Incentive Mechanisms in Strategic Classification and Regression Problems

Kun Jin, Xueru Zhang, Mohammad Mahdi Khalili, Parinaz Naghizadeh and Mingyan Liu

Introduction

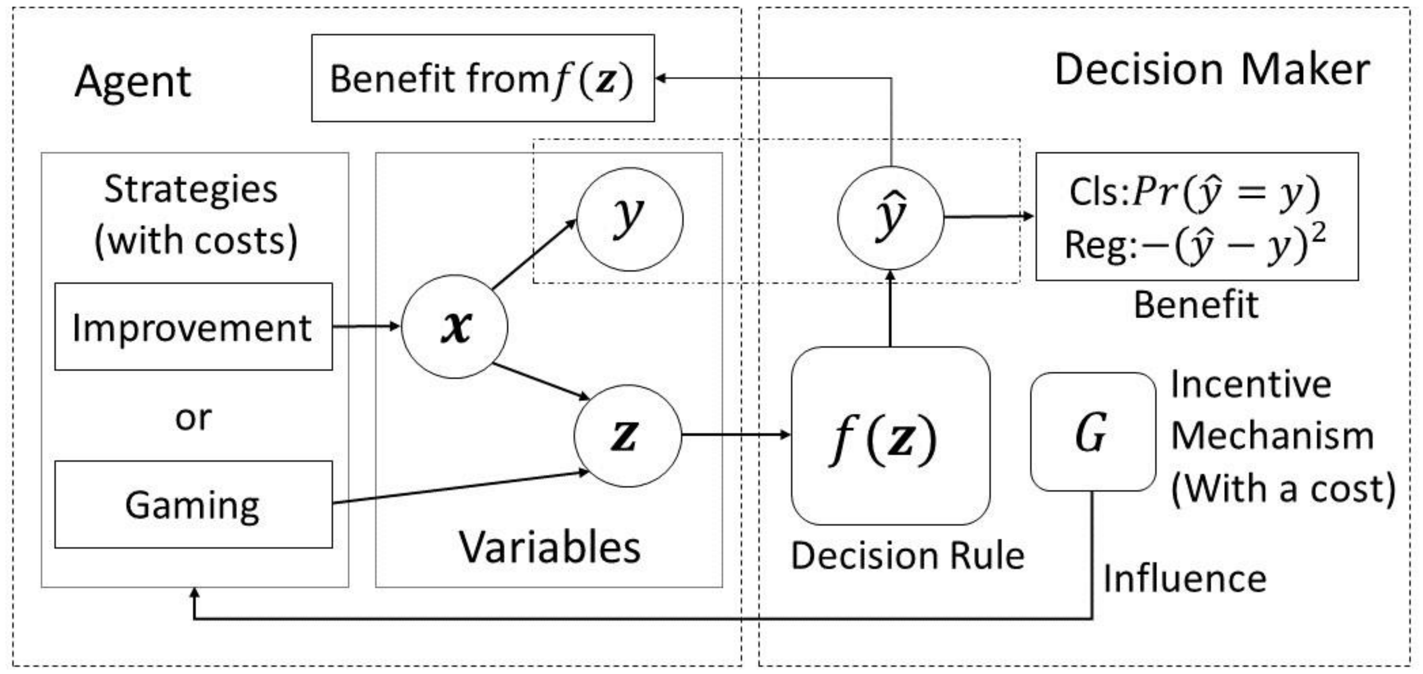
We study the design of subsidy mechanisms in strategic learning problems, which are modeled as Stackelberg games. In these games, the decision maker designs and commits to a decision rule first, and the agents strategically best respond to the rule by manipulating their features. The manipulation that does not change the agents' labels are gaming actions while those that improve the feature and labels simultaneously are improvement actions. We show that subsidizing the improvement actions can ***benefit both sides*** in the game and study the how the mechanism designer's objective influence the system outcome.

