Decisions, Counterfactual Explanations, and Strategic Behavior

By: Stratis Tsirtsis and Manuel Gomez-Rodriguez

Presented by: Ahsan Sanaullah

Summary

- Automated decision makers play an increasingly important role in society
- People increasingly demand explanations of the decision making process from automated decision makers
- Tsirtsis *et al.* provide approximation algorithms for finding the optimal decision policy and set of counterfactual explanations in terms of maximimizing decision maker utility.

Automated Decision Makers

- Automated decision makers play many parts in society today, these include:
  - Banks giving out loans
  - Investment firms trading stocks
  - Companies hiring employees
  - Many more
- Automated Decision Makers are increasingly asked to explain their decisions
  - The EU has passed a *right-to-explanation* law for any individual subject to an automated decision making process

Counterfactual Explanations

- Tsirtsis *et al.* attack this problem by attempting to give counterfactual explanations.

- A counterfactual explanation is an example of something that an individual could change that would guarantee they receive a beneficial decision
  - E.G. if you reduce your debt by 20%, we will give you the loan you request.

Problems

- The authors attack 3 main problems:

1. Find an optimal set of counterfactual explanations for a given decision policy.
2. Find an optimal decision policy for a set of counterfactual explanations.
3. Jointly find the optimal decision policy and set of counterfactual explanations.

Proofs

- The authors show that problems 1 and 3 are NP-Hard with a reduction of Set Cover.

- They show that problem 2 can be done in polynomial time and provide an algorithm with $O(km)$ time complexity. k is the size of the set of counterfactual explanations. m is the number of possible feature values.

Approximation

- However, the authors show that problem 1 can be approximated with an $1 - \frac{1}{e}$ approximation factor using a standard greedy algorithm (Nemhauser *et al.*, 1978).

- This means that the utility the decision maker achieves with the approximated set of counterfactual explanations is $\geq \left(1 - \frac{1}{e}\right) \times$ the utility of the optimal set.

Approximation cont.

- The authors also show that problem 3 can be approximated. An approximation factor of $\frac{1}{e}$ is achieved using a recent randomized algorithm (Buchbinder *et al.*, 2014).

- This means that the utility of the decision maker with the approximated decision policy and set of counterfactual explanations is $\leq \frac{1}{e} \times$ the utility with the optimal policy and set.

Experiments

- The authors performed experiments on simulated and real data, the simulated data was generated randomly.

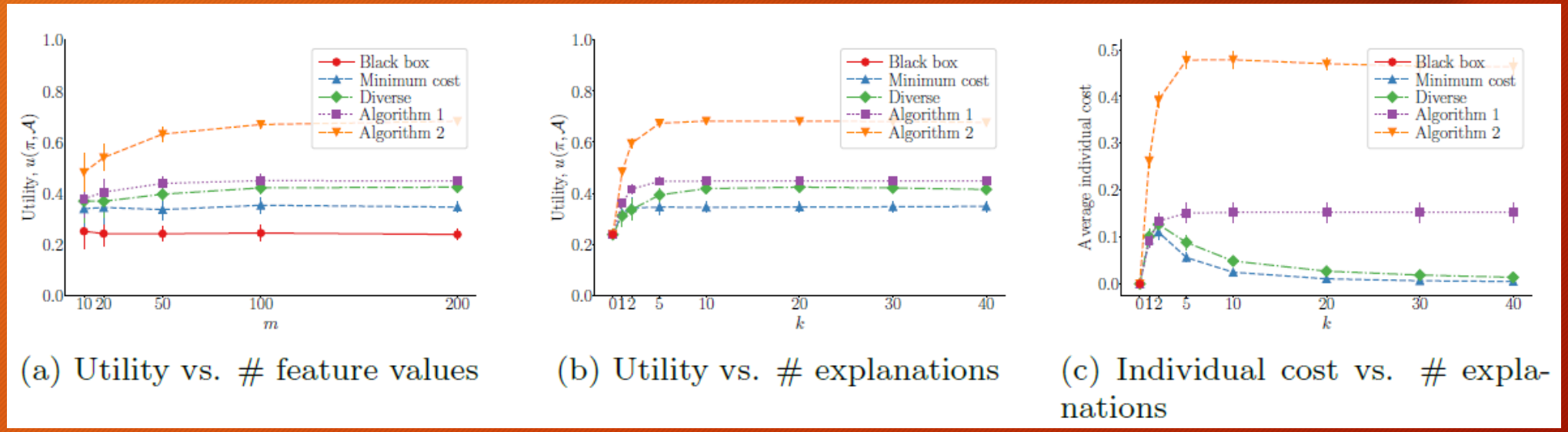
- The real data was taken from LendingClub data (a dataset of loans, applicant information, and their payments). A decision tree was trained as a predictor of defaulting. It was also used to generate approximations of some missing features.

