Feature Selection in Environments with Limited Voluntary Information Sharing

Nicholay Topin, Karan K. Budhraja, Tim Oates

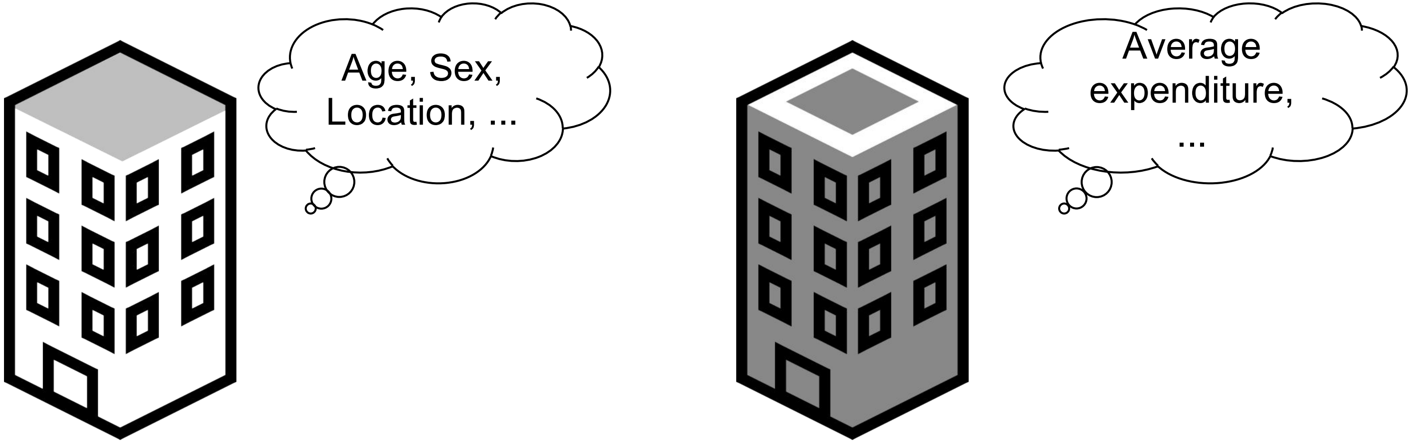
Contents

- ➔ Motivation
- ➔ Related work
- ➔ Problem definition
- ➔ Proposed method
- ➔ Evaluation
- ➔ Applications and future work

Motivation

➔ Real-world contains *vertically partitioned* data

- E.g. company *A* has demographic data for a group
- E.g. company *B* has shopping data for the same group
- Companies know certain features about a common population



Motivation

➔ Example classification task: modeling personal behaviour

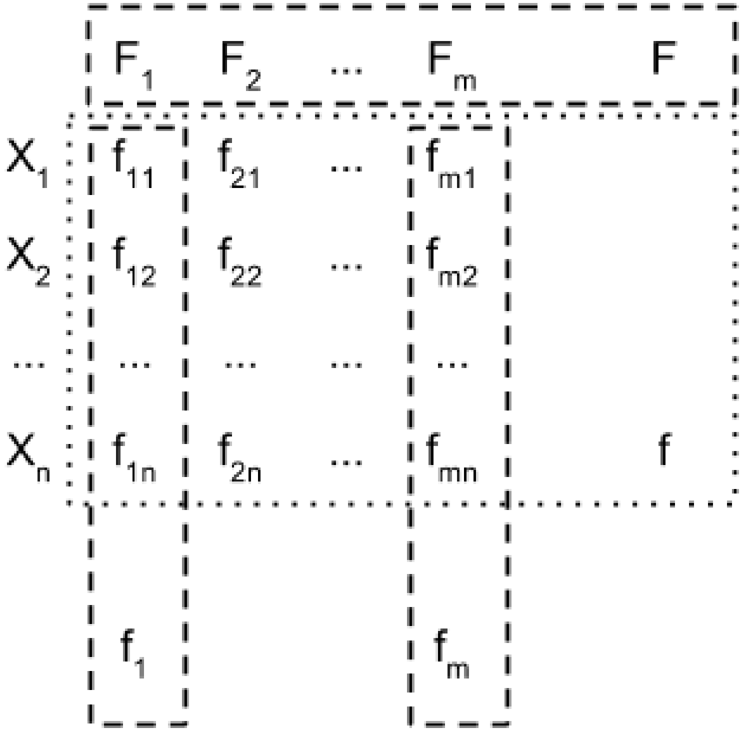
- Required feature data spread across companies
  - Quality of data varies per feature
  - Data access is priced
- Features accessed should be *worth paying for*
  - This is a feature selection problem

