Classification with Strategic Data Sources

Yang Liu and Yiling Chen
{yangl, yiling}@seas.harvard.edu
Harvard University

Abstract

We consider a binary classification problem where the labels of training data are elicited from crowd workers who are strategic at exerting costly effort for coming up with a label or at reporting the resultant label. The goal of the learner is to learn an un-biased classifier that minimizes 0-1 loss. For a non-strategic setting, when the error rates of the two classes in the training data are known, Natarajan *et al.* [2013] have developed techniques to learn an un-biased classifier from the noisy training data. Moreover, they have shown that the smaller the error rates are, the better the performance of the classifier. When data sources are strategic, the error rates of the two classes are a result of their strategic behavior. We hence need to overcome several interwinding challenges: How to incentivize effort exertion so that workers produce labels with low errors? How to incentivize workers to truthfully report the labels that they have obtained? Even when workers follow a known strategy of effort exertion and label reporting, the error rates of the two classes are unknown to the learner. How to estimate the error rates so that the method of Natarajan *et al.* [2013] can be adapted to train an unbiased classifier?

We solve this problem by proposing a peer-prediction-style mechanism, where a worker's reported label of a data point is scored against the label given by a reference classifier that is constructed using techniques similar to those of Natarajan *et al.* [2013]. The mechanism has several desirable properties: (1) Every worker exerting effort and truthfully reporting is a Bayesian Nash equilibrium (BNE). (2) Colluding on reporting the same label disregard of their obtained labels is not an equilibrium strategy. (3) The obtained classifier's 0-1 loss converges to the optimal unbiased one. (4) We do not need completely redundant assignment of tasks; we assign a relatively much smaller number of tasks to more than one worker in our mechanism, in contrast to reassigning all tasks for many peer-prediction mechanisms. (5) Our mechanism is also robust to permutation strategies. In addition to the above technical merits, we view our work as an initial move towards using machine learning techniques for designing peer prediction mechanisms.

1 Introduction and formulation

Suppose the learner is targeting a decision function $f : \mathcal{X} \to \{-1, +1\}$ for a binary classification task, that takes a feature vector $\mathbf{x} \in \mathcal{X}$ as input and outputs a prediction on label y. $(\mathbf{x}, y)$ has an unknown joint distribution $\mathcal{D}$. Denote the 0-1 loss of f as $R_{\mathcal{D}}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[1(sgn(f(\mathbf{x})) \neq y)]$, and the minimum risk over concept class $\mathcal{F}$ as $R^* := \min_{f\in\mathcal{F}} R_{\mathcal{D}}(f)$. We define $f^* := \text{argmin}_{f\in\mathcal{F}} R_{\mathcal{D}}(f)$. The learner would like to obtain a f such that $R_{\mathcal{D}}(f)$ approaches R^* as the size of the training data grows. Before running an algorithm, the learner needs to collect a set of training data. Suppose a total of N data points are independently drawn according to $\mathcal{D}$. The learner however only knows their feature vectors $\{\mathbf{x}_i\}_{i=1}^N$ but not the corresponding labels $\{y_i\}_{i=1}^N \in \{-1, +1\}^N$. The learner resorts to a crowd of workers for labeling these data.

We have all-together T homogeneous crowd workers $\{1, 2, ..., T\}$. We assign one data point $\mathbf{x}_i$ to each worker i and ask for a label $\hat{y}_i$. A data point may be assigned to more than one worker (but at most two). Once assigned a task, worker i observes a signal $\tilde{y}_i$. Such observations come from a flipping error model. Workers are effort sensitive: they can choose to exert effort $e_i = 1$ to improve the quality of their outputs, or they can shirk from doing so ($e_i = 0$). Exerting effort incurs cost $c > 0$, and this is a common knowledge to all workers and the learner. Different effort levels lead to different label flipping error rates: $\Pr(\tilde{y}_i = -1|y_i = +1, e_i) = p_+(e_i)$, $\Pr(\tilde{y}_i = +1|y_i = -1, e_i) = p_-(e_i)$, and exerting effort leads to better error rate: $p_+(1) < p_+(0)$, $p_-(1) < p_-(0)$. Assuming $p_+(1) + p_-(1) < 1$. This condition is equivalent with Bayesian

