Towards Bridging the Gaps between the Right to Explanation and the Right to be Forgotten

Satyapriya Krishna, Jiaqi Ma, Himabindu Lakkaraju

Motivation/Previous Work

- Right to an explanation: Given a model prediction, how can we provide a reason for this prediction to a user? (Algorithmic Recourse)
  - Multi-Objective Counterfactual (Dandel et al 2020)
  - MACE (model-agnostic counterfactual explanations) (Karimi et al 2020)

- Right to be forgotten: User can ask to have personal data removed from databases and models
  - Descent-to-delete (Neel et al 2021)
  - Approximate data deletion (Izzo et al 2021)

Motivation/Previous Work

- Pawelczyk et al. 2022: tradeoff between right to explanation and to be forgotten
  - Forgetting ⇒ Model Changing ⇒ Explanation not valid ⇒ Right to explanation not met
- Trade-off stems from fact current explanation methods do not take into account underlying model changing
  - But! Wait! We know there are methods that do this like ROAR by Upadhyay et al. 2021
  - Authors say these methods assume certain model changes, but we can't know how a model may change as a result of forgetting
- Pawelczyk et al. highlighted the problem there is a need for a solution!

Contributions

- First algorithmic framework (ROCERF) to address tradeoff between right to explanation and to be forgotten

- ROCERF not only is able to bridge the gap but also outperform existing counterfactual explanation methods

ROCERF Framework

- **RO**bust **C**ounterfactual **E**xplanations under the **R**ight to be **F**orgotten
  - CFE as an optimization problem
  - CFE as an optimization problem under data removal
  - This is impractical! Efficient approximation
  - Theoretical bound on cost, and validity

ROCERF Framework: Notation

Training Data $D = \{(x_i, y_i)\}_{i=1}^{n}, x_i \in X, y_i \in \{-1, +1\}$

Data weight vector $\boldsymbol{w} \in \{0, 1\}^n$

Classifier Family $f_\theta : X \to \mathbb{R}, \theta \in \Theta$

$$y = \begin{cases} +1 & f_\theta(x) \geq 0 \\ -1 & f_\theta(x) < 0 \end{cases}$$

$$w_i = \begin{cases} 1 & \text{data point in training data} \\ 0 & \text{data point not in training data} \end{cases}$$

$\boldsymbol{w} = \mathbf{1} \Rightarrow$ no data point is being removed

ROCERF Framework: Notation

f_{θ_1} - trained on all data D, f_{θ_w} - trained on D_w

$$\theta_1 = \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} l_i(\theta)$$

$$\theta_w = \arg\min_{\theta \in \Theta} \frac{1}{||w||_1} \sum_{i=1}^{n} w_i l_i(\theta)$$

$$g_i(\theta) := \frac{\partial l_i(\theta)}{\partial \theta}$$

$$h_i(\theta) := \frac{\partial g_i(\theta)}{\partial \theta^T}$$

$$H := \frac{1}{n} \sum_{i=1}^{n} h_i(\theta_1)$$

Counterfactual Explanation

- Optimisation problem - find valid counterfactual with minimum cost

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \|x - x_0\|_2 \\ \text{subject to} \quad & f_{\hat{\theta}_1}(x) \geq 0. \end{aligned}$$

K-Removal Robust CFE

- k-RR CFE := CFE that is valid upon removal of any k data points

$$\mathcal{W}^{(k)} = \{w \in \{0,1\}^n : \|w\|_1 = n - k\}$$

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \|x - x_0\|_2 \\ \text{subject to} \quad & f_{\hat{\theta}_w}(x) \geq 0, \forall w \in \mathcal{W}^{(k)}. \end{aligned}$$

- Naive way - Train $\binom{n}{k}$ classifiers, and optimise with $\binom{n}{k}$ constraints
- Computationally impractical!

k-Robust Removal CFE: Efficient Approximation

- Note that classifier $f_{\hat{\theta}_w}$ is a function of both x and data weight vector **w**
- Fix x and approximate $f_{\hat{\theta}_w}$ using Taylor series with respect to **w**
- First order approximation!

$$\tilde{f}_{\hat{\theta}_w}(x) = f_{\hat{\theta}_1}(x) + \frac{\partial f_{\hat{\theta}_w}(x)}{\partial w}\Big|_{w=\mathbf{1}} (w - \mathbf{1})$$

