TẠP CHÍ KHOA HỌC
TRƯỜNG ĐẠI HỌC SƯ PHẠM TP HỒ CHÍ MINH
Tập 17, Số 12 (2020): 2251-2261
ISSN: 1859-3100
HO CHI MINH CITY UNIVERSITY OF EDUCATION
JOURNAL OF SCIENCE
Vol. 17, No. 12 (2020): 2251-2261
Website: *http://journal.hcmue.edu.vn*

Bài báo nghiên cứu

XÂY DỰNG CÂY HỒI QUY ĐẢM BẢO TÍNH RIÊNG TƯ CHO TẬP DỮ LIỆU HUẤN LUYỆN BẰNG RIÊNG TƯ SAI BIỆT

Vũ Quốc Hoàng*, Nguyễn Đình Thúc

Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Thành phố Hồ Chí Minh, Việt Nam

**Tác giả liên hệ: Vũ Quốc Hoàng – Email: vqhoang@fit.hcmus.edu.vn*

Ngày nhận bài: 28-4-2020; ngày nhận bài sửa: 22-5-2020, ngày chấp nhận đăng: 28-12-2020

TÓM TẮT

Mô hình hóa dữ liệu là bài toán quan trọng trong phân tích dữ liệu cũng như trong học máy. Có nhiều phương pháp giải quyết bài toán mô hình hóa này, trong đó, cây hồi quy là phương pháp có nhiều ưu điểm so với các phương pháp hồi quy khác. Bên cạnh độ chính xác, khả năng giải thích của mô hình kết quả thì vấn đề đảm bảo tính riêng tư cho tập dữ liệu huấn luyện cũng rất quan trọng và đặt ra cấp thiết, đặc biệt với các dữ liệu cá nhân, nhạy cảm. Bài báo này đề xuất các phương pháp và thuật toán cơ bản để xây dựng cây hồi quy đảm bảo tính riêng tư dựa trên các kĩ thuật riêng tư sai biệt. Kết quả thử nghiệm cho thấy tính khả thi đồng thời cũng mở ra những thách thức cần tiếp tục nghiên cứu, cải tiến.

Từ khóa: riêng tư sai biệt; phân tích dữ liệu đảm bảo tính riêng tư; hồi quy; cây hồi quy

1. Giới thiệu

Khai thác dữ liệu, học máy và học sâu đang ngày càng phát triển nhờ nguồn dữ liệu phong phú, khổng lồ. Tuy nhiên, đi kèm với những lợi ích chúng mang lại là vấn đề riêng tư của dữ liệu, đặc biệt là các dữ liệu có tính cá nhân, nhạy cảm như dữ liệu tài chính, y tế, sinh học... Ngoài hai mục tiêu quan trọng là độ chính xác và tính tự giải thích, thì các mô hình, thuật toán phân tích dữ liệu cũng cần phải chú ý đến tính riêng tư của dữ liệu, là tính chất đặc biệt quan trọng khi các bộ luật về bảo vệ dữ liệu cá nhân của nhiều nước được áp dụng.

Có nhiều kĩ thuật hỗ trợ việc bảo vệ tính riêng tư cho dữ liệu được phân tích. Trong đó, riêng tư sai biệt là kĩ thuật đảm bảo tính riêng tư có thể chứng minh được về mặt toán học. Kĩ thuật này cũng rất tổng quát, áp dụng được cho mọi dạng dữ liệu và thuật toán phân tích mà không phụ thuộc vào thông tin thêm về dữ liệu của người tấn công. Nó cũng lượng hóa được mất mát riêng tư qua các tham số.

Cây quyết định, với nhiều ưu điểm, đã được dùng từ rất sớm trong khai thác dữ liệu với 2 bài toán chính là phân lớp và hồi quy. Mặc dù đã có các phương pháp được đề xuất để xây dựng cây phân lớp hỗ trợ riêng tư sai biệt nhưng lại chưa có phương pháp tương tự trên

Cite this article as: Vu Quoc Hoang, & Nguyen Dinh Thuc (2020). Differentially private regression tree and forest. *Ho Chi Minh City University of Education Journal of Science, 17*(12), 2251-2261.

cây hồi quy. Bài báo này đề xuất các phương pháp đơn giản làm cơ sở ban đầu cho việc xây dựng cây hồi quy hỗ trợ riêng tư sai biệt. Các phương pháp chúng tôi đưa ra dựa trên các thuật toán và phương pháp riêng tư sai biệt và các thuật toán tương tự trên cây phân lớp.