Motivation

- ➔ Estimating data price
  - ◆ Price ∝ Data importance
- ➔ Environment contains multiple buyers, multiple sellers
  - ◆ No governing third-party
- ➔ *Voluntary information sharing*
  - ◆ Incentive based
  - ◆ Closer to real-world setting
- ➔ Limited information sharing
  - ◆ Limit on number of agents contacted

Related Work

- ➔ Adversarial feature selection
  - Data actively manipulated per column (F)
  - Adversary causes misclassification
  - *Game* between buyer and seller



Related Work

- ➔ Existing information sharing models
  - ◆ Trust modeling
  - ◆ Fuzzy logic based
  - ◆ Model reputation
- ➔ All require *compulsory information sharing*
- ➔ We do not currently focus on dishonesty / trust modeling

Problem Definition

- Defined as a game
  - Players: *Buyer* (*B*), *Seller* (*S*)
  - Multiple buyers and sellers present in same environment
- Cost definition
  - *B*: *cost of accessing features* (*v*)
  - *S*: not applicable
- Utility (*U*) definition
  - *B*: proportional to classification accuracy using a given set of features
  - *S*: (monetary) value gained on sale of a feature to a buyer
    - Cost incurred by the buyer

Problem Definition

- Utility to Cost Ratio (UCR) defined for buyer
  - Fixed for a given buyer
  - Seller tries to identify to optimize sales

Proposed Method

- ➔ Buyer Strategy
  - Buy *useful features* to improve performance on task
  - Classification task
- ➔ Seller Strategy
  - Learn UCR of various buyers
  - Sell features at *highest costs* possible
- ➔ Turn based game
  - The order in which sellers take turns is fixed
  - The order in which buyers are offered features is random

Proposed Method

- ➔ Types of buyers
  - ◆ Simple buyer (B)
    - Utility includes a discount factor
      (i.e., a feature is less useful when buyer already has high total utility)
- ➔ Types of sellers
  - ◆ Binary search seller (BS)
    - Use binary search to identify buyer UCR
  - ◆ Learning with binary search seller (LWBS)
    - Use binary search to identify buyer UCR
    - Use Q-learning to *learn order of selling features*

Proposed Method

Algorithm 1 A round of buyer and seller interaction with no information sharing.

Input: Buyer and seller models and utility functions
Initialize buyer and seller utility values to zero
Initialize seller estimates of buyer models
for all Sellers (in fixed order) **do**
  for all Buyers (in random order) **do**
    Seller offers a feature to the buyer at a price
    Buyer accepts or rejects offer.
    if Offer accepted **then**
      Seller sells to the buyer
      Seller updates its model of the buyer
      End of this sellers turn (go to next seller)
    else
      Seller updates its model of the buyer
    end if
  end for
end for
Return: Utilities of buyers and sellers

Proposed Method

➔ Information sharing

- Request up to a fixed number of sellers to share information
  - Gain information about a buyer
  - A seller providing information is a “witness”
- Potential witness considers gains from requesting information from this seller in the future
  - *No immediate gain* for witness
- Future gain based on knowledge gap between requester and witness
- Fractional update to personal knowledge

$$WS = \sigma(|KS - KS_w| / |KS_w|) \qquad PE = \alpha * PE_w + (1 - \alpha) * PE$$

Proposed Method

Algorithm 2 Witness selection and information gathering in an environment with limited voluntary information sharing.

