries target, and in this set of experiments, the difference between the best and worst experts may be rather small. As in the results for the noisily wrong strategy, we see AGR often failed to return an estimate. The error of MV increases as the number of adversaries increases. The error of MLE increases when a single adversary is added, but then stays roughly the same

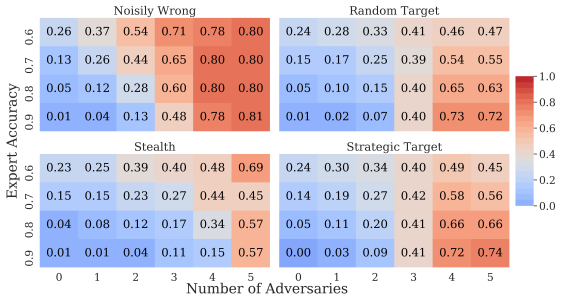

Figure 6: MV: Error $|\hat{\alpha}^C - \alpha^C|$ of MV for each of the four adversary strategies and different combinations of number of adversaries and expert accuracy. The value in each cell is the median error (rounded to two decimal places) over 100 trials.

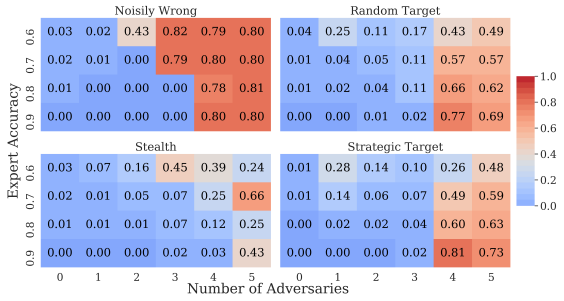

Figure 7: IWMV: Error $|\hat{\alpha}^C - \alpha^C|$ of IWMV for each of the four adversary strategies and different combinations of number of adversaries and expert accuracy. The cell values are as in Figure 6.

until a fourth adversary is added. Both COV and IWMV perform well until there are four adversaries.

Finally, in Figure 5, we show the results for the stealth adversary strategy. For all estimators, the stealth strategy has less of an effect than the other strategies. Across the number of adversaries, MV is usually most affected by the stealth strategy, perhaps because the stealth strategy was specifically designed to change the majority vote of the experts. The errors of the other methods slowly grow as more adversaries are added. The one exception is MLE, the error of which decreases slightly from four to five adversaries.

Across the strategies, then, it is clear that MV and AGR are more strongly affected by the presence of adversaries than IWMV, MLE, or COV. The error of MV rises considerably when there is even a single adversary and AGR often fails to return a value, due to some of its underlying assumptions not being satisfied. On the other hand, IWMV and COV are usually less affected until a third or even fourth adversary is added. MLE is on par with IWMV and COV for the noisily wrong and stealth adversary strategies, but performs worse for the random and strategic target strategy. Note that IWMV, MLE, and COV handle the noisily wrong adversary strategy particularly well and have very small errors until a third or fourth adversary is added. This is likely because the noisily wrong adversaries better conform to the statistical model assumed by the estimators in that they are conditionally independent of the experts.

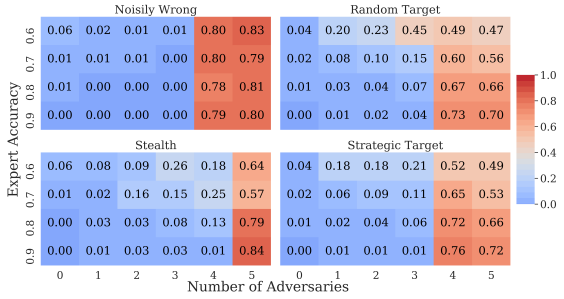

Figure 8: COV: Error $|\hat{\alpha}^C - \alpha^C|$ of COV for each of the four adversary strategies and different combinations of number of adversaries and expert accuracy. The cell values are as in Figure 6.

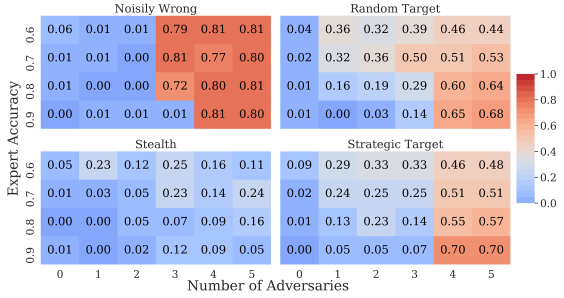

Figure 9: MLE: Error $|\hat{\alpha}^C - \alpha^C|$ of MLE for each of the four adversary strategies and different combinations of number of adversaries and expert accuracy. The cell values are as in Figure 6.

4.2 Effect of Expert Accuracy

Recall that in the previous set of experiments, we set the base expert accuracy at $\theta = 0.80$. Here, we vary it by setting it to $\theta = 0.60$, $\theta = 0.70$, $\theta = 0.80$, and $\theta = 0.90$, to gauge the effect of expert accuracy on estimator error.