2. Cơ sở lí thuyết

2.1. Riêng tư sai biệt

Phần này nêu lại các định nghĩa và định lí quan trọng của riêng tư sai biệt sẽ được dùng cho các phần sau (chi tiết trong Dwork và Roth (2014)). Ta nói hai tập dữ liệu $x, y \in D^n$ là lân cận nếu chúng khác nhau không quá 1 điểm dữ liệu. Kí hiệu D^n ở đây chỉ tập tất cả các tập dữ liệu gồm n điểm dữ liệu, mỗi điểm là một phần tử của D. Ý tưởng cơ bản của riêng tư sai biệt là dựa vào ngẫu nhiên với yêu cầu khó phân biệt cho các phân phối xác suất của kết quả truy vấn trên các tập dữ liệu lân cận.

Định nghĩa 1. (Riêng tư sai biệt, Dwork et al., 2006)

Thuật toán ngẫu nhiên $M: D^n \to R$ được gọi là thỏa riêng tư sai biệt ε ($\varepsilon \geq 0$) nếu với mọi tập dữ liệu lân cận $x, y \in D^n$ và mọi $S \subset R$, ta có:

$$\Pr[M(x) \in S] \leq \exp(\varepsilon)\Pr[M(y) \in S] \quad (1)$$

Định nghĩa 2. (Độ nhạy toàn cục, Dwork et al. (2006))

Độ nhạy của hàm $f: D^n \to \mathbb{R}$ được định nghĩa là:

$$\Delta f = \max_{x,y \in D^n \text{ lân cận}} |f(x) - f(y)| \quad (2)$$

Định lí 1. (Hậu xử lí, Dwork et al., 2006)

Nếu $M: D^n \to R$ riêng tư sai biệt ε và $f: R \to R'$ thì $f \circ M: D^n \to R'$ cũng riêng tư sai biệt ε.

Định lí 2. (Kết hợp tuần tự, McSherry & Talwar, 2007)

Nếu $M_i: D^n \to R_i$ riêng tư sai biệt ε_i $(1 \leq i \leq k)$ thì thuật toán dùng các M_i trên cùng tập dữ liệu $x \in D^n$ thỏa riêng tư sai biệt $\sum_{i=1}^{k} \varepsilon_i$.

Định lí 3. (Kết hợp song song, McSherry, 2009)

Nếu $M_i: D^n \to R_i$ riêng tư sai biệt ε_i $(1 \leq i \leq k)$ thì thuật toán dùng các M_i trên các tập con rời nhau của tập dữ liệu $x \in D^n$ thỏa riêng tư sai biệt $\max_{1 \leq i \leq k} \varepsilon_i$.

Định lí 4. (Cơ chế Laplace, Dwork et al., 2006)

Cho hàm $f: D^n \to \mathbb{R}$ có độ nhạy Δf, cơ chế $M: D^n \to \mathbb{R}$ được định nghĩa như sau thỏa riêng tư sai biệt ε:

$$M(x) = f(x) + L, x \in D^n \quad (3)$$

trong đó, $L \sim Lap\left(\frac{\Delta f}{\varepsilon}\right)$ là biến ngẫu nhiên có phân phối Laplace với kì vọng 0 và tỉ lệ $\frac{\Delta f}{\varepsilon}$.

Định lí 5. (Cơ chế mũ, McSherry & Talwar, 2007)

Cho hàm $u: D^n \times R \to \mathbb{R}$ có độ nhạy Δu, cơ chế $M: D^n \to R$ được định nghĩa như sau thỏa riêng tư sai biệt ε:

$$\Pr[M(x) = r] \propto \exp\left(\frac{\varepsilon u(x,r)}{2\Delta u}\right), x \in D^n, r \in R \quad (4)$$

trong đó, hàm u thường được gọi là hàm tiện ích và $\Delta u = \max_{r \in R} \max_{x,y \in D^n \text{ lân cận}} |u(x,r) - u(y,r)|$.