Input: Maximum number of witnesses K_s
Initialize number of witnesses $W_s = 0$
while $W_s < K_s$ **and** Candidate witness available **do**
  Request candidate seller to be a witness
  if Request accepted **then**
    Get information from witness
    Update witness information-based buyer model PE_w
    $W_s = W_s + 1$
  end if
end while
Return: PE_w

Evaluation

- ➔ Time complexity improvement
  - ◆ Improvement over adversarial feature selection
    - N: number of data items / rows (X)
    - J: number of features / columns (F)
    - $O(N^2J)$
  - ◆ Preprocess features
    - Evaluating utility for a given set of features is now $O(1)$
      - Store *residuals*
    - Running a round with information sharing is $O(S^2J)$
      - S is number of sellers

Evaluation

- ➔ Experimental setup
  - ◆ Preliminary evaluation: *proof of concept*
  - ◆ UCI letter recognition dataset
    - 16 attributes, 20,000 instances
  - ◆ Average over 1000 executions, 3000 games each, 40 rounds per game
  - ◆ Each seller begins with same features
  - ◆ Arbitrary values for Q-learning
    - $\alpha=0.1$, $\gamma=0.99$, $\varepsilon=0.0$
  - ◆ No information sharing / information sharing configurations

Evaluation

➔ Experimental results

◆ No information sharing (2 BS sellers, 1 buyer), seller utility

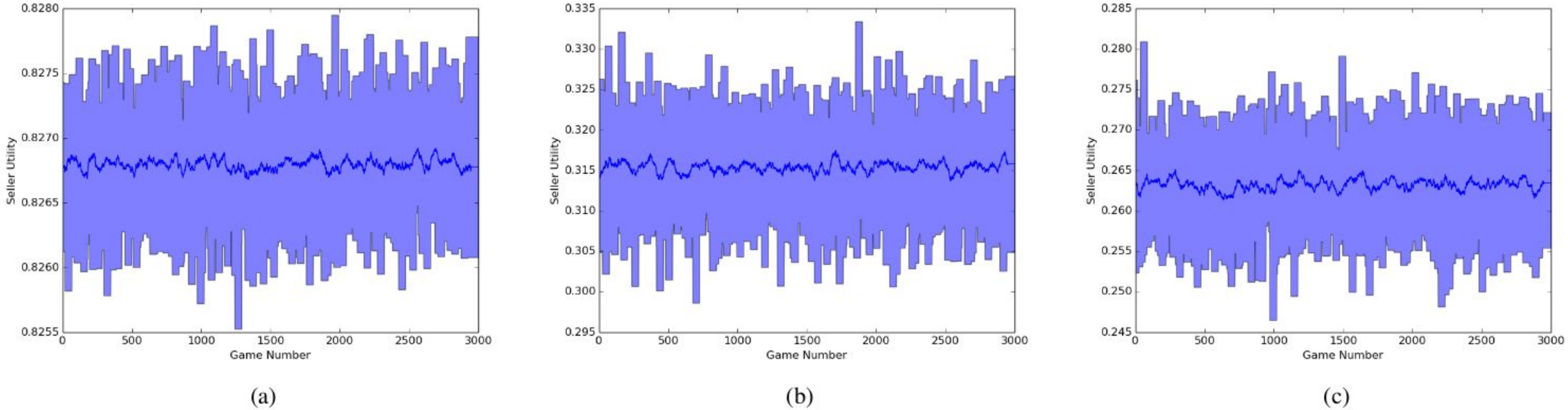


Fig. 2: Utility of a BS seller in a single buyer environment with no information sharing. Subfigure (a) shows utility of a single BS seller (with no seller competition). Subfigures (b) and (c) depict utility of two competing BS sellers (with turns in that order). Consistency of utility on average reflects BS sellers not learning between games.