Automated Decision Makers

The following decision makers were tested:

1. Black Box: The optimal decision policy in a non-strategic setting. No counterfactual explanations were given to individuals.
2. Minimum Cost: The optimal decision policy in a non-strategic setting, the counterfactual explanation with minimum cost to the individual was given.
3. Diverse: The optimal decision policy in a non-strategic setting, a diverse set of counterfactual explanations with minimum cost to the individual was used. Similar to previous work (Russel, 2019; Mothilal *et al.*, 2020).
4. Algorithm 1: Approximates problem 1, the optimal decision policy in a non-strategic setting was used.
5. Algorithm 2: Approximates problem 2.

Simulated



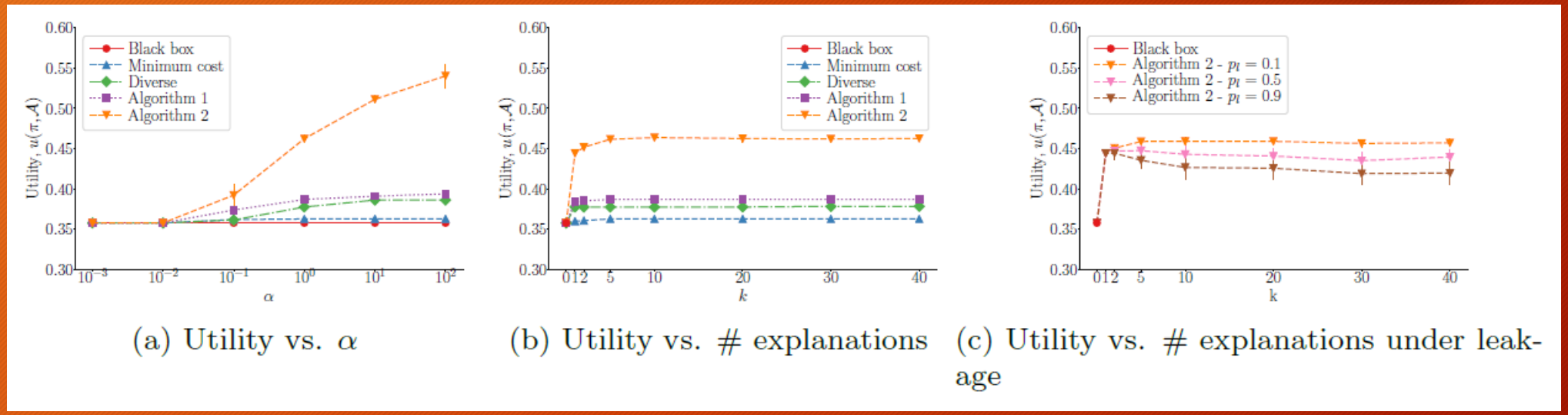
(a) Utility vs. # feature values

(b) Utility vs. # explanations

(c) Individual cost vs. # explanations

Figure taken from Tsirtsis *et al*. This figure shows the utility of the decision maker. It is clear that Algorithm 2 vastly outperforms other decision makers and Algorithm 1 outperforms all decision makers except Algorithm 2. Furthermore, increasing the size of the set of counterfactual explanations results in a higher utility (b). Lastly, individual cost increases with increasing utility of the decision maker (c). The authors argue the individual still benefits from this.

Real



(a) Utility vs. α (b) Utility vs. # explanations (c) Utility vs. # explanations under leakage

Figure taken from Tsirtsis *et al*. This figure shows the utility of the decision maker on real data. (a) shows that the gap between Algorithm 2 and other decision makers vastly increases when individuals are more likely to adapt. (b) shows that the decision makers perform similarly on real and simulated data. (c) shows that when individuals have a high probability of counterfactual explanations leaking, the decision maker is better off sharing providing less counterfactual explanations. This is meant to simulate communication between individuals

Conclusion

Tsirtsis *et al*. did relevant and impactful work in the following areas:

- Finding optimal set of counterfactual explanations for a decision policy. Approximation factor: $1 - \frac{1}{e}$. Proved NP-Hard.
- Finding optimal decision policy for a set of counterfactual explanations. Polynomial time.
- Jointly finding optimal decision policy and set of counterfactual explanations. Approximation factor: $\frac{1}{e}$. Proved NP-Hard.

Future Work

Their work has paved the way for many more researchers, some possible future works include:

- The same algorithms without a given cost function
- Real valued feature values.
- Counterfactual explanations with multiple feature values.
- Information sharing between individuals

References

- Niv Buchbinder, Moran Feldman, Joseph Naor, and Roy Schwartz. Submodular maximization with cardinality constraints. In *Proceedings of the twenty-fifth annual ACM-SIAM symposium on Discrete algorithms*, pages 1433-1452. SIAM, 2014.
- Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 2020.
- George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical programming*, 14(1):265-294, 1978.

References cont.

- Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 20-28, 2019.
- Stratis Tsirtsis and Manuel Gomez-Rodriguez. Decisions, Counterfactual Explanations and Strategic Behavior. *arXiv preprint arXiv:2002.04333*, 2020.