Report on "Decisions, Counterfactual Explanations and Strategic Behavior"

Ahsan Sanaullah

April 2020

1. Abstract

This paper covers "Decisions, Counterfactual Explanations and Strategic Behavior" by Stratis Tsirtsis and Manuel Gomez-Rodriguez. In this paper, Tsirtsis *et al*. propose the use of *counterfactual explanations* for maximization of utility with automated decision makers. They show that finding the optimal set of counterfactual explanations for a decision policy is NP-Hard. They also show that finding the optimal decision policy and set of counterfactual explanations simultaneously is NP-Hard. This is done by reducing Set Cover to these problems. However, Tsirtsis *et al*. showed that a greedy algorithm offers a $1 - \frac{1}{e}$ approximation factor for the first problem and a recent randomized algorithm offers a $\frac{1}{e}$ approximation factor for the second problem. Lastly, Tsirtsis *et al*. run experiments on real and simulated data to test the affect their approximation algorithms on utility.

2. Introduction

Important decisions are increasingly made by computers and automated decision makers. Automated decision makers typically have two stages: first, a predictive model predicts what would happen if it took each decision, then the decision maker picks a decision that maximizes a utility function according to the predictions. Many important institutions are using automated decision makers to assist or fully control the decision making process. These include banks giving out loans, judges granting bail, companies hiring employees, large investing firms trading stocks, and many more. These decisions have a big impact on society; therefore, decision makers are often pressured or legally obligated to provide reasons for their decisions. Because of the increased prevalence of automated decision makers, a law in the EU to be passed that gives individuals that have been part of a semi-automated decision process a *right-to-explanation*. This has caused a lot of work in the area of *interpretable machine learning*. Interpretable machine learning

attempts to explain the prediction of a predictive model. However, this is only one side of the coin. Tsirtsis *et al*. instead attacked the problem by attempting to provide optimal *counterfactual explanations*. A counterfactual explanation is a feature an individual can improve in order to receive a beneficial decision. Once an individual receives a counterfactual explanation, it is clear that they may take the advice and improve their features, this may increase the utility of the decision maker. Therefore, an optimal counterfactual explanation is one that maximizes the utility of the decision maker.

In their paper, Tsirtsis *et al*. attacked three main problems. They first showed that finding an optimal set of counterfactual explanations under a fixed decision policy is an NP-Hard problem. They did this with a reduction of the Set Cover problem. However, they show that a standard greedy algorithm achieves a $1 - \frac{1}{e}$ approximation factor on this problem. Then, they show that finding an optimal decision policy with a fixed set of counterfactual explanations can be done in polynomial time. Then they show that the problem of finding an optimal decision policy and set of counterfactual explanations is NP-Hard using a reduction of the Set Cover problem again. However, they show that a recent randomized algorithm provides an approximation factor of $\frac{1}{e}$. Lastly, they perform some experiments on their algorithms with some real and simulated data.

3. Definitions

An individual that desires a decision from an automated decision maker has a feature vector $x \in \{1, \dots, n\}^m$, a ground truth label $y \in \{0, 1\}$, and a decision $d(x) \in \{0, 1\}$. $d(x) = 1$ is the beneficial decision. Each decision ($d(x)$) is sampled from a decision policy (π), i.e. $d(x) \sim \pi(d \mid x)$. The ground truth (y) is sampled from a conditional probability distribution (P), i.e. $y \sim P(y \mid x)$. Let X be the set of all possible feature values. $X = \{x_1, x_2, x_3, \dots x_m\}$ where m is the number of possible feature values. X is indexed such that the contribution of feature value to the ground truth is in decreasing order, i.e. $\forall\, 1 \leq i, j \leq m,\ i < j \rightarrow P(y = 1 \mid x_i) \geq P(y = 1 \mid x_j)$. When an individual with an attribute x_i receives a counterfactual explanation $\varepsilon(x_i)$ from the decision maker, it is a guarantee that the individual will receive a beneficial