- Accounting for approximation error in the optimization problem

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \|x - x_0\|_2 \\ \text{subject to} \quad & \tilde{f}_{\hat{\theta}_w}(x) \geq \delta, \forall w \in \mathcal{W}^{(k)}, \end{aligned}$$

k-RR CFE: Efficient Approximation

$$\tilde{f}_{\hat{\theta}_w}(x) = f_{\hat{\theta}_1}(x) + \frac{\partial f_{\hat{\theta}_w}(x)}{\partial w}\Big|_{w=\mathbf{1}}(w-\mathbf{1})$$

$$= f_{\hat{\theta}_1}(x) + \frac{\partial f_\theta(x)}{\partial \theta}\Big|_{\theta=\hat{\theta}_1} \frac{\partial \hat{\theta}_w}{\partial w}\Big|_{w=\mathbf{1}}(w-\mathbf{1}),$$

$$\frac{\partial \hat{\theta}_w}{\partial w}\Big|_{w=\mathbf{1}}(w-\mathbf{1}) = \frac{1}{n}\sum_{i:w_i=0} H^{-1} g_i(\hat{\theta}_\mathbf{1}). \qquad \beta(x) := \left(\frac{\partial f_\theta(x)}{\partial \theta}\Big|_{\theta=\hat{\theta}_1}\right)^T$$

Giordano et al. (2019) “A swiss army infinitesimal jackknife”
The proof is left as an exercise to the reader!

$$\tilde{f}_{\hat{\theta}_w}(x) = f_{\hat{\theta}_1}(x) + \frac{1}{n}\sum_{i:w_i=0} \beta(x)^T H^{-1} g_i(\hat{\theta}_\mathbf{1}),$$

k-RR CFE: Efficient Approximation

$$\tilde{f}_{\hat{\theta}_w}(x) = f_{\hat{\theta}_1}(x) + \frac{1}{n} \sum_{i:w_i=0} \beta(x)^T H^{-1} g_i(\hat{\theta}_1),$$

- We still have $\binom{n}{k}$ constraints
- The term $\beta(x)^T H^{-1} g_i(\hat{\theta}_1)$ is independent of data weight vector **w**
- Pick the one tightest constraint!

k-RR CFE: Efficient Approximation

$$\tilde{f}_{\hat{\theta}_w}(x) = f_{\hat{\theta}_1}(x) + \frac{1}{n} \sum_{i:w_i=0} \beta(x)^T H^{-1} g_i(\hat{\theta}_1),$$

$$\mathcal{A}(x) := \{\beta(x)^T H^{-1} g_i(\hat{\theta}_1)\}$$

$$f_{\mathcal{A}}^{(k)}(x) := f_{\hat{\theta}_1}(x) + \frac{1}{n} \min_{\mathcal{B} \subseteq \mathcal{A}(x), |\mathcal{B}|=k} \sum_{b \in \mathcal{B}} b,$$

$$\begin{aligned} \min_{x \in \mathcal{X}} \quad & \|x - x_0\|_2 \\ \text{subject to} \quad & f_{\mathcal{A}}^{(k)}(x) \geq \delta. \end{aligned}$$

k-RR CFE: Efficient Approximation

- One last step!
- Solving the constrained optimization problem.
- Penalty method

$$\phi(z) := \max(z, 0)^2$$

$$\min_{x \in \mathcal{X}} \quad J_t(x) = \lambda_t \phi(\delta - f_{\mathcal{A}}^{(k)}(x)) + \|x - x_0\|_2,$$

k-RR CFE: Practical considerations

$$f_{\mathcal{A}}^{(k)}(x) := f_{\hat{\theta}_1}(x) + \frac{1}{n} \min_{\mathcal{B} \subseteq \mathcal{A}(x), |\mathcal{B}|=k} \sum_{b \in \mathcal{B}} b,$$

- Computational cost
  - O(n)
- Hyperparameters
  - k - number of data removals
  - δ - approximation error
- Linear Models
  - We can avoid backward pass!

$$\tilde{f}_{\hat{\theta}_w}(x) = f_{\hat{\theta}_1}(x) + \frac{1}{n} \sum_{i:w_i=0} \beta(x)^T H^{-1} g_i(\hat{\theta}_1), \longrightarrow \tilde{f}_{\hat{\theta}_w}(x) = \hat{\theta}_1^T x + \frac{1}{n} \sum_{i:w_i=0} x^T H^{-1} g_i(\hat{\theta}_1),$$

