ĐẠI HỌC QUỐC GIA THÀNH PHỐ HỒ CHÍ MINH
TRƯỜNG ĐẠI HỌC KHOA HỌC TỰ NHIÊN
KHOA CÔNG NGHỆ THÔNG TIN

CÂY GIẢI THÍCH PHẢN THỰC

Báo cáo nghiên cứu Cây quyết định

Trương Quốc Cường – 23127333
Trần Cẩm Huy – 23127056
Nguyễn Tấn Phát – 23127449
Lưu Huy Minh Quang – 23127016

Ngày 28 tháng 8 năm 2025

Nội dung

Nền tảng: Giải thích phản thực (CE)

Giải thích "hành động"để đạt được kết quả dự đoán mong muốn

Phương pháp giải thích hậu nghiệm để đưa ra các **giải thích cục bộ** cho một bộ phân loại.

Giải thích phản thực (Wachter **andothers**, 2018): Đối với mỗi trường hợp $x \in \mathcal{X}$, CE tìm một **hành động** a^* sao cho kết quả dự đoán của **bộ phân loại** $f : \mathcal{X} \to \mathcal{Y}$ **thay đổi thành nhãn mục tiêu** $y^* \in \mathcal{Y}$, đồng thời tối thiểu hóa chi phí thực hiện hành động:

$$a^* = \arg\min_{a \in \mathcal{A}} \boxed{c(a|x)} \quad \text{với điều kiện} \quad f(x + a) = y^*$$

hàm chi phí
(Max Percentile Shift Ustun **andothers**, 2019)

Nội dung

Mục tiêu nghiên cứu: Giải thích phản thực cho "nhiều" trường hợp

Gán hành động cho nhiều trường hợp $X \subset \mathcal{X}$ đồng thời

Các hành động a tối ưu cho từng cá thể x **không nhất thiết** do chính cá thể đó thực hiện (Karimi **andothers**, 2021).

Ví dụ: *Dự đoán nguy cơ nghỉ việc*. Một công ty gán các hành động cho nhân viên để giảm nguy cơ nghỉ việc.

Một hành động a cho một cá thể x (ví dụ: tăng lương) **có thể ảnh hưởng đến các cá thể khác** (ví dụ: thay đổi hệ thống lương trong công ty).

$\Longrightarrow$ Trong trường hợp này, việc tối ưu hành động riêng lẻ cho từng cá thể là chưa đủ.

Nội dung

Tổng quan đóng góp

Một framework CE mới có khả năng gán hành động cho nhiều trường hợp

Nghiên cứu đề xuất **Cây Giải thích phản thực (CET)**, có khả năng gán hành động cho các trường hợp đầu vào bằng **cây quyết định**.

Đặc tính:

- **Tính minh bạch:** Giải thích được hành động áp dụng cho tất cả các trường hợp trong không gian đầu vào.
- **Tính nhất quán:** Không tồn tại xung đột trong lý do gán hành động giữa các trường hợp.

Tổng quan đóng góp (tiếp)

Đề xuất một **thuật toán hiệu quả để xây dựng CET** từ các trường hợp cho trước, dựa trên tìm kiếm cục bộ ngẫu nhiên và MILO.
Thực nghiệm xác nhận sự **hiệu quả và khả năng diễn giải** của CET.

Định hướng nghiên cứu:

- Mở rộng CET cho dữ liệu phi tuyến và nhiều đặc trưng.
- Kết hợp với các loại giải thích khác.
- Tối ưu hóa hiệu năng.

Nội dung

So sánh với các hướng tiếp cận trước đó

Counterfactual Explanation (CE):

- Cung cấp hành động khả thi cho từng cá thể.
- Hạn chế: chỉ **cục bộ**, thiếu **phủ toàn cục** và đôi khi thiếu ổn định.

Framework toàn cục:

- **AReS** – tập luật hai cấp, nhưng không bao phủ đầy đủ, thiếu minh bạch/nhất quán.
- **MAME**, **GIME** – phân cụm, mô hình chủ đề $\rightarrow$ có minh bạch nhưng không nhất quán.