decision if it changes its feature from x_i to $\varepsilon(x_i)$. $c\big(x, \varepsilon(x_i)\big)$ is the cost an individual x incurs by switching from x_i to $\varepsilon(x_i)$. $b(\pi, x)$ is the immediate benefit x receives from a decision policy π. The *Region of adaptation* ($R(x)$) is the feature values it would change if told to, i.e. $R(x_i) \coloneqq \{x_j \in X \mid b(\pi, x_j) - c(x_i, x_j) \geq b(\pi, x_i)\}$. An individuals best response is to change their feature value from x_i to $\varepsilon(x_i)$ if and only if $\varepsilon(x_i)$ if in its region of adaption of x_i, i.e. $\varepsilon(x_i) \in R(x_i)$. $u(\pi, A)$ is the expected utility of the decision maker. The authors assume the decision maker is rational and wants to maximize their utility. Therefore, given a feature value x_i and a set of counterfactual explanation A, if the region of adaptation of x_i includes something in A, the decision maker will choose $\varepsilon(x_i)$ that maximizes their expected utility assuming individuals adapt x_i to $\varepsilon(x_i)$. If the region of adaptation does not intersect with the set of counterfactual explanations, the decision maker arbitrarily picks an explanation since the individuals will not adapt regardless of choice. Tsirtsis *et al*. attack the problems of finding the optimal set of counterfactual explanations for a given policy, finding the optimal policy for a given set of counterfactual explanations, and jointly finding the optimal policy and set of counterfactual explanations.

4. Finding the optimal set of counterfactual explanations for a decision policy

Tsirtsis *et* al. attempt to solve the following problem: given a decision policy π and an upper bound on the size of the set of counterfactual explanations k, find a set of counterfactual explanations A^* s.t. $|A^*| \leq k$ and $u(\pi, A^*)$ is maximized. The authors show that assuming the decision maker is rational, if the decision policy is outcome monotonic[1] or deterministic, this is an NP-Hard problem. They do this by providing a polynomial time reduction from set cover to it, i.e. Set Cover $\leq_p$ our problem. However, the authors then prove that the problem has nice properties that allow a standard greedy algorithm to achieve a $1 - \frac{1}{e}$ approximation factor on it. These properties are non-negativity, monotonicity, and submodularity. The

[1] Outcome monotonic means that if an individual's ground-truth probability is higher than another's, then its probability of getting a beneficial decision is higher as well. This constraint is reasonable in a perfect information setting, however it may not always be the case in the real world.

ALGORITHM 1: Greedy algorithm

Input: Ground set of counterfactual explanations $\mathcal{P}_\pi$, maximum number of counterfactual explanations k and utility function f

Output: Set of counterfactual explanations $\mathcal{A}$

$\mathcal{A} \leftarrow \varnothing$
while $|\mathcal{A}| \leq k$ **do**
  $x^* \leftarrow \text{argmax}_{x \in \mathcal{P}_\pi \setminus \mathcal{A}} f(\mathcal{A} \cup \{x\}) - f(\mathcal{A})$
  $\mathcal{A} \leftarrow \mathcal{A} \cup \{x^*\}$
end while
return $\mathcal{A}$

greedy algorithm of Nemhauser *et al*. provides a set A s.t. $u(\pi, A) \geq \left(1 - \frac{1}{e}\right) u(\pi, A^*)$, where A^* is the optimal set of counterfactual explanations (Nemhauser *et al*., 1978). Algorithm 1 shows their procedure, it has time complexity $O(k^2m^2)$ because of $O(km)$ utility function calls which can be calculated in $O(km)$ per call.

5. Finding the optimal decision policy and set of counterfactual explanations

The authors show that the optimal decision policy for a given set of counterfactual explanations is deterministic and can be found in polynomial time. However, the authors believe that jointly optimizing decision policy and set of counterfactual explanations offers greater gains in utility for the decision maker than optimizing one at a time. Therefore, the authors attack the following problem: given an upper bound k on the size of the set of counterfactual explanations, find π^* (optimal decision policy) and A^*(optimal set of counterfactual explanations) such that $|A^*| \leq k$ and $u(\pi^*, A^*)$ is maximized. The authors claim this can be proved using a small modification of their other reduction. Therefore, Set Cover $\leq_p$ this problem. Define the function h as $h(A) = u(\pi_A^*, A)$ where π_A^* is the optimal decision policy for the set of counterfactual explanations A. The authors have shown that the π_A^* can be computed in $O(km)$ time. The authors show that the function h is non-negative and submodular. Unfortunately, the standard greedy algorithm can't be applied to h since it is not outcome monotonic. However, a recent randomized algorithm from Bunchbinder *et al*. provides a $\frac{1}{e}$ approximation factor, i.e. the algorithm returns a set A s.t. $h(A) \leq \frac{1}{e} h(A^*)$, where A^* and $\pi_{A^*}^*$ are the optimal set of counterfactual explanations and decision policy respectively (Buchbinder *et al*.,