2.2. Hồi quy và cây hồi quy

Cho $D = \{(x_i, y_i)\}_{i=1}^n$ là tập dữ liệu huấn luyện với $x_i \in A = A_1 \times A_2 \times ... \times A_k$ và $y_i \in Y = \mathbb{R}$. Các $A_1, A_2, ..., A_k$ được gọi là thuộc tính và Y được gọi là mục tiêu. Hồi quy là bài toán từ tập dữ liệu huấn luyện D, xây dựng quan hệ giữa mục tiêu Y với các thuộc tính A. Quan hệ này có thể được dùng để giải thích hoặc dự đoán giá trị mục tiêu của điểm dữ liệu mới khi biết các giá trị thuộc tính.

Có nhiều mô hình và thuật toán hồi quy khác nhau, trong đó, mô hình hồi quy bằng cây quyết định có nhiều ưu điểm như: có tính giải thích cao, dễ hiểu với người phân tích, chi phí tính toán thấp, phi tham số, mô tả được quan hệ phi tuyến, dùng được cho các thuộc tính rời rạc lẫn liên tục… Nhược điểm chính của cây hồi quy là nhạy cảm với dữ liệu và dễ xảy ra quá khớp.

Thuật toán xây dựng cây cơ bản là quá trình lặp lại các bước sau xuất phát từ tập dữ liệu huấn luyện:

- Kiểm tra điều kiện dừng, nếu thỏa thì tạo nút lá với giá trị tương ứng;
- Nếu không, chọn thuộc tính và giá trị chia nhánh tốt nhất cho tập dữ liệu hiện tại;
- Phân hoạch tập dữ liệu thành các nhóm theo giá trị của thuộc tính đã chọn, tạo nút nội với các nhánh đệ quy cho từng nhóm dữ liệu đã chia.

Hai tiêu chí đánh giá thường dùng để chọn thuộc tính và giá trị tốt nhất cho nút nội là: trung bình bình phương lỗi và trung bình trị tuyệt đối lỗi với giá trị tương ứng ở nút lá là trung bình và trung vị các giá trị mục tiêu của các điểm dữ liệu ở nút đó (chi tiết trong Han et al. (2012), Breiman et al. (2017)).

3. Đối tượng và Phương pháp nghiên cứu

3.1. Cây phân lớp thỏa riêng tư sai biệt

Cây quyết định cũng có thể được dùng làm mô hình phân lớp mà khi đó thường được gọi là cây phân lớp. Cụ thể, từ tập dữ liệu huấn luyện $D = \{(x_i, y_i)\}_{i=1}^n$ với $y_i \in Y$ là tập nhãn lớp, cây quyết định được xây dựng bằng thuật toán tham lam tương tự trên với giá trị của nút lá là nhãn của lớp chứa nhiều điểm dữ liệu nhất trong nút. Các tiêu chí đánh giá hay dùng để chọn thuộc tính và giá trị tốt nhất cho nút nội là độ lợi thông tin, tỉ suất lợi, chỉ số Gini (Han et al., 2012; Breiman et al., 2017).

Nhiều thuật toán xây dựng cây phân lớp thỏa riêng tư sai biệt đã được đề xuất với mục tiêu vừa cho kết quả dự đoán tốt vừa đảm bảo tính riêng tư cho tập dữ liệu huấn luyện

(Fletcher, 2016). Các yếu tố chính cần xem xét khi xây dựng cây thỏa riêng tư sai biệt là (Fletcher, 2016):

- Tối thiểu số lần truy cập dữ liệu,
- Dùng các truy vấn có độ nhạy thấp,
- Cấp phát quỹ riêng tư một cách hợp lí.

Sớm nhất (Blum et al., 2005) dùng cơ chế Laplace (Định lí 4) cho các truy vấn đếm để kiểm tra điều kiện dừng và tính độ lợi thông tin. Sau đó, Friedman và Schuster (2010) dùng cơ chế mũ (Định lí 5) để chọn thuộc tính và giá trị chia nhánh tốt nhất. Cải tiến này giúp tiết kiệm quỹ riêng tư nhờ truy vấn trên các tập dữ liệu rời nhau (Định lí 3). Friedman và Schuster (2010) cũng thử nghiệm các tiêu chí phân nhánh với độ nhạy khác nhau.