So sánh các hướng tiếp cận

Phương pháp	Mục tiêu	Phủ toàn cục	Minh bạch	Nhất quán
CE (Wachter **andothers**, 2018; Karimi **andothers**, 2021)	Local	Không	Có	Có
CE (Ustun **andothers**, 2019; Mothilal **andothers**, 2020)	Local	Không	Có	Có
AReS (Rawal **and** Lakkaraju, 2020)	Global	Có, chưa đủ	Không	Không
MAME (Ramamurthy **andothers**, 2020)	Global	Có	Có	Không
GIME (Gao **andothers**, 2021_08)	Global	Có	Có	Không
CET	Global	Có	Có	Có

Kết luận: CET vượt trội nhờ đồng thời đảm bảo **minh bạch, nhất quán và phủ toàn cục**.

Vị trí của CET trong bối cảnh nghiên cứu

- **Minh bạch:** cấu trúc cây, dễ hiểu, mỗi lá = một hành động.
- **Nhất quán:** mỗi cá thể chỉ gán đúng một hành động.
- **Phủ toàn cục:** toàn bộ không gian đầu vào đều được bao phủ.

Kết luận: CET kết hợp **cây quyết định cổ điển** và **CE hiện đại**, tạo framework vừa **có thể giải thích được**, vừa **có thể hành động toàn cục**.

Nội dung

Yêu cầu của CE cho nhiều trường hợp

Học một mô hình gán hành động minh bạch và nhất quán

- **Tính minh bạch (Transparency) (Rawal and Lakkaraju, 2020):** Cần giải thích được *cách các hành động được quyết định cho toàn bộ cá thể*.
- **Tính nhất quán (Consistency) (Rudin and Shaposhnik, 2023):** Cần cung cấp lý do của hành động *không mâu thuẫn giữa các cá thể*. *Ví dụ:* Quy tắc "Age $>$ 35 & Dept.=Sales" có thể gây mâu thuẫn nếu hai nhân viên thoả điều kiện nhưng được gán hành động khác nhau.

Cách tiếp cận

Phát triển mô hình CE đáp ứng tính minh bạch và nhất quán

- **Ý tưởng 1:** Thiết kế một mô hình gán hành động hiệu quả trên toàn bộ không gian đầu vào $\mathcal{X}$ theo cách *minh bạch* và *nhất quán*.
- **Ý tưởng 2:** Xây dựng một thuật toán để học mô hình đó từ tập nhiều trường hợp $X \subset \mathcal{X}$.

Nội dung

Định nghĩa và Ưu điểm của CET

CET: Cây quyết định gán hành động hiệu quả trên toàn bộ không gian đầu vào

Định nghĩa: Với một tập hành động khả thi $\mathcal{A}$, **Counterfactual Explanation Tree (CET)** là cây quyết định $h : \mathcal{X} \to \mathcal{A}$ gán hành động cho mỗi trường hợp $x \in \mathcal{X}$.

Ưu điểm:

- Giải thích mỗi hành động dưới dạng *quy tắc* (minh bạch).
- Gán duy nhất một cặp *(hành động, quy tắc)* cho mỗi trường hợp (nhất quán).

Xây dựng Counterfactual Explanation Tree (CET)

Xây dựng CET từ $X \subset \mathcal{X}$ dựa trên điểm không hợp lệ (Invalidity score)

$$i_\gamma(a \mid x) := c(a \mid x) + \gamma \cdot l(f(x+a), y^*)$$

Trong đó:

- $c(a \mid x)$: **Chi phí** khi thực hiện hành động a.
- $l(f(x+a), y^*)$: **Hàm mất mát (Loss)** liên quan đến ràng buộc $f(x+a) = y^*$.
- Hành động $a = h(x)$ được gán là *hiệu quả* nếu có chi phí thấp và đạt được kết quả dự đoán mong muốn.

Thuật toán xây dựng CET

Xây dựng CET từ tập các trường hợp bằng stochastic local search

Bài toán tối ưu CET: Cho tập trường hợp $X \subseteq \mathcal{X}$ và tham số $\gamma, \lambda > 0$, tìm CET h^* sao cho:

$$h^* = \arg\min_{h \in \mathcal{H}} \frac{1}{|X|} \sum_{x \in X} i_\gamma(h(x) \mid x) + \lambda \cdot |\mathcal{L}(h)|$$

- $\mathcal{H}$: tập tất cả các CET $h : \mathcal{X} \to \mathcal{A}$.
- $\mathcal{L}(h)$: tập lá của cây h ($|\mathcal{L}(h)|$ = số hành động được gán).
- Điều chỉnh *trade-off* giữa **hiệu quả hành động** và **tính diễn giải**.