informativeness in that $p_+(1) + p_-(1) < 1 \Leftrightarrow \Pr(y_i = s|\tilde{y}_i = s) > \text{Prior}(s)$, $s \in \{+, -\}$, when worker i exerts effort.[1] That is it requires that a worker's observation is informative with respect to the true label. Interestingly we find this condition is also needed for the de-bias technique for classification. The error rates $p_+(e_i)$ and $p_-(e_i)$ are common knowledge to all workers, but are unknown to the learner. After observing the signals , each worker i decides on which signal to report. The reporting could depend on its observations, such as truthfully reporting his observation, or reverting it, or randomizing over two options. A particularly undesirable reporting strategy is an *uninformative* strategy, which has the same report distribution for all observed signals. Denote worker i's report as $\hat{y}_i$. Workers can be incentivized via payment, and they would like to maximize their expected payment minus cost.

When the workers are non-strategic ($\hat{y}_i = \tilde{y}_i$), and their label flipping errors p_+, p_- are known to the learner, the problem of learning a good un-biased classifier from data with flipping errors has been resolved in Natarajan *et al.* [2013]. The naive approach of directly minimizing empirical 0-1 loss over reported data not only is technically difficult but also would give a biased classifier. Natarajan *et al.* [2013] tackle the problem by first finding a convex and *classification-calibrated*[2] loss function $l(t, y)$ (with prediction t and label y as the inputs), then defining the following surrogate loss function $\tilde{l}$: [3]

$$\tilde{l}(t, y) := \frac{(1 - p_{-y})l(t, y) - p_y l(t, -y)}{1 - p_+ - p_-},$$

and finally finding a classifier $\tilde{f}^*_{\tilde{l}}$ via minimizing the empirical risk with respect to the surrogate loss:

$$\tilde{f}^*_{\tilde{l}} = \text{argmin}_f \frac{1}{N} \sum_{j=1}^{N} \tilde{l}(f(\mathbf{x}_j), \tilde{y}_j).$$

They show that the 0-1 risk of $\tilde{f}^*_{\tilde{l}}$ converges to R^* with guarantee; further the smaller the flipping errors are, the faster the convergence of the classifier. The introduce of a convex loss function l helps to remove computational difficulties, and the surrogate loss function counters the bias. More specifically, by showing $\mathbb{E}_{\tilde{y}}[\tilde{l}(t, \tilde{y})] = l(t, y)$, where $\tilde{y}$ is the contribution from non-strategic workers with known flipping erros, and y is the ground-truth label, they establish that $\tilde{f}^*_{\tilde{l}}$ converges to the unbiased minimizer $f^*_l := \text{argmin}_f R_{l,\mathcal{D}}(f)$. Since l is classification calibrated, this gives the convergence of the 0-1 loss for f^*_l, and thus for $\tilde{f}^*_{\tilde{l}}$. Note f^*_l cannot be obtained directly, due to the errors in the training data.

For our setting with strategic workers, we need to overcome several interwinding challenges for our problem: How to incentivize effort exertion so that workers produce labels with low errors? How to incentivize workers to truthfully report the labels that they have obtained? Even when workers follow a known strategy of effort exertion and label reporting, the error rates of the two classes are unknown to the learner. How to estimate the error rates so that the method of Natarajan *et al.* [2013] can be adapted to train an unbiased classifier? This set of questions can potentially be answered by separating the effort elicitation (and possibly the estimation of flipping error) problem from the classification task, and apply several recent peer prediction mechanisms (Dasgupta and Ghosh [2013]; Kong and Schoenebeck [2016]; Shnayder *et al.* [2016]) to elicit labels. In these mechanisms, exerting effort and truthful reporting form an equilibrium (and in many cases a highest-payoff equilibrium for the workers). We depart from above line of works and propose a solution to resolve the learning and information elicitation problems jointly. Further we demonstrate the benefits of leveraging the learning-specific structure and properties for designing information elicitation mechanisms. Specifically we show that by scoring workers via comparing their answers to a reference prediction given by a reference classifier (instead of a peer worker), we are able to (1) provide better incentives (collusion is not an equilibria, in contrast to it being a bad equilibria[4]), and (2) make the elicitation step more efficient (less redundant assignments). (3) Under certain cases, our mechanism is shown to be robust to permutation strategies, which remained as a concern for adversarial classification (e.g., spam classification).