ALGORITHM 2: Randomized algorithm

Input: Ground set of counterfactual explanations $\mathcal{Y}$, maximum number of counterfactual explanations k and utility function f

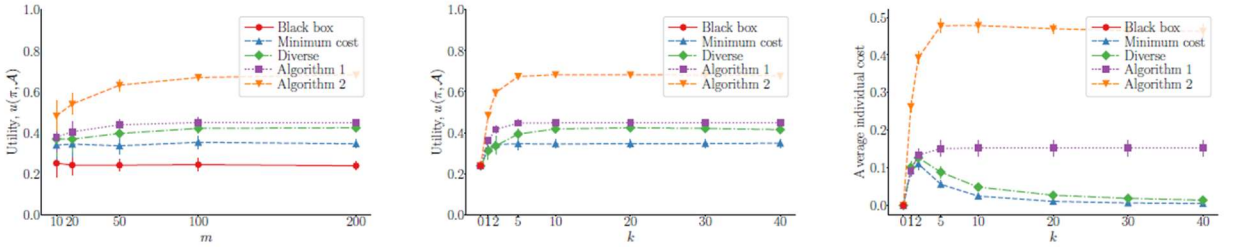
Output: Set of counterfactual explanations $\mathcal{A}$

$\mathcal{A} \leftarrow \varnothing$
while $|\mathcal{A}| \leq k$ **do**
  $\mathcal{B} \leftarrow \text{GetTopK}(\mathcal{Y}, \mathcal{A}, f)$
  $x^* \sim \mathcal{B}$
  $\mathcal{A} \leftarrow \mathcal{A} \cup \{x^*\}$
end while
return $\mathcal{A}$

2014). Algorithm 2 shows the algorithm for this problem, its time complexity follows from the same reasoning as Algorithm 1's, there are $O(km)$ calls to the utility function, therefore the time complexity is $O(k^2m^2)$.

6. Results

The authors performed experiments on simulated and real data. These experiments measured the utility of the decision maker with 5 different decision makers. They tested black box, minimum cost, diverse, algorithm 1, and algorithm 2. The black box decision maker was the optimal decision policy in a setting without communication between decision maker and individual (referred to as a non-strategic setting), no counterfactual explanations were given to individuals. The minimum cost decision maker used the optimal decision policy in a non-strategic setting and provided counterfactual explanations that minimized the cost to the individual. The diverse decision maker used the optimal decision policy in a non-strategic setting and provided a diverse set of counterfactual explanations that minimized the cost to the individual, this is similar to previous work (Russel, 2019; Mothilal *et al*., 2020). The decision maker that used algorithm 1 used the optimal decision policy of a non-strategic setting, the set of counterfactual explanations was computed using algorithm 1. Individuals were given counterfactual explanations that maximized the probability of their ground truth being beneficial. Lastly, the decision maker that used algorithm 2 used the optimal decision policy for the set of counterfactual explanations returned by algorithm 2, individuals were again given counterfactual explanations that maximized the probability of their ground truth being beneficial.

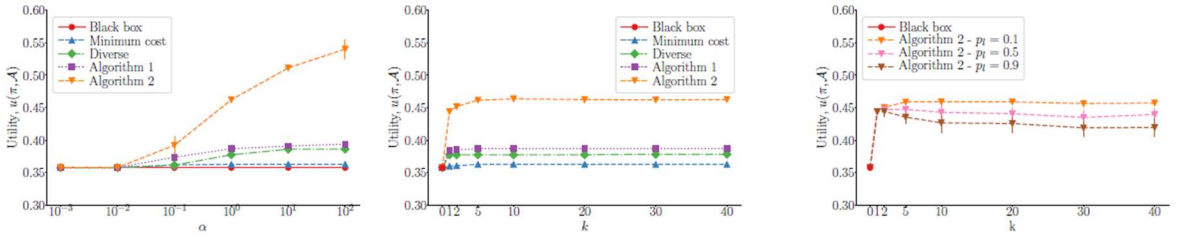


(a) Utility vs. # feature values (b) Utility vs. # explanations (c) Individual cost vs. # explanations

Figure 1.