Định lý 1: Giới hạn số lá của CET

Kích thước cây tối ưu bị chặn trên bởi tham số γ, λ

$$|\mathcal{L}(h^*)| \leq \frac{\gamma + \lambda}{\lambda}$$

Ý nghĩa:

- Số lá của CET tối ưu h^* luôn bị chặn trên bởi $\frac{\gamma + \lambda}{\lambda}$.
- Cho thấy sự phụ thuộc trực tiếp giữa độ phức tạp của cây và tham số trade-off.

Thuật toán: Stochastic Local Search

Ý tưởng: Cập nhật các quy tắc phân nhánh trong cây CET hiện tại $h^{(t)}$ bằng một số thao tác ngẫu nhiên.
Các thao tác chỉnh sửa:

- *insert*: chèn thêm một nút phân nhánh
- *delete*: xoá một nút phân nhánh
- *replace*: thay thế quy tắc tại một nút

Bước con (subroutine): Hành động gán cho các lá được tối ưu bằng cách giải MILO mở rộng (cf. (Ustun **andothers**, 2019), (Kanamori **andothers**, 2022)).

Tham khảo: (Wang, 2019), (Pan **andothers**, 2020)

Thực nghiệm (IBM Attrition dataset)

CET có thể gán hành động hiệu quả theo cách dễ diễn giải

So sánh với AReS (Rawal and Lakkaraju, 2020) dựa trên **rule set**:

- **Định lượng:** hiệu quả của hành động (cost, loss, invalidity).
- **Định tính:** mức độ dễ diễn giải với người dùng (user study).

Kết quả:

- CET gán hành động **hiệu quả hơn** AReS về chi phí, mất mát và invalidity, đồng thời vẫn đảm bảo minh bạch và nhất quán.
- Hành vi của CET dễ dàng được người dùng hiểu.
- **CET đạt được hành động hiệu quả và có thể diễn giải!**

Nội dung

Kết quả thực nghiệm chi tiết

So sánh Cost, Loss, Invalidity:

Dataset	Method	Cost	Loss	Invalidity
Train	AReS	0.436 ± 0.06	0.435 ± 0.07	0.871 ± 0.04
	CET	$\mathbf{0.349 \pm 0.1}$	$\mathbf{0.4 \pm 0.11}$	$\mathbf{0.749 \pm 0.05}$
Test	AReS	0.45 ± 0.08	0.298 ± 0.09	0.748 ± 0.09
	CET	$\mathbf{0.383 \pm 0.12}$	$\mathbf{0.318 \pm 0.09}$	$\mathbf{0.701 \pm 0.12}$

So sánh User Accuracy và Thời gian:

Method	User Acc.	Time [s]
AReS	95.12%	784.8 ± 202
CET	**100.0%**	$\mathbf{674.0 \pm 392}$

Kết luận: Kết quả thực nghiệm khác với công bố gốc, chủ yếu do nhóm giảm bớt các tham số và hạn chế về cấu hình máy. Tuy nhiên, kết quả vẫn phản ánh đúng xu hướng phương pháp và cần được kiểm chứng thêm với điều kiện tối ưu hơn.

Thực nghiệm (IBM Attrition dataset) – AReS (**Rawal2020**)

Rule	**Action**
If OverTime=True AND Performance=False	OverTime=False
If BusinessTravel=1 AND OverTime=False	BusinessTravel<1 AND OverTime=False
If JobLevel<2 AND MonthlyIncome<2275	MonthlyIncome≥15170 AND OverTime=False
If OverTime=True AND $2 \leq$ YearsInCurrentRole < 3	MonthlyIncome≥15170 AND OverTime=False
Default	MonthlyIncome≥15170 AND OverTime=False

Cost: 47.0%
Loss: 8.7%
User Acc.: 95.1%
Time: 784.8 sec.

Thực nghiệm (IBM Attrition dataset) – CET (Ours)

Rule	**Action**
If OverTime=True	OverTime=False
If YearsInCurrentRole < 1	BusinessTravel=-1, MonthlyIncome+=8502
If OverTime=False AND YearsInCurrentRole$\geq$1	MonthlyIncome+=2276, OverTime=False
If OverTime=True AND YearsInCurrentRole$\geq$1	BusinessTravel=-1, MonthlyIncome+=2231
Else	OverTime=False, PercentSalaryHike+=1

Cost: 41.0%
Loss: 4.3%
User Acc.: 100%
Time: 674.0 sec.