Jagannathan và cộng sự (2012), mở đầu hướng nghiên cứu dùng kết hợp nhiều cây thỏa riêng tư sai biệt bằng cách chia đều quỹ riêng tư cho mỗi cây (Định lí 2). Các công trình sau đó của Patil và Singh (2014), Rana và cộng sự (2015), Fletcher và Islam (2015), Fletcher và Islam (2017), tiếp tục hướng này. Gần đây, Xin và cộng sự. (2019) dùng nhiều cây riêng tư xây dựng trên các tập con rời nhau của tập dữ liệu huấn luyện đã cho kết quả rất tốt nhờ Định lí 3.

3.2. Cây hồi quy tham lam thỏa riêng tư sai biệt

3.2.1. Thuật toán

Mặc dù có nhiều công trình nghiên cứu việc xây dựng cây phân lớp thỏa riêng tư sai biệt nhưng chưa có công trình nào như vậy cho cây hồi quy. Bảng 1 trình bày khung thuật toán để xây dựng cây hồi quy tham lam thỏa riêng tư sai biệt do chúng tôi đề xuất.

Bảng 1. *Thuật toán xây dựng cây hồi quy tham lam hỗ trợ riêng tư sai biệt*

1.	**procedure** DPGreedyRegTree($X, Y, A, l_{max}, N_{split}, N_{leaf}, \varepsilon$)
2.	**Input**: X, Y – tập dữ liệu huấn luyện, A – tập thuộc tính, l_{max} – mức tối đa của cây, N_{split} – số điểm dữ liệu tối thiểu để chia nút, N_{leaf} – số điểm dữ liệu tối thiểu của nút lá, ε – tham số riêng tư.
3.	$\beta = \frac{\varepsilon}{2l_{max}+1}$
4.	**return** cây hồi quy có nút gốc là Build_Tree($X, Y, A, l_{max}, N_{split}, N_{leaf}, \beta$)
5.	**end procedure**
6.	**procedure** Build_Tree($X, Y, A, l, N_{split}, N_{leaf}, \varepsilon$)
7.	$N = \text{LapMech}(\lvert X\rvert, \varepsilon)$
8.	**if** $l = 0$ **or** $N < N_{split}$ **then**
9.	**return** nút lá với giá trị $\text{LapMech}(\bar{Y}, \varepsilon)$
10.	$A^*, v^* = \text{ExpMech}(X, Y, A, \varepsilon)$
11.	Chia (X, Y) ra 2 phần $(X_l, Y_l), (X_r, Y_r)$ theo giá trị v^* của thuộc tính A^*
12.	$N_l, N_r = \text{LapMech}(\lvert X_l\rvert, \varepsilon), \text{LapMech}(\lvert X_r\rvert, \varepsilon)$
13.	**if** $N_l < N_{leaf}$ **or** $N_r < N_{leaf}$ **then**

14. **return** nút lá với giá trị LapMech($\bar{Y}, \varepsilon$)
15. **return** nút nội với nhãn (A^*, v^*) và 2 cây con trái, phải tương ứng là Build_Tree($X_l, Y_l, A, l-1, N_{split}, N_{leaf}, \varepsilon$) và Build_Tree($X_r, Y_r, A, l-1, N_{split}, N_{leaf}, \varepsilon$)
16. **end procedure**

Tương tự như Friedman và Schuster (2010), chúng tôi dùng cơ chế cấp phát quỹ riêng tư đều với tham số riêng tư ε được chia đều cho các mức của cây theo Định lí 2. Ở mỗi mức, do các nút truy cập đến các tập rời nhau của dữ liệu huấn luyện nên theo Định lí 3, quỹ riêng tư không cần chia cho các nút khác nhau. Trong thủ tục DPGreedyRegTree, tham số l_{max} xác định mức tối đa của cây, và vì ở mỗi nút có 2 truy vấn (trừ mức l_{max} chỉ có 1 truy vấn) nên ε được chia cho các nút như ở Dòng 3.