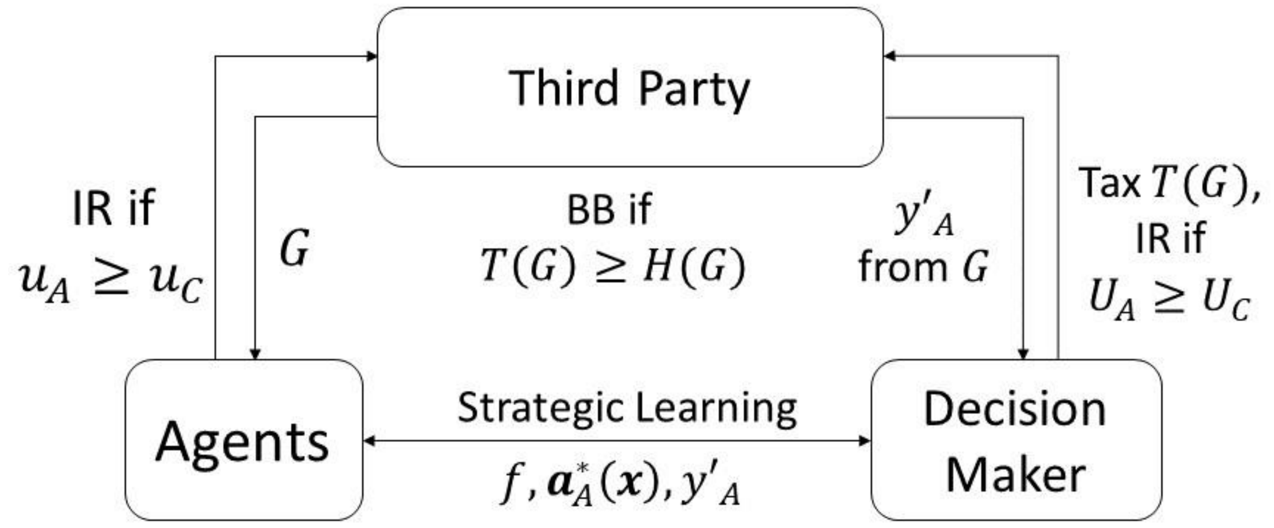
Model

N-dim attribute $\boldsymbol{x} \sim p(\boldsymbol{x})$, ***private*** information
Action $\boldsymbol{a} = (\boldsymbol{a}_+, \boldsymbol{a}_-)$ (improvement/gaming)
Post response feature is $\boldsymbol{z} = \boldsymbol{x} + P\boldsymbol{a}$, public information, P is projection matrix
Decision rule $f(\boldsymbol{z}) = \mathbf{1}\{\boldsymbol{w}^T\boldsymbol{z} \geq \tau\}$,
Post-response attribute $\boldsymbol{x}' = \boldsymbol{x} + P_+\boldsymbol{a}_+$, ***private*** information
Post-response label $y' = l(\boldsymbol{\theta}^T\boldsymbol{x})$, known to the decision maker if agent is accepted

Augmented Strategic Classification

Augmented strategic learning system uses an augmented mechanism that combines the ***subsidy and the decision rule***. We consider two cases, (1) decision maker designs the subsidy, and (2) a third-party designs the subsidy with social well-being objectives.





Subsidy G partly covers the cost of actions, and the subsidized agent utility is $u_A(\boldsymbol{x}, \boldsymbol{a}) = f(\boldsymbol{x} + P\boldsymbol{a}) - h_A(\boldsymbol{a})$ where $h_A(\boldsymbol{a}) = h(\boldsymbol{a}) - \Delta\boldsymbol{c}^T\boldsymbol{a}$ is the subsidized cost. Denote the best response as $\boldsymbol{a}^*_t(\boldsymbol{x}), t \in \{A, C\}$
Decision maker optimizes $U_A(f) = \boldsymbol{E}[y'_A = f(\boldsymbol{z}_A)] - H(G)$, H is subsidy cost

Finding Optimal Subsidy is Hard

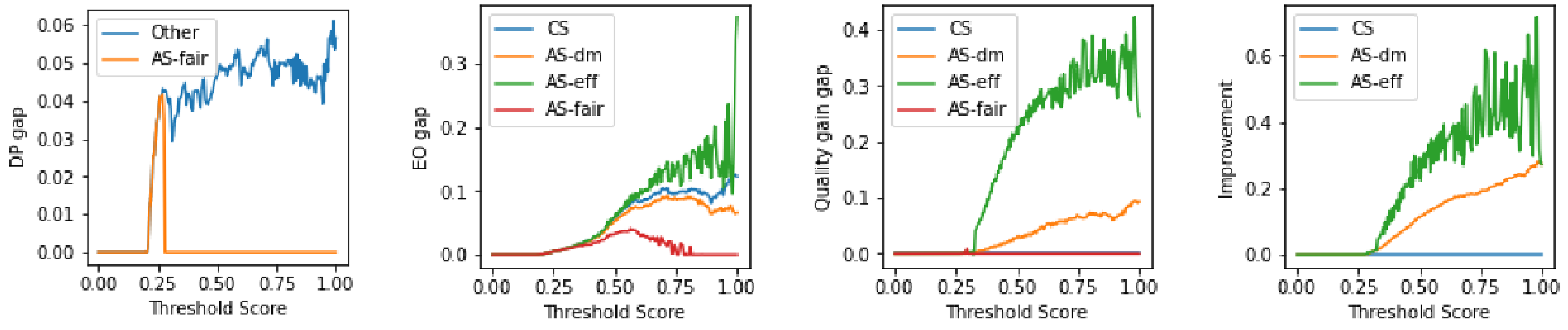
We aim to find subsidies that are individually rational (IR), incentive compatible (IC) and budget balanced (BB). For general $\boldsymbol{w}$ in $f(\boldsymbol{z}) = \mathbf{1}\{\boldsymbol{w}^T\boldsymbol{z} \geq \tau\}$, $p(\boldsymbol{x})$, and $l(\boldsymbol{\theta}^T\boldsymbol{x})$, finding the optimal IC, IR and BB subsidy is hard. But in a special but realistic special case when $\boldsymbol{w} = \boldsymbol{\theta}$, and l is convex on $[0, \tau]$, we can have ***closed-form representations*** of the optimal subsidy. $\boldsymbol{w} = \boldsymbol{\theta}$ is the optimal choice when it's impossible to incentivize improvement with $f(\boldsymbol{z})$ alone.

Multiple Demographic Groups

Agents from different demographic groups are distinguished by a ***sensitive attribute*** $d \in \{1,2\}$, which is ***private*** information. We study unified classifier f that is not allowed to use d as a feature, but group specific subsidies G^d, which can induce the agents to reveal d

Social Well-being Analysis

The third-party mechanism designer can focus either on efficiency-oriented objective $\boldsymbol{E}[y'_A]$ (the social quality) or the fairness-oriented objective. The fairness metrics we study include: (1) ***quality gain gap*** that measures the group-wise difference of expected label improvement pre- and post-response, (2) ***EO gap*** (TPR difference), and (3) ***DP gap*** (PR difference).
We show in the numerical experiment with the FICO dataset that while the augmented mechanism always improves efficiency-oriented objectives, the decision maker and the efficiency-oriented third-party can make fairness issues worse since they prefer to subsidize the advantaged group. On the other hand, a ***fairness-oriented third-party*** can achieve ***improvement on the efficiency, the fairness, the algorithm robustness and benefit all parties***.



(a) DP gap (b) EO gap (c) Quality gain gap (d) Improvement