NEWDIGS Incentives

A Double Auction for Pharmaceutical Predictions

Andy Haupt

Massachusetts Institute of Technology
NEWDIGS

March 20, 2020

Outline

Setup

- Pharma companies would like to collaborate with each other and algorithm designers to create predictions of patients/patient timeseries
- Status Quo: Decentralized negotiations are inefficient
- Predictions not excludable, and this feature of the market is a constraint also for future designs
- There is, however, interest in predictions for heterogenous patient populations

$\rightarrow$ Heterogeneity might allow for surplus in centralised design

Constraints/Assumptions

- As soon as predictors are shared, information will reach other market participants: Effectively, a public good is allocated
- The fact of someone else knowing something does not give a relevant negative externality in itself
- Can merge prediction tasks that might have synergies into a single one

Additional Challenges

- Design an Auction which requires little prior information on participants' willingness to pay
- Allow for heterogeneity sub-populations of interest that partly overlap and partly don't.
- Control computational and communication complexity

Related Work and Examples

- Information Pricing Based on the Shapley Value Agarwal, Dahleh, and Sarkar 2019; Ohrimenko et al. 2019; Ghorbani et al. 2020; Yona et al. 2019.
  - → Challenge: Uses a concept from cooperative game theory which is manipulable. Papers justify in the use case of consumer web data.
- Information Pricing with a probabilistic model Bergemann et al. 2018; Howard 1962.
  - → Challenge: Uses extensive prior modelling of either downstream decision-making based on data, or prior knowledge on willingness to pay.
- Incentive Auctions Milgrom et al. 2020 (and more earlier papers)
  - → Used to reallocate spectrum (most recently) from TV channels to IoT and Wireless companies. about $10 Bn revenue.

FCC Auction

- FCC Explanation (2min)
- Similarities
  - Heterogenous Buyers and Sellers
  - Computational Challenges
  - Communication Challenges
- Differences
  - There is no concept of staying
  - More generally, there are no “winners” / “losers” of the auction

Definitions

Data Suppliers There are m data suppliers that each have data in the form of pairs $(\hat{X}_i, \hat{Y}_i)$, $i = 1, 2, \ldots, m$; $\hat{Y}_i | \hat{X}_i \sim p$

Data Demanders There are n prediction demanders that each have a test set in the form of pairs (X_i, Y_i), $i = 1, 2, \ldots, n$; $Y_i | X_i \sim p$

Algorithms There are k algorithms, which are functions $f_k \colon X \mapsto f(X)$ that can be trained using data $(\hat{X}, \hat{Y})$

Preferences
- Suppliers and algorithm designers have a fixed reservation price for all of their data.
- Demanders are interested in the prediction performance a predictor f on their distribution (X_i, Y_i) and have a willingness to pay that is linear in percent accuracy increase (compare Agarwal, Dahleh, Shah, et al. 2019)

Goal Design a truthful auction

Overview of the Proposed Auction

1. There is a reverse auction that starts with a high sales prices and decreases prices in (potentially individualised increments).
2. Data contributors can stay in the auction or exit.
3. Before a price is decremented, check whether leaving out this data could "fund itself" (see example below):
   - Train decorrelated algorithms (see below) to consider different patient sub-populations separately.
   - For each sub-population, determine, whether several agents derive utility from accuracy on this population (the public goods case) or only one agent is interested.
   - Calculate the normalised value of accuracy.
   - Set accuracy in the public goods case accoding to the agent with the highest normalised value of accuracy.
4. If no decrease can be made, but the revenue from selling prediction does not fund buying the data, reverse the auction.
5. Otherwise, allocate the predictions and issue payments.

Example

- There three distinct patient populations A, B, C.
- Two data suppliers $1, 2$ hold equal amounts of A, B, only 2 holds data C.
- There are two algorithms f_1 and f_2.
- One data demander, a is interested in A 80% of the cases, 20% in B, the other 70% in C and 30% in B.
- Willingness to pay per $\%$ accuracy is $10k for a, $8k for b.
- We compute the normalised values of accuracy on patient populations: For A, $ 8k. For B, $ 3k resp. $ 2.4k, and for C

Example II

- Whenever we lower prices for 1 or 2, we look whether such a decrease could "pay itself" in the following way:
    - We decorrelate the algorithms (see below) and consider patient sub-populations separately
    - If a single demander is interested in a patient population, determine accuracy for them alone
    - If several are interested, compute the value per patient sub-population (see above) and consider the maximum in setting the accuracy
    - If the decrease in payment due to less data is smaller than the gain by decreasing the price, propose to lower the price

Decorrelating Algorithms

- To reduce constraints between patient subpopulations, we want to train machine learning algorithms to perform poorly on test data outside of a certain domain.
  - Variant 1 to achieve this: Add noise to prediction in certain parts of the space: Fragile, could be reverse engineered
  - Variant 2 to achieve this: Give noisy training data for points outside a certain training population.
- Limited coverage in machine learning literature, despite other relevant applications (no criminal use of, e.g., deepfake technology)

Extensions and Limitations

- Unexplored Questions:
  - Analysis of investment incentives and welfare, Akbarpour et al. 2020
- Limitations
  - Cannot ensure that agents want to participate in this mechanism; e.g. data demander that is interested in exactly the same questions as another agent that one thinks is more interested in question cannot be incentivized to participate in any mechanism.
  - Optimising paramters of the auction are hard to optimise and need domain knowledge even if any choice makes it truthful.
- Open Questions:
  - How to quantify heterogeneity of interests.
  - Evaluate theory for quantifying entanglement of different sub-populations.

Discussion

- Are the assumptions made realistic?
- Is there some value for you?
- If yes, what would you like more concrete soon?
- Next steps
  - Next meeting?
  - Structure of further collaboration in COVID-19.
- Thank you!