Theoretical Analysis of Validity and Cost

Validity - Accuracy of the counterfactual explanation with respect to the retrained model.

$$\frac{1}{M}\sum_{w\in\mathcal{V}}\frac{1}{|\mathcal{T}|}\sum_{x\in\mathcal{T}}\mathbb{1}[f_{\hat{\theta}_w}(c(x))=1],$$

Cost - Distance measure for how far the the new input is from the original to satisfy the counterfactual requirement.

$$\frac{1}{|\mathcal{T}|}\sum_{x\in\mathcal{T}}\|c(x)-x\|_2.$$

Analysis on Linear Models

For regularized logistic regression model with loss function defined:

$$l_i(\theta) = \log(1+\exp(-y_i\theta^T x_i))+\gamma\|\theta\|_2^2$$

The result states that the additional cost needed to achieve robust validity has an upper bound of O(k/n) with theoretical guarantees for validity.

$$\|\tilde{x}_0^{(k)} - x_0\|_2 \leq \|\tilde{x}_0 - x_0\|_2 + \frac{kC}{n\|\hat{\theta}_1\|_2},$$

Assumptions for Nonlinear models

$\sup_{\theta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \|h_i(\theta)\|_F \le C_1;$

$\sup_{\theta\in\Theta} \|g_i(\theta)\|_2 \le C_2, i = 1, \ldots, n;$

$\sup_{x\in\mathcal{X}} \|\beta(x)\|_2 \le C_3;$

$H(\theta) := \frac{1}{n}\sum_{i=1}^{n} h_i(\theta)$ is nonsingular and

$\sup_{\theta\in\Theta} \|H(\theta)^{-1}\|_{\text{op}} \le C_4;$

there exists suitable $\Delta > 0$, such that

$$\sup_{\|\theta-\hat{\theta}_1\|_2<\Delta} \frac{1}{n}\sum_{i=1}^{n} \|h_i(\theta) - h_i(\hat{\theta}_1)\|_F \le C_5\|\theta - \hat{\theta}_1\|_2.$$

Analysis on Nonlinear Models

With certain regularity assumptions, the result provides theoretical guarantees where C is a constant independent of n.

$$\begin{aligned} &\|\tilde{x}_0^{(k)} - x_0\|_2 \\ \leq &\|\tilde{x}_0 - x_0\|_2 + \min_{\substack{x \in \mathcal{X}, \\ f_{\hat{\theta}_1}(x) - f_{\hat{\theta}_1}(\tilde{x}_0) \geq \frac{kC}{n}}} \|x - \tilde{x}_0\|_2, \end{aligned}$$

If the function is Mu-strongly convex then,

$$\|\tilde{x}_0^{(k)} - x_0\|_2 \leq \|\tilde{x}_0 - x_0\|_2 + \frac{2kC}{n\mu}.$$

Experimental Setup

- Datasets - Three real-world binary classification datasets collected from high-stakes decision making scenarios.

- Predictive Models - Regularized logistic regression and 3-layer fully-connected feedforward neural network.

- Baseline Methods for Comparison - SCFE, C-CHVAE, and ROAR

Experimental Setup

- Validity is evaluated by randomly removing a small fraction (α) of the training data points (varies between 0.5% to 5%).

- Process is repeated M times (fixed to 100).