Nội dung

Phát hiện & Đóng góp

- Giới thiệu CET – cây quyết định gán hành động phản thực.
- Minh bạch, nhất quán cho từng cá thể.
- Học bằng Stochastic Local Search + pruning.
- Hiệu quả, dễ hiểu qua thử nghiệm và khảo sát.

Ưu điểm & Hạn chế

Ưu điểm:

- Minh bạch, nhất quán.
- Phủ toàn cục, dễ hiểu.

Hạn chế:

- Khó mô tả quan hệ phi tuyến phức tạp.
- Tốn chi phí tính toán khi dữ liệu lớn.

Ý nghĩa & Ứng dụng

- Hướng mới cho CE: minh bạch + toàn cục.
- Hữu ích trong tài chính, y tế, pháp lý.
- Thu hẹp khoảng cách: hiệu quả vs giải thích.

Tài liệu tham khảo I

[1] S. Wachter, B. Mittelstadt **and** C. Russell, Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR, 2018. arXiv: 1711.00399 [cs.AI]. **url**: https://arxiv.org/abs/1711.00399.

[2] B. Ustun, A. Spangher **and** Y. Liu, "Actionable Recourse in Linear Classification," **in**Proceedings of the Conference on Fairness, Accountability, and Transparency **jourser** FAT* '19, ACM, **january** 2019, **pages** 10–19. DOI: 10.1145/3287560.3287566. **url**: http://dx.doi.org/10.1145/3287560.3287566.

Tài liệu tham khảo II

[3] A.-H. Karimi, G. Barthe, B. Schölkopf **and** I. Valera, A survey of algorithmic recourse: definitions, formulations, solutions, and prospects, 2021. arXiv: 2010.04050 [cs.LG]. **url**: https://arxiv.org/abs/2010.04050.

[4] R. K. Mothilal, A. Sharma **and** C. Tan, "Explaining machine learning classifiers through diverse counterfactual explanations," **in**Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency **jourser** FAT* '20, ACM, **january** 2020, **pages** 607–617. DOI: 10.1145/3351095.3372850. **url**: http://dx.doi.org/10.1145/3351095.3372850.

Tài liệu tham khảo III

[5] K. Rawal **and** H. Lakkaraju, Beyond Individualized Recourse: Interpretable and Interactive Summaries of Actionable Recourses, 2020. arXiv: 2009.07165 [cs.LG]. **url**: https://arxiv.org/abs/2009.07165.

[6] K. N. Ramamurthy, B. Vinzamuri, Y. Zhang **and** A. Dhurandhar, Model Agnostic Multilevel Explanations, 2020. arXiv: 2003.06005 [cs.LG]. **url**: https://arxiv.org/abs/2003.06005.

Tài liệu tham khảo IV

[7] J. Gao, X. Wang, Y. Wang, Y. Yan **and** X. Xie, "Learning Groupwise Explanations for Black-Box Models," **in**Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence International Joint Conferences on Artificial Intelligence Organization, **august** 2021_08, **pages** 2396–2402. DOI: 10.24963/ijcai.2021/330. **url**: http://dx.doi.org/10.24963/ijcai.2021/330.

[8] C. Rudin **and** Y. Shaposhnik, "Globally-Consistent Rule-Based Summary-Explanations for Machine Learning Models: Application to Credit-Risk Evaluation," Journal of Machine Learning Research, **jourvol** 24, **number** 16, **pages** 1–44, 2023. **url**: http://jmlr.org/papers/v24/21-0488.html.

Tài liệu tham khảo V

[9] K. Kanamori, T. Takagi, K. Kobayashi **and** Y. Ike, "Counterfactual Explanation Trees: Transparent and Consistent Actionable Recourse with Decision Trees," **in**Proceedings of the 25th International Conference on Artificial Intelligence and Statistics 2022, **pages** 1846–1870.

[10] T. Wang, Gaining Free or Low-Cost Transparency with Interpretable Partial Substitute, 2019. arXiv: 1802.04346 [cs.LG]. **url**: https://arxiv.org/abs/1802.04346.

[11] D. Pan, T. Wang **and** S. Hara, Interpretable Companions for Black-Box Models, 2020. arXiv: 2002.03494 [stat.ML]. **url**: https://arxiv.org/abs/2002.03494.