Strategyproof Classification

RESHEF MEIR and JEFFREY S. ROSENSCHEIN
The Hebrew University of Jerusalem

Experts reporting the labels used by a learning algorithm cannot always be assumed to be truthful. We describe recent advances in the design and analysis of strategyproof mechanisms for binary classification, and their relation to other mechanism design problems.

Categories and Subject Descriptors: I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent Systems*

General Terms: Algorithms, Theory, Economics

Additional Key Words and Phrases: Mechanism design, Classification, Game theory

1. INTRODUCTION

The field of mechanism design deals with problems that involve multiple parties, or agents, with potentially conflicting interests. The goal is typically to design interaction rules such that rational behavior of the agents will lead to an outcome that is "good" according to a certain criterion. Such a criterion may be the welfare of the agents themselves, or some other achievement about which the designer cares.

Most machine learning problems do not fall into that category. Prior research has traditionally addressed many issues related to the quality of learning (such as noise, biased sampling, partial information, and even multiple experts), but the issue of incentives has received much less attention.

However, when multiple experts are involved, game-theoretic considerations become increasingly important, especially when the agents (i.e., experts) have a direct interest in the outcome of the learning algorithm. More specifically, agents may *lie* so as to bias the outcome closer to their own opinion.

In a SIGecom Exchanges letter a few years ago, Ariel Procaccia [2008] reviewed several *strategyproof learning mechanisms*—that is, learning mechanisms[1] in which agents have no incentive to lie. Unfortunately such mechanisms cannot guarantee an optimal result (in terms of the minimal total error), and thus we evaluate them according to their worst-case approximation ratio, when compared to the optimal outcome.

Procaccia presented truthful approximation mechanisms in two highly important supervised learning domains, namely *regression* and *binary classification.* His letter also called for a synthesis of mechanism design and machine learning, and predicted that such a joint approach will benefit both communities.

In the three years that have passed since the aforementioned letter, both fields have advanced significantly. Furthermore, it turns out that strategyproof learning

[1] We use the term *mechanism* as a higher abstraction level than that of an algorithm; a mechanism focuses on information passed, and incentives, rather than on implementation details.

Authors' addresses: reshef24@cs.huji.ac.il, jeff@cs.huji.ac.il

does not have to be treated as a standalone mechanism design problem, but that it is deeply related to other kinds of problems as well. In this paper, we describe recent advances in strategyproof classification, and explain some of its unexpected connections to the problems of *facility location* and *judgment aggregation.*

2. STRATEGYPROOF CLASSIFICATION MECHANISMS

We begin with some formal definitions. A *classifier* or *concept* c is a function from some input space $\mathcal{X}$ to *labels* $\{-,+\}$. A *concept class* $\mathcal{C}$ is a set of concepts. Each agent $i \in I$ controls a set of data points X_i, where $Y_i : X_i \rightarrow \{+,-\}$ reflects the true label of each data point (known only to agent i). Let $S_i = \{\langle x, Y_i(x)\rangle \ : \ x \in X_i\}$ be the partial *dataset* of agent i, and let $S = \langle S_1, \ldots, S_n\rangle$ denote the complete *dataset.*

In a classification problem, we are given a dataset S and a concept class C, and need to return some $c \in C$ which best classifies the data. To evaluate a classifier, we simply count the number of errors (the 0–1 loss). That is, $\mathrm{R}_i(c, S) = \sum_{x \in X_i} [\![c(x) \neq Y_i(x)]\!]$ (where $[\![A]\!] = 1$ iff A is true and 0 otherwise). The *global risk* is similarly defined as $\mathrm{R}_I(c, S) = \sum_{i \in I} w_i \mathrm{R}_i(c, S)$, where w_i is the weight of agent i.

A *classification mechanism* $\mathbf{M}$ is a function (deterministic or randomized) mapping each dataset S to a classifier $c \in \mathcal{C}$; we do not allow a mechanism to make payments. We say that a mechanism is *strategyproof* (SP) if no agent can gain by lying. Formally, if for every S, i and S'_i (where the labels in S_i, S'_i may differ) it holds that $\mathrm{R}_i(\mathbf{M}(S), S) \leq \mathrm{R}_i(\mathbf{M}(S'_i, S_{-i}), S)$.