The authors first experimented on synthetic data. This data was generated as follows: $x = \{0, \dots, m-1\}, P(x=i) = \frac{p_i}{\sum_j p_j}$. p_i is sampled from a Gaussian distribution $N(\mu = 0.5, \sigma = 0.1)$ truncated below 0. $P(y=1|x) \sim U[0,1], c(x_i, x_j) \sim U[0,1]$ for half of the pairs and $c(x_i, x_j) = 2$ for the rest of the pairs. Then the utility of the decision maker was calculated after the procedures described previously ran. The experiment was run 20 times. The results (taken from the authors' paper) can be seen in Figure 1. Fig. 1a shows the utility of the decision maker vs m, Fig. 1b shows the utility of the decision maker vs k, and Fig 1c shows the cost incurred by the individual vs k. The figure clearly shows that Algorithm 2 vastly outperforms all other decision makers in terms of maximizing decision maker utility. Furthermore, Algorithm 1 outperforms all decision makers except for Algorithm 2. Figure 1b shows that decision makers benefit from being more open about the decision making process. Lastly, increased utility of the decision maker is directly related to increased cost to the individual, however, the authors argue this is beneficial to the individual by making the probability their ground truth is beneficial higher.

The authors also experimented on real data. They used the LendingClub dataset, this is a public dataset available at https://www.kaggle.com/wordsforthewise/lending-club/version/3. At the time, it had information accepted and rejected loans in LendingClub from 2007 to 2018. The dataset has various features including payment information, FICO credit scores, and current loan status for accepted applicants. The authors first built a decision tree classifier that predicts if an applicant will fully pay off a loan based on loan amount, employment length, state of residence, debt to income ratio, zip code, and FICO score. Their classifier achieved a 90% accuracy. The feature of an individual was the leaf of the decision tree it



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

Figure 2.

maps to. The probability $P(y = 1|x)$ is approximated using the decision tree. The cost $c(x_i, x_j)$ is calculated using the features that x_i and x_j are mapped to. The cost is also scaled by $\frac{1}{\alpha}$. Then, the same experiment was run as previously. The results can be seen in Figure 2. Figure 2a shows the utility of the decision maker vs α, higher α means individuals are more willing to change their features. Figure 2b shows the utility vs the number of counterfactual explanations (k). Figure 3c tests a new idea, in this experiment, individuals receive a random extra counterfactual explanation with probability p_l (probability of leakage). Individuals take the explanation that benefits them the most. This is meant to simulate the idea of individuals communicating their counterfactual explanations with each other. Figure 1a shows that the benefit of simultaneously optimizing decision policy and counterfactual explanations increases if individuals are more likely to adapt. Figure 1b confirms that the decision makers perform similarly on real data and simulated data. Figure 1c tells us that if the probability of leakage is high, the decision maker is better off giving fewer counterfactual explanations.

7. Conclusion

The work of Tsirtsis *et al*. is very important, it impacts important aspects of society and begins the path towards further explanation of automated decision making. Automated decision making will only become a bigger part of society, explaining its decision making process and how to achieve a better decision is a crucial human aspect of decision making that is currently missing. Their work is very replicable and has strong theoretical foundations. The authors could have made their work more readable. Certain concepts were not clearly explained. When formally defining the problem, the authors skipped into defining the input

to the algorithm. Instead, they should have started with an intuitive, and then a formal definition of the problem.

Their paper leaves a multitude of avenues for further research. The authors outlined some of these. For example, the authors assume only one feature value per individual (one of the things not clear in the definition), this could be extended for multiple feature values per individual. Secondly, the cost of switching feature values is usually not given in the real world, it would be interesting to see this assumption removed. Also, the current method uses discrete feature values, a method for real valued feature values would be more useful. Lastly, the authors suggest that a method that allows information sharing between individuals is an interesting and useful goal. Further ideas for extension of this work include counterfactual explanations with multiple features and a decision maker that doesn't limit itself to a set of counterfactual explanations, instead finding the optimal counterfactual explanation per individual.

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 TransparePncy*, 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.

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.