- Hyperparameters - k as 0.5% of the training set size and fix δ = 0

Experimental Results

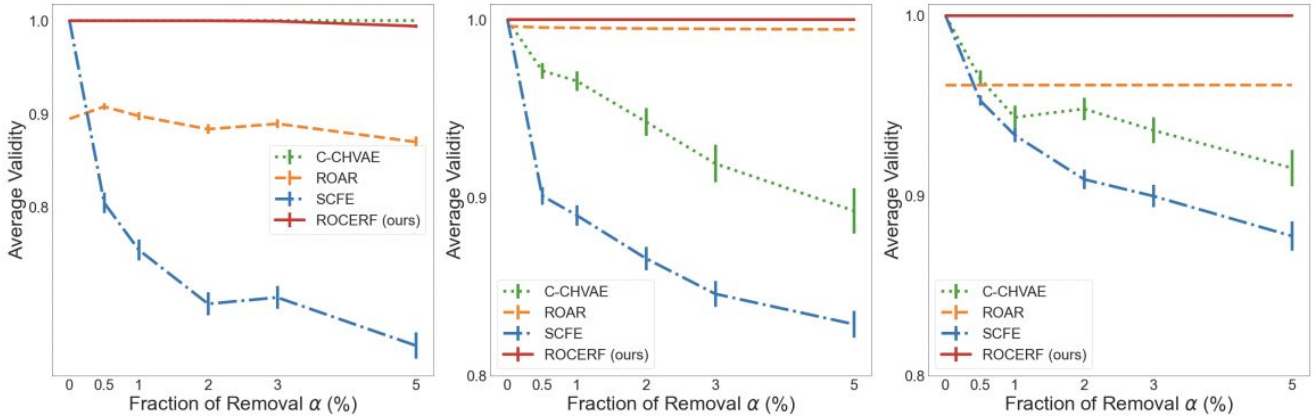


Figure 1. Average validity of different counterfactual explanation methods applied to logistic regression models on three datasets. In each figure, the x-axis corresponds to the fraction of data removal α and the y-axis corresponds to the average validity. The error bars indicate the standard errors across $M = 100$ trials with each trial having an α fraction of training data points randomly removed.

Experimental Results

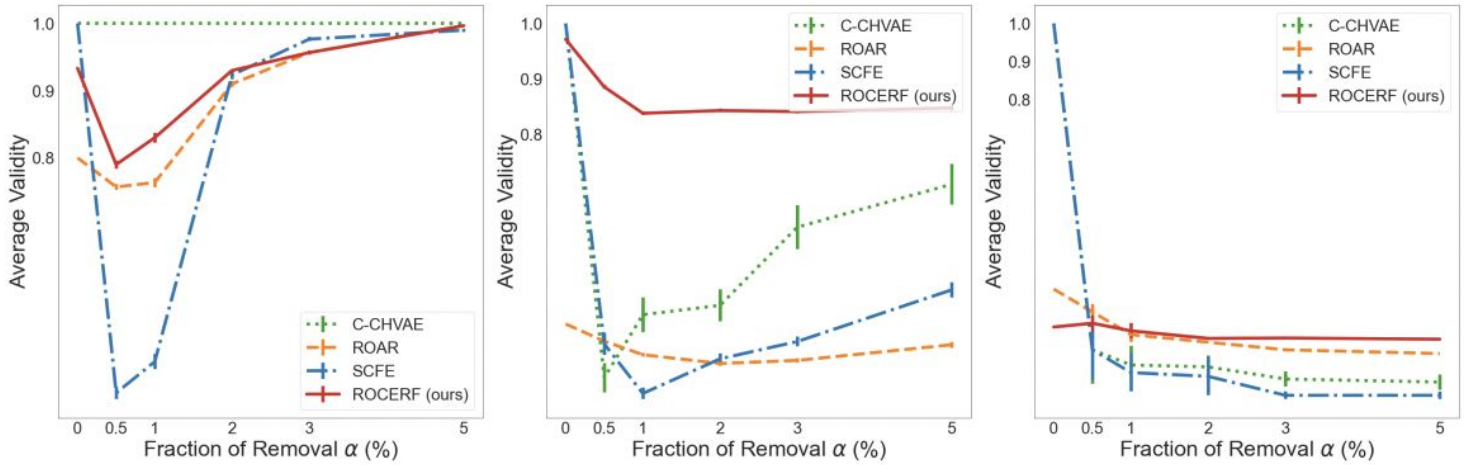


Figure 2. Average validity of different counterfactual explanation methods applied to neural network models on three datasets. See Figure 1 for more details about the plot setting.