Evaluation

➔ Experimental results

- No information sharing (2 LWBS sellers, 1 buyer), seller utility

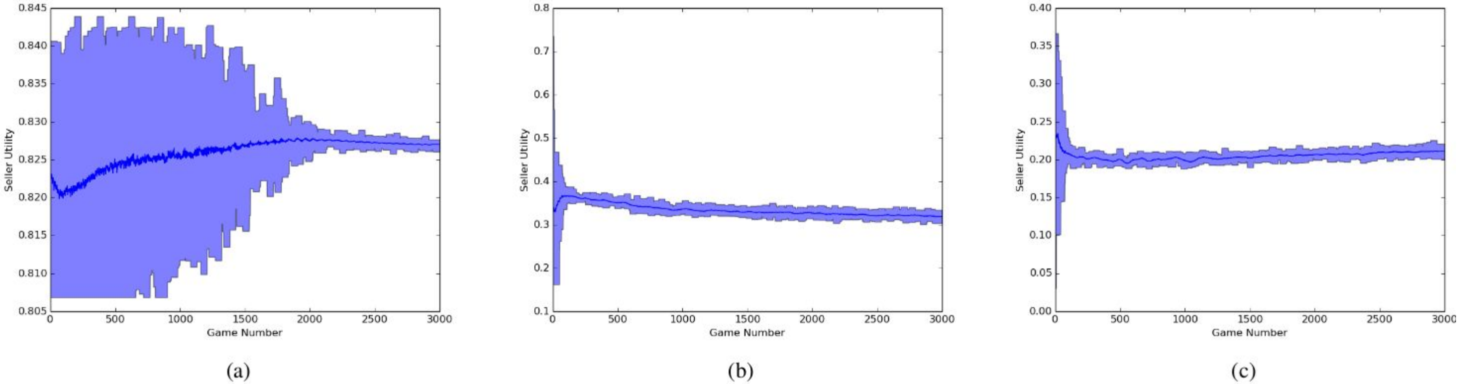


Fig. 3: Utility of an LWBS seller in a single buyer environment with no information sharing. Subfigure (a) shows utility of a single LWBS seller (with no seller competition). Subfigures (b) and (c) depict utility of two competing LWBS sellers (with turns in that order). Because the overall utility that can be extracted from a buyer is fixed for a given buyer, a gradual decrease in utility (to stabilization) of the first LWBS seller is reflected by a gradual increase in utility (to stabilization) of the second LWBS seller. Such gradual change and stabilization of seller utility is suggestive of a situation among the sellers similar to Nash-equilibrium [15].

Evaluation

- ➔ Experimental results
  - ◆ No information sharing (1 BS, 1 LWBS seller, 1 buyer)
  - ◆ BS Seller utility
  - ◆ LWBS seller utility

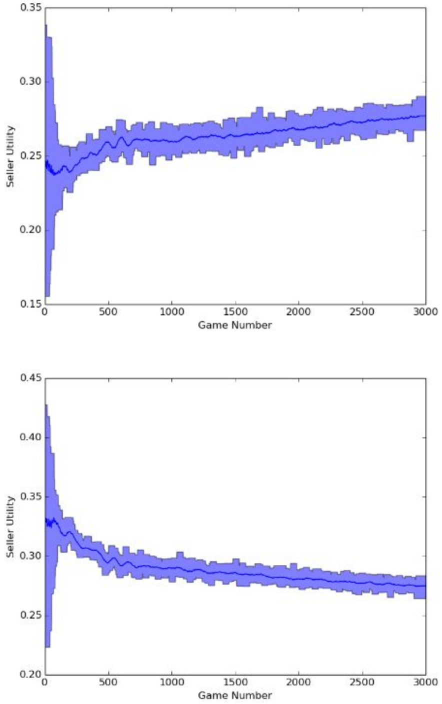


Fig. 4: Utility of competing BS (above) and LWBS (below) sellers in a single buyer environment with no information sharing. LWBS utility in a competitive environment gradually decreases to stabilization, allowing BS utility to gradually increase to stabilization.