The classifier with the lowest global risk is called the *empirical risk minimizer* (ERM), and is denoted by $c^*(S)$. The optimal risk is denoted by $\mathbf{opt}(S) = \mathrm{R}_I(c^*(S), S)$. Finally, $\mathbf{M}$ is an α-*approximation* mechanism if for every dataset S, $\mathrm{R}_I(\mathbf{M}(S), S) \leq \alpha \cdot \mathbf{opt}(S)$.

The goal of the strategyproof classification agenda is the design and analysis of SP mechanisms with good (i.e., low) approximation ratios.

The setting reported by Procaccia [2008] (originally published in [Meir et al. 2008]) was a very simple one, and assumed that there are only two possible classifiers, i.e., that $|\mathcal{C}| = 2$. Under this extreme limitation, the authors provided a deterministic 3-approximation SP mechanism, and showed that no better mechanisms exist. Allowing randomization can improve the approximation to 2, which is again a tight bound.

When considering the most general classification setting, no deterministic SP mechanism can guarantee any reasonable outcome [Meir et al. 2010] in terms of approximation.[2] This result holds for widely used concept classes like linear classifiers and boolean conjunctions. This strong negative result means that some restriction is necessary to obtain good mechanisms.

2.1 Shared inputs

A natural restriction is to assume that all agents are labeling the same set of data points X, i.e., that all X_i are equal. This is the case, for example, in online surveys, where everyone is answering the same set of questions. Quite interestingly, this simple restriction makes the problem much easier to handle. In fact, selecting

[2] A similar negative result was provided for randomized mechanisms, but it requires additional technical assumptions.

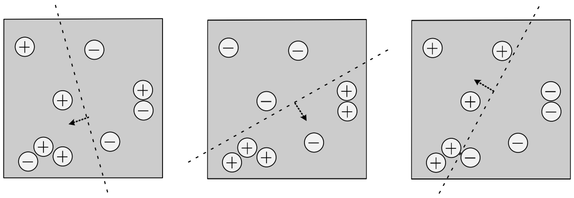

Fig. 1. An instance of a dataset with shared inputs. Here, $\mathcal{X} = \mathbb{R}^2$, $\mathcal{C}$ is the class of linear separators over $\mathbb{R}^2$, and $n = 3$. The data points X of all three agents are identical, but the labels are different. The best classifier from $\mathcal{C}$ with respect to each S_i is also shown (the arrow marks the positive halfspace). Only the rightmost dataset is realizable.

an agent at random and using him as a dictator guarantees a $3 - \frac{2}{n}$ approximation ratio (and is clearly SP) [Meir et al. 2009]. When agents are weighted, the same approximation ratio can be achieved with a more subtle randomization [Meir et al. 2011]. The latter paper also proves that no SP mechanism can do better.

2.2 Realizable datasets

A particular case of interest is when the dataset of an agent can be classified perfectly, i.e., there is some $c_i \in \mathcal{C}$ s.t. $c_i(x) = Y_i(x)$ for all $x \in X$. If this is the case for every agent, we say that the data is *individually realizable* (IR); see Figure 1. It turns out that IR can improve the approximation ratio even further, from $3 - \frac{2}{n}$ to $2 - \frac{2}{n}$ [Meir et al. 2009; 2011]. Interestingly, we must know in advance whether our dataset is IR or not in order to apply the correct mechanism—selecting the mechanism after observing the data is no longer SP.

2.3 Generalizing from samples

A crucial requirement from supervised learning algorithms, and classification algorithms in particular, is that rules learned from sampled data can be applied to new data. Formally, we want the *empirical error* (on the dataset) to be close to the real error (measured on the entire distribution). Unfortunately, the SP requirement w.r.t. the real error cannot be obtained even if there is only one agent, due to the small chance that the sample does not reflect the agent's true opinion. In such cases we need to make some assumptions on the behavior of the agents. The *truthful approach* asserts that agents will only lie if they gain at least ϵ from doing so. In contrast, the *pure rationality approach* assumes that an agent will use a dominant strategy when one is available to him. Notably, for concept classes of a bounded VC dimension, all the algorithms mentioned above can be applied to sampled data under either of these assumptions, guaranteeing an approximation ratio that is arbitrarily close to $3 - \frac{2}{n}$.