3.2.2. Các truy vấn hỗ trợ riêng tư sai biệt

Để đếm số lượng điểm dữ liệu trong nút, chúng tôi dùng cơ chế Laplace (Định lí 4) với độ nhạy của truy vấn là $\Delta f = 1$ (Blum et al., 2005). Cụ thể:

$$\text{LapMech}(|X|, \varepsilon) = |X| + Lap\left(\frac{1}{\varepsilon}\right) \quad (5)$$

Lựa chọn thường dùng cho giá trị của nút lá là trung bình các giá trị mục tiêu của các điểm dữ liệu trong nút. Để hỗ trợ riêng tư sai biệt, phạm vi của mục tiêu phải bị chặn, ở đây, chúng tôi giả sử $Y = [0, 1]$. Giả sử này hợp lí vì giá trị mục tiêu thường được chuẩn hóa về khoảng này trong thực tế. Khi đó, chúng tôi dùng cơ chế Laplace cho truy vấn trung bình với độ nhạy của truy vấn là $\Delta f = \frac{1}{n}$ (Dwork et al., 2006). Vì số lượng điểm dữ liệu tối thiểu trong nút lá là N_{leaf} nên số lượng điểm dữ liệu trong nút $n \geq N_{leaf}$ nên $\Delta f \leq \frac{1}{N_{leaf}}$. Lưu ý, các thông số $l_{max}, N_{split}, N_{leaf}$ được chọn trước và không phụ thuộc vào tập dữ liệu. Tóm lại, giá trị cho nút lá là:

$$\text{LapMech}(\bar{Y}, \varepsilon) = \bar{Y} + Lap\left(\frac{1}{N_{leaf}\varepsilon}\right) \quad (6)$$

Lựa chọn khác cho giá trị của nút lá là trung vị các giá trị mục tiêu của các điểm dữ liệu trong nút. Để hỗ trợ riêng tư sai biệt, chúng tôi dùng cơ chế mũ như trong Sarwate và Chaudhuri (2013). Cụ thể, sắp xếp các giá trị mục tiêu trong nút tăng dần theo các khoảng $y_0 = 0, y_1, \dots, y_n, y_{n+1} = 1$, đặt $F_n(y) = \frac{\#\{y_i \leq y\}}{n}$ là hàm phân phối tích lũy thực nghiệm của các giá trị mục tiêu, chọn hàm tiện ích $u(y) = -|0.5 - F_n(y)|$ thì $\Delta u = \frac{1}{n} \leq \frac{1}{N_{leaf}}$. Giá trị cho nút lá là:

$$\text{ExpMech}(\text{median}(Y), \varepsilon) = Uniform(y_{K-1}, y_K) \quad (7)$$

với:

$$\Pr[K = k] \propto |y_k - y_{k-1}| \exp\left(\frac{-\varepsilon|0.5 - F_n(y_k)|N_{leaf}}{2}\right), k \in \{1, 2, \dots, n+1\} \quad (8)$$

Để chọn thuộc tính và giá trị chia nhánh tốt nhất cho tập dữ liệu tại nút nội, chúng tôi dùng cơ chế mũ với hàm tiện ích là đối của trung bình bình phương lỗi hay đối của trung bình trị tuyệt đối lỗi, tương ứng với lựa chọn giá trị cho nút lá là trung bình hay trung vị. Độ nhạy của các truy vấn này đều bị chặn bởi $\Delta f \leq \frac{1}{N_{split}}$ do số lượng điểm dữ liệu tối thiểu để chia nhánh là N_{split}.

Friedman và Schuster (2010), có đề xuất cách xử lí thuộc tính liên tục. Tuy nhiên, cách này không hiệu quả khi số lượng điểm dữ liệu trong nút nhiều nên chúng tôi dùng cách đơn giản hơn là rời rạc hóa bằng N_{range} điểm đại diện phân cách đều trong khoảng $[0, 1]$. Dĩ nhiên, các thuộc tính liên tục cũng phải được chuẩn hóa về khoảng $[0, 1]$ trước đó. Lưu ý, các điểm đại diện này không phụ thuộc vào dữ liệu.