Figures 6, 7, 8, and 9 show the results of this experiment for MV, IWMV, COV, and MLE. AGR often failed to return an estimate, except when there was a single adversary and the expert base accuracy was $\theta = 0.90$, so we omit its results here. For smaller numbers of adversaries, it is clear that increasing the accuracy of the experts tends to decrease the error of the estimators. This pattern is most noticeable for MV, which can clearly tolerate more adversaries when the accuracy of the experts is high. Across expert accuracies and number of adversaries, COV and IWMV were less impacted, in general, by adversaries. Both usually maintain reasonably low error rates until a fourth adversary is added and are more robust to low expert accuracy. MLE generally

performed fairly well, but tended to have larger errors than COV or IWMV, especially for the random and strategic target adversary strategies.

As a brief note, recall that we set the classifier accuracy at 0.90 in all experiments. We also performed the experiments described in this section with different classifier accuracies, but found that while the magnitude of errors changed somewhat, the overall trends were similar.

5 Conclusion

In this work, we proposed four adversary labeling strategies and used them to assess the impact of adversarial labels on several classifier accuracy estimation algorithms. We now review our major findings.

First, our experiments clearly show that there is a relationship between the presence of adversarial labels and estimator performance. The presence of adversaries generally causes an increase in estimator error. Depending on the adversary strategy and the accuracy of the experts, our findings show that even a single adversary can considerably increase the error of each estimator.

Second, we find that the estimators were affected differently both by the presence of adversaries, and by the strategy employed by those adversaries. In general, IWMV, MLE, and COV are less affected by the presence of adversaries than either AGR or MV. AGR in particular was severely impacted by the presence of adversaries. In our simulations (apart from the stealth strategy), AGR only successfully completed in the presence of an adversary if there was only a single adversary and expert accuracy was high. We hypothesize that this sharp difference in the methods is caused by the fact that AGR and MV weight all expert input equally. On the other hand, IWMV, COV, and MLE all infer the accuracy of each expert and as such, are able to "down-weight" the input of the adversaries, up until the point that there are about as many adversaries as experts.

Finally, we find that, for all estimators, more accurate experts help to mitigate the effect of adversaries.

5.1 Future Work

There are several opportunities for future work in this area. Since our results are only for simulated data, it would be interesting to investigate the effects of adversarial labeling on real datasets, perhaps by injecting simulated adversarial labels into them.

Given the apparent impact of adversarial labels on the estimators, we might also consider whether it is possible to design estimator methods that are more robust to adversaries. One potential idea is to combine algorithms for detecting adversaries, such as those by Jagabathula, Subramanian, and Venkataraman (2017), with the estimators.

References

Akhtar, N., and Mian, A. 2018. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access* 6:14410–14430.

Carlini, N.; Mishra, P.; Vaidya, T.; Zhang, Y.; Sherr, M.; Shields, C.; Wagner, D.; and Zhou, W. 2016. Hidden voice commands. In *25th USENIX Security Symposium*, 513–530.

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

Difallah, D. E.; Demartini, G.; and Cudré-Mauroux, P. 2012. Mechanical cheat: Spamming schemes and adversarial techniques on crowdsourcing platforms. In *CrowdSearch*, 26–30.

Donmez, P.; Lebanon, G.; and Balasubramanian, K. 2010. Unsupervised supervised learning I: Estimating classification and regression errors without labels. *Journal of Machine Learning Research* 11(Apr):1323–1351.

Hu, W., and Tan, Y. 2018. Black-box attacks against rnn based malware detection algorithms. In *AAAI Workshops*.

Jaffe, A.; Nadler, B.; and Kluger, Y. 2015. Estimating the accuracies of multiple classifiers without labeled data. In *Artificial Intelligence and Statistics*, 407–415.

Jagabathula, S.; Subramanian, L.; and Venkataraman, A. 2017. Identifying unreliable and adversarial workers in crowdsourced labeling tasks. *Journal of Machine Learning Research* 18(93):1–67.

Lehner, P. E. 2015. Estimating the accuracy of automated classification systems using only expert ratings that are less accurate than the system. *Journal of Modern Applied Statistical Methods* 14(1):13.

Li, H., and Yu, B. 2014. Error Rate Bounds and Iterative Weighted Majority Voting for Crowdsourcing. *ArXiv e-prints*.

Parisi, F.; Strino, F.; Nadler, B.; and Kluger, Y. 2014. Ranking and combining multiple predictors without labeled data. *Proceedings of the National Academy of Sciences* 111(4):1253–1258.

Platanios, E. A.; Blum, A.; and Mitchell, T. M. 2014. Estimating accuracy from unlabeled data. In *Conference on Uncertainty in Artificial Intelligence*, 682–691.

Platanios, E. A.; Dubey, A.; and Mitchell, T. 2016. Estimating accuracy from unlabeled data: A bayesian approach. In *International Conference on Machine Learning*, 1416–1425.

Sinha, V. B.; Rao, S.; and Balasubramanian, V. N. 2018. Fast Dawid-Skene: A Fast Vote Aggregation Scheme for Sentiment Classification. In *Workshop on Issues of Sentiment Discovery and Opinion Mining*.

Whitehill, J.; Wu, T.-f.; Bergsma, J.; Movellan, J. R.; and Ruvolo, P. L. 2009. Whose vote should count more: Optimal integration of labels from labelers of unknown expertise. In *Advances in Neural Information Processing Systems*, 2035–2043.

Yuan, X.; He, P.; Zhu, Q.; Bhat, R. R.; and Li, X. 2017. Adversarial examples: Attacks and defenses for deep learning. *ArXiv e-prints*.