3. A UNIFIED APPROACH TO SP CLASSIFICATION AND MECHANISM DESIGN

In addition to the technical advances mentioned above, many conceptual links have been drawn between the broad framework of mechanism design without money (i.e., without payments), and the problem of SP classification (the case of shared inputs in particular).

3.1 Judgment aggregation

A fairly intuitive connection is with the problem of *judgment aggregation* (JA) [Dokow and Holzman 2010]. In JA there is an agenda consisting of several logical expressions, and each agent has some opinion over the agenda. The different opinions, or judgments, should be aggregated to a single consistent assignment to all logical atoms. A simple mapping between the problems would identify each issue of the agenda with a data point, where every assignment vector corresponds to a binary classifier. The set of all legal (consistent) assignments then corresponds to the concept class of the learning problem. Note that in JA there is usually a requirement that the opinion of each agent itself be logically consistent. This requirement coincides with the IR requirement in the classification setting.

3.2 Facility location

In the *facility location* (FL) problem, agents report their location (usually in some metric space), and the mechanism outputs a location for a facility that is close, on average, to all agents [Procaccia and Tennenholtz 2009].

Consider a dataset labeled by several agents, and a binary cube where each dimension corresponds to a data point. We can now identify the label vector of each agent with a specific vertex of this cube. Similarly, any concept class (which defines the allowed labeling) corresponds to a set of vertices that constitutes the allowed locations. The IR condition in the FL setting is translated to the restriction that agents' locations are limited vertices where the facility can be placed. Moreover, the optimal location in FL corresponds to the ERM classifier.

4. CONCLUSION AND FUTURE DIRECTIONS

The above correspondences imply that questions of incentives and truthfulness in the distinct settings of JA, FL and learning can be studied in a unified model.

The design of learning algorithms that preclude or handle strategic behavior is advancing quickly, but certain obstacles still hinder its successful application to problems in the real world. First, the current models are quite general, overlook many intricacies that are featured in data from particular domains, and focus on worst-case analysis. Second, strong requirement of strategyproofness constrains the possible set of algorithms, whereas weaker strategic requirements may allow for much better results.

The first obstacle should be tackled with the help of experimental and empirical analysis of real data. As for the second, we believe that the emerging integration with the wider area of mechanism design will supply the necessary conceptual and technical tools to develop the proper solution concepts.

REFERENCES

Dokow, E. and Holzman, R. 2010. Aggregation of binary evaluations. *Journal of Economic Theory 145*, 495–511.

Meir, R., Almagor, S., Michaely, A., and Rosenschein, J. S. 2011. Tight bounds for strategyproof classification. In *Proceedings of the 10th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. Taipei, Taiwan, 319–326.

Meir, R., Procaccia, A. D., and Rosenschein, J. S. 2008. Strategyproof classification under constant hypotheses: A tale of two functions. In *Proceedings of the 23rd AAAI Conference on Artificial Intelligence (AAAI)*. 126–131.

MEIR, R., PROCACCIA, A. D., AND ROSENSCHEIN, J. S. 2009. Strategyproof classification with shared inputs. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI)*. 220–225.

MEIR, R., PROCACCIA, A. D., AND ROSENSCHEIN, J. S. 2010. On the limits of dictatorial classification. In *Proceedings of the 9th International Joint Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)*. 609–616.

MEIR, R., PROCACCIA, A. D., AND ROSENSCHEIN, J. S. 2011. Algorithms for strategyproof classification. manuscript.

PROCACCIA, A. D. 2008. Towards a theory of incentives in machine learning. *ACM SIGecom Exchanges 7,* 2.

PROCACCIA, A. D. AND TENNENHOLTZ, M. 2009. Approximate mechanism design without money. In *Proceedings of the 10th ACM Conference on Electronic Commerce (ACM-EC)*. 177–186.