[1] We will use $+1/-1$ and $+/-$ interchangeably for the signals.

[2] l is *classification-calibrated* if there exists a convex, invertible, nondecreasing transformation ϕ_l with $\phi_l(0) = 0$ s.t. $\phi_l(R_{\mathcal{D}}(f) - R^*) \leq R_{l,\mathcal{D}}(f) - \min_f R_{l,\mathcal{D}}(f)$, where $R_{l,\mathcal{D}}(f) = \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[l(f(\mathbf{x}), y)]$.

[3] It is required that $p_+ + p_- < 1$.

[4] Dominant strategy incentive compatibility for effort exertion can also be established.

2 Mechanism and results

Mechanism 1 (`ML_Peer`)

For each worker i:

1. Estimate flipping errors $\tilde{p}_{i,+}, \tilde{p}_{i,-}$ based on reported data from workers $j \neq i$ (will be detailed later).
2. Define the following surrogate loss function $\tilde{l}(\cdot,\cdot)$ based on above estimations:
$$\tilde{l}(t,y) := \frac{(1-\tilde{p}_{i,-y})l(t,y) - \tilde{p}_{i,y}l(t,-y)}{1-\tilde{p}_{i,+}-\tilde{p}_{i,-}}.$$
3. Train a classifier $\tilde{f}^*_{\tilde{l},-i}$ using $\{\mathbf{x}_j, \hat{y}_j\}_{j\in\mathcal{U}\setminus\{i\}}$ via empirical risk minimization:[5]
$\tilde{f}^*_{\tilde{l},-i} = \text{argmin}_f \frac{1}{N-1}\sum_{j\in\mathcal{U}\setminus\{i\}} \tilde{l}(f(\mathbf{x}_j), \hat{y}_j)$, where $\mathcal{U}$ is the set that contains unique samples.
4. Score worker i: $C_p \cdot 1(\tilde{f}^*_{\tilde{l},-i}(\mathbf{x}_i) = \hat{y}_i)$, where $C_p > 0$ is a configurable constant.

Assume $R_{\mathcal{D}|y}(f^*|y) < 1/2, \forall y \in \{-1,+1\}$, i.e., f^* achieves a 0-1 loss that is less than 1/2 for each of the label class; and l is class-dependent classification calibrated fro each $y \in \{-1,+1\}$. Assume workers have perfect knowledge of the mechanism. Denote by $\Delta := \min_y\{1/2 - R_{\mathcal{D}|y}(f^*|y)\} > 0$, $\Delta_p := \min\{p_+(0)-p_+(1), p_-(0)-p_-(1)\} > 0$, and (`ML_Peer`) achieves the following results (where we use *Const(·)* to denote a constant that depends only on its inputs):

Theorem 2.1. *(1) When $K \geq$ Const1$(\mathcal{P}_+, \mathcal{P}_-, \delta_p, \Delta_p, \Delta, N)$, $N \geq$ Const2$(\mathcal{P}_+, \mathcal{P}_-, \delta_p, \Delta_p, \Delta)$, $C_p \geq$ Const3$(K, N, \delta_p, \Delta_p, \Delta, c)$, , every worker exerting effort and truthfully reporting is a BNE. (2) Reporting symmetric uninformative pure signal is not an equilibrium. (3) Reporting symmetric uninformative mixed signal with randomization probability $p \neq 1/2$ is also not an equilibrium. (4) Under certain conditions, permutation strategy is also not an equilibrium.*