Evaluation

➔ Experimental Results

- ◆ Information sharing
  (8 BS sellers, 8 buyers)
- ◆ BS seller utility

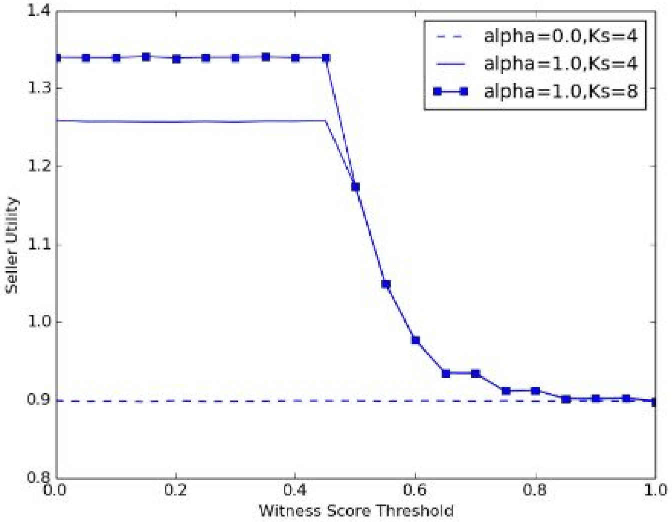


Fig. 6: Utility of a BS seller in an 8 buyer, 8 BS seller environment with information sharing. BS sellers benefit from information sharing ($\alpha = 1.0$).

Evaluation

➔ Experimental Results

- Information sharing
  (8 LWBS sellers, 8 buyers)
- LWBS seller utility

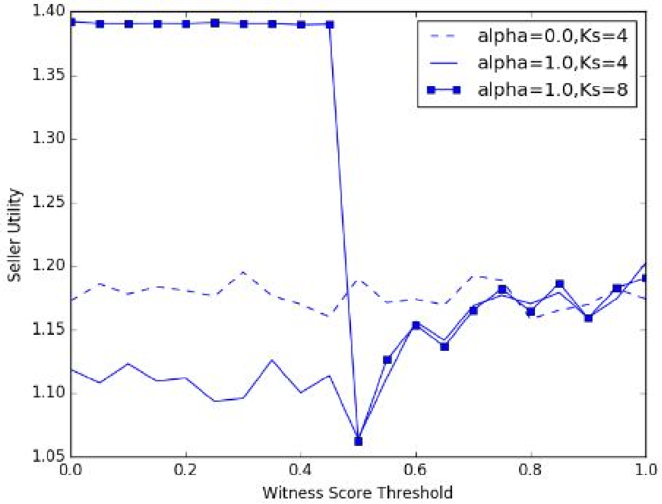


Fig. 7: Utility of an LWBS seller in an 8 buyer, 8 LWBS seller environment with information sharing. LWBS sellers benefit from maximized information sharing (witness score threshold ≤ 0.5) but not further limited information sharing (witness score threshold > 0.5).

Evaluation

➔ Experimental results

- Information sharing (4 BS, 4 LWBS sellers, 8 buyers), seller utility

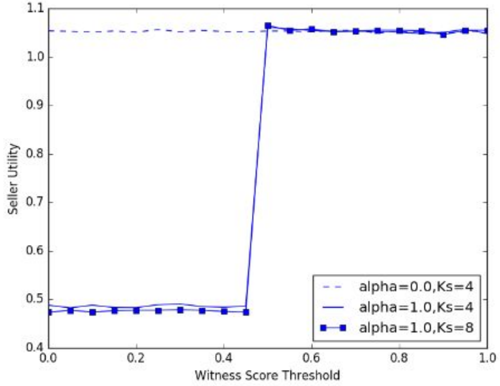


Fig. 8: Utility of a BS seller in an 8 buyer, 4 BS seller, 4 LWBS seller environment with information sharing. The BS sellers extract utility from the buyers, limited by the utilities of the LWBS sellers. This figure is therefore negatively proportional to Figure 9.

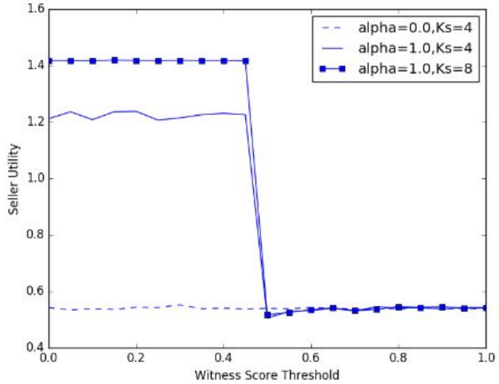


Fig. 9: Utility of an LWBS seller in an 8 buyer environment with 4 BS and 4 LWBS sellers using information sharing. LWBS sellers benefit from full information sharing ($\alpha = 1.0, K_s = 8$ with witness score threshold ≤ 0.5) but not limited information sharing ($\alpha = 1.0, K_s = 4$).