Experimental Results

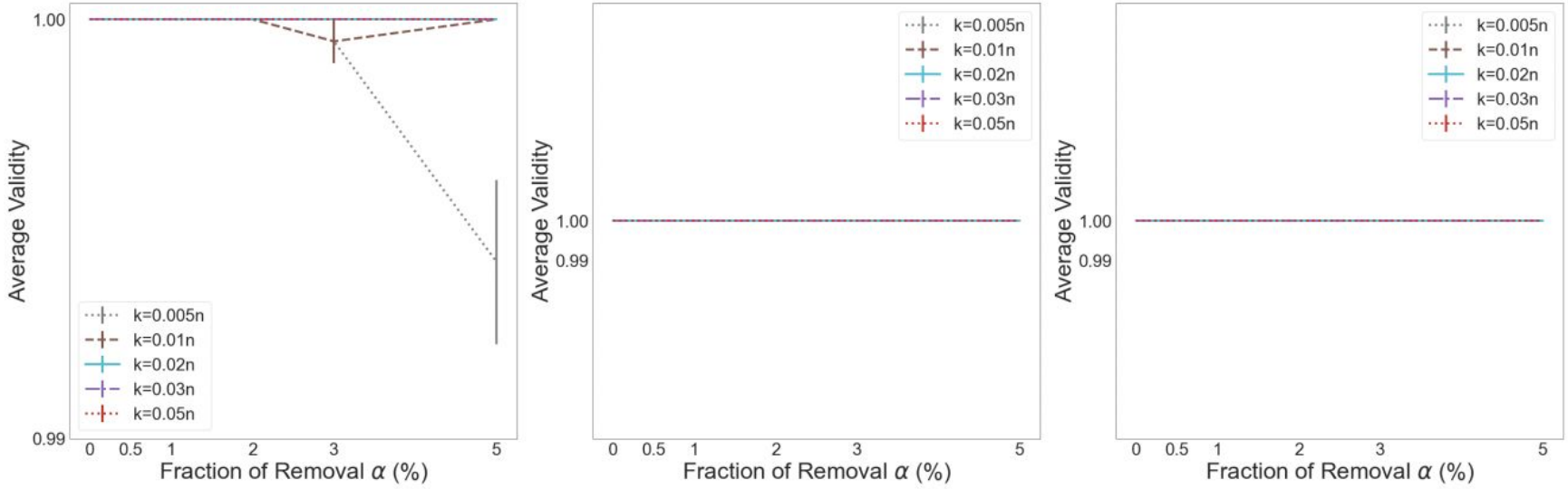
Methods	German Credit	COMPAS	Adult
SCFE	0.82 ± 0.12	0.78 ± 0.02	1.04 ± 0.006
C-CHVAE	8.51 ± 0.38	5.93 ± 0.11	3.79 ± 0.013
ROAR	1.45 ± 0.09	1.08 ± 0.01	1.07 ± 0.006
ROCERF (ours)	1.35 ± 0.14	0.87 ± 0.02	1.14 ± 0.006

Table 1. Average cost of different recourse methods applied to logistic regression models on three datasets. The cost is measured in terms of L2 norm.

Methods	German Credit	COMPAS	Adult
SCFE	1.18 ± 0.08	0.97 ± 0.11	1.00 ± 0.09
C-CHVAE	4.45 ± 0.18	5.98 ± 0.12	8.83 ± 0.31
ROAR	3.84 ± 0.33	1.09 ± 0.13	4.07 ± 0.55
ROCERF (ours)	2.76 ± 0.22	3.07 ± 0.08	4.06 ± 0.52

Table 2. Average cost of different recourse methods applied to neural network models on three datasets. The cost is measured in terms of L2 norm.

Experimental Results



(a) German Credit. (b) COMPAS. (c) Adult.

Figure 3. Sensitivity analysis with respect to the hyperparameter k.

Conclusions

- Propose ROCERF: Initial attempt to bridge the gap between right to be forgotten and right to explanation
- ROCERF is provably robust to model updates as triggered by data deletion requests
- ROCERF not only comes with theoretical guarantees on validity and cost but also outperforms baselines so is able to bridge this gap

Limitations/Future Work

- Non-linear models: Uses linear approximations under certain conditions, do these reflect non-linear model behavior? (common throughout recourse literature)
  - Some results show our assumptions may not hold for more complex non-linear neural networks (validity improves in the models as more data points are "forgotten")
- ROAR and ROCERF track closely for some of the results in this paper
  - Authors use ROAR with same hyperparameters as original paper, one could argue that ROAR could encapsulate ROCERF
- Why would the performance of ROCERF not vary more with k? Test higher alphas?
  - Intuitively as more points are removed in the framework the model should be less robust but seems like it doesn't matter?

Discussion

- This is an example of when we were able to bridge the gap (as compared to privacy vs explanation where they seem more opposed). Any other tradeoffs that we learned about or that you think are able to be bridged?

- Given ROAR exists, do we need ROCERF? Could one not argue that the deletion of data parameters causes a perturbation of model parameters?
    - Why didn't the previous paper (Pawelczyk et al. 2022) use ROAR?

- Why has recourse literature focused on linear approximations when results show that non-linear models exhibit non-intuitive behavior?