Intuitive reasoning: when workers $j \neq i$ are exerting effort and truthfully reporting, and when the estimations on p_+, p_- are accurate enough, we can show the trained classifier $\tilde{f}^*_{\tilde{l},-i}$'s 0-1 loss converges to R^*, so $\tilde{f}^*_{\tilde{l},-i}$ labels data correctly with probability being strictly more than 1/2. As $\tilde{y}_i$ is compared to an answer that is more likely to be correct, when the payment constant C_p is large enough, worker i is better off exerting effort.

Now we briefly explain the steps on estimating $\tilde{p}_{i,+}, \tilde{p}_{i,-}$. Our estimation is based on the following two quantities that can be computed from the contributed data directly:

- *Matching probability*: this is the probability when a task is assigned to two different workers, the chance of the outputs match each other. Denote this quantity as q. In order to evaluate the matching on the same task, we need to re-assign the tasks. First we separate the tasks into two sets: we will randomly select K (e.g., $K = \mathrm{O}(\log N)$ to ensure enough samples for an accurate estimation) tasks to re-assign. Denote the set of re-assignment as S, and denote the uniquely assigned set as $\mathcal{U}$ (i.e., $\mathcal{U}$ contains each task exactly once, $|\mathcal{U}| = N$). We assume $T = |\mathcal{U}| + |S|$ for simplicity, so each worker is assigned exactly one task. Estimate for worker i: $\tilde{q}_i := \sum_{t\in S\setminus\{i\}} 1(\text{there is a match on task } t)/|S\setminus\{i\}|$.
- *Fraction of +1/-1 labels* that we observe, denoting as P_+, P_-; denote $\tilde{P}^i_+, \tilde{P}^i_-$ as the corresponding estimations for worker i: $\tilde{P}^i_- := \sum_{t\in\mathcal{U}\setminus\{i\}} 1(t \text{ has label } 0)/|\mathcal{U}\setminus\{i\}|$.

Denote the prior distribution of labels as $\mathcal{P}_+, \mathcal{P}_-$ (known by learner), we have the following sets of equations:
$$(1)\text{: } \mathcal{P}_+(p^2_{i,+} + (1-p_{i,+})^2] + \mathcal{P}_-(p^2_{i,-} + (1-p_{i,-})^2] = \tilde{q}_i,\ (2)\text{: } \mathcal{P}_+ p_{i,+} + \mathcal{P}_{-1}(1-p_{i,-}) = \tilde{P}^i_-.$$
When workers are arbitrarily reporting, the solution for above system of equations leads to meaningless numbers, or simply there does not exist a solution. However when workers are reporting accordingly to certain symmetric strategies, the first equation characterizes the probability of observing a matching signal, while the second one characterizes the fraction of observed negative samples. Note that the second equation is also equivalent with $\mathcal{P}_+(1-p_{i,-}) + \mathcal{P}_{-1}p_{i,-} = \tilde{P}^i_+$. Particularly we can characterize its solution for the following three cases, which correspond to our three sets of results, namely BNE, uninformative strategy proof and permutation proof. For details please refer an extended version of this work Liu and Chen [2016].

References

Anirban Dasgupta and Arpita Ghosh. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd international conference on World Wide Web*, pages 319–330. International World Wide Web Conferences Steering Committee, 2013.

Yuqing Kong and Grant Schoenebeck. A framework for designing information elicitation mechanisms that reward truth-telling. *arXiv preprint arXiv:1605.01021*, 2016.

Yang Liu and Yiling Chen. Strategic Classification with Crowdsourcing. *http://alturl.com/yubno*, October 2016.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

V. Shnayder, A. Agarwal, R. Frongillo, and D. C. Parkes. Informed Truthfulness in Multi-Task Peer Prediction. *ACM EC*, March 2016.