Evaluation

➔ Experimental results

◆ Witness information sharing summary

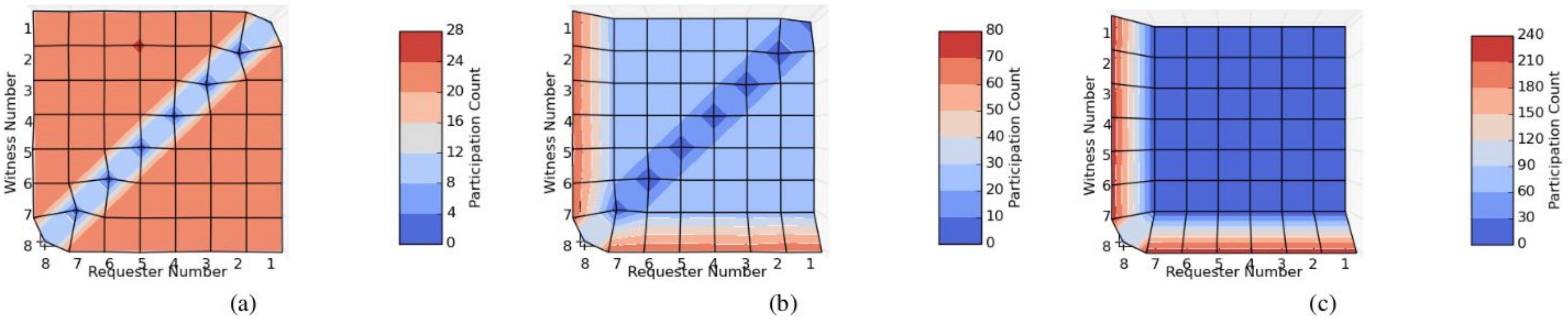


Fig. 5: Witness information sharing summary. With no score threshold requirement as in (a) (witness score threshold = 0), all sellers exchange information irrespective of incentive. With the introduction of incentive requirement as in (b) (witness score threshold = 0.1), witness participation becomes more one-sided. This is because there is not enough incentive for sellers to reciprocate to witnesses when requested by the respective witnesses in future rounds. An extreme case of (b) is represented in (c) with witness score threshold = 1.0. Sellers are not witnesses for themselves and diagonal elements are therefore zero.

Applications

- ➔ Modeling real-world transaction environment
  - ◆ *Incentive-based* communication
- ➔ No third-party requirement
  - ◆ *Self-regulating* environment of agents
- ➔ Study environment
  - ◆ Study buying / selling strategies in environment

Current and Future Work

- Future work
  - Integrate *dishonesty*
    - Trust modeling
  - More complex buyers
  - More complex sellers

Selected References

Budhraja, Karan Kumar, and Tim Oates. "Adversarial feature selection." *2015 IEEE International Conference on Data Mining Workshop (ICDMW)*. IEEE, 2015.

Chang, Liu, et al. "Multi-criteria decision making based on trust and reputation in supply chain." *International Journal of Production Economics*147 (2014): 362-372.

Zhang, Jie, Robin Cohen, and Kate Larson. "A trust-based incentive mechanism for e-marketplaces." *International Workshop on Trust in Agent Societies*. Springer Berlin Heidelberg, 2008.

Tran, Thomas. "Protecting buying agents in e-marketplaces by direct experience trust modelling." *Knowledge and information systems* 22.1 (2010): 65-100.

Acampora, Giovanni, Arcangelo Castiglione, and Autilia Vitiello. "A fuzzy logic based reputation system for E-markets." *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. IEEE, 2014.

Pan, Yong. "E-marketing under the adverse selection environment: model and case study." *International Journal of Networking and Virtual Organisations* 14.1-2 (2014): 111-128.

Thank you!

Please send questions to:

ntopin1@umbc.edu

karanb1@umbc.edu

oates@cs.umbc.edu