3.3. Rừng hồi quy phân vùng thỏa riêng tư sai biệt

Tương tự Xin và cộng sự (2019), chúng tôi cũng đề xuất việc dùng nhiều cây hồi quy riêng tư xây dựng trên các tập con rời nhau của tập dữ liệu huấn luyện như trong thuật toán ở Bảng 2.

Bảng 2. *Thuật toán xây dựng rừng hồi quy phân vùng hỗ trợ riêng tư sai biệt*

1.	**procedure** DPPartRegForest($\tau, X, Y, A, l_{max}, N_{split}, N_{leaf}, \varepsilon$)
2.	**Input**: τ – số cây, X, Y – tập dữ liệu huấn luyện, A – tập thuộc tính, l_{max} – mức tối đa của cây, N_{split} – số điểm dữ liệu tối thiểu để chia nút, N_{leaf} – số điểm dữ liệu tối thiểu của nút lá, ε – tham số riêng tư.
3.	Chia tập dữ liệu (X, Y) ra làm τ vùng rời nhau $(X_1, Y_1), (X_2, Y_2), \dots, (X_\tau, Y_\tau)$
4.	**return** rừng gồm τ cây là kết quả của DPGreedyRegTree($X_i, Y_i, A, l_{max}, N_{split}, N_{leaf}, \varepsilon$) với $i = 1, \dots, \tau$
5.	**end procedure**

Số điểm dữ liệu trong mỗi phân vùng được chọn xấp xỉ nhau. Vì mỗi cây hồi quy được xây dựng trên các tập dữ liệu rời nhau nên quỹ riêng tư vẫn giữ cho mỗi cây theo Định lí 3. Khi dự đoán, giá trị trung bình của kết quả dự đoán của các cây được dùng mà không phải dùng thêm cơ chế riêng tư nào nhờ Định lí 1.

4. Kết quả và thảo luận

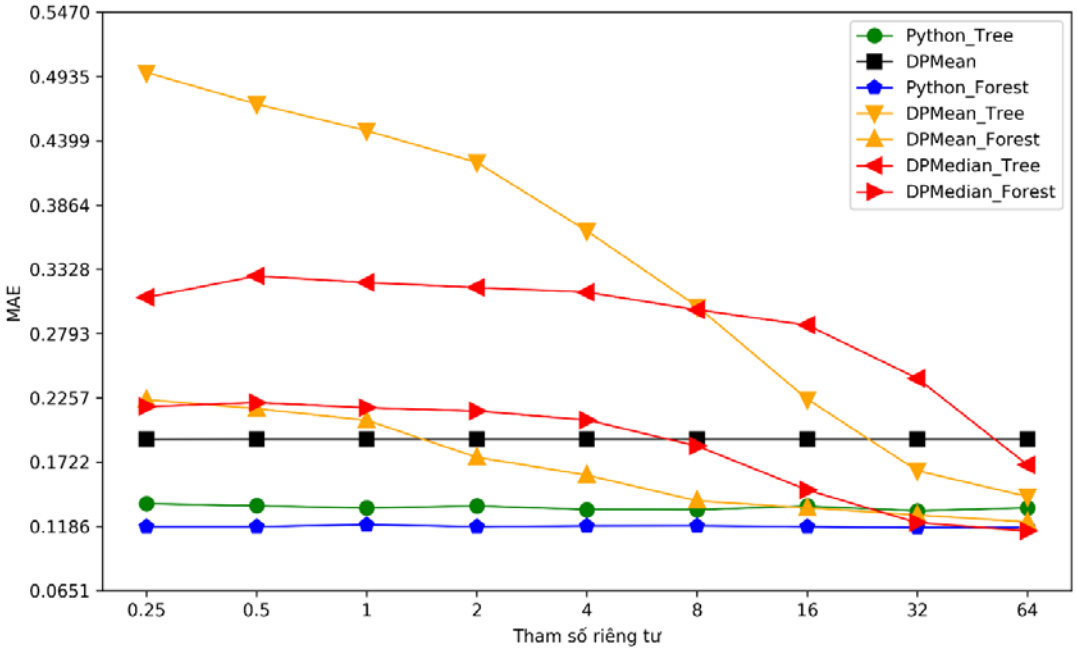
Để đánh giá các phương pháp, chúng tôi dùng tập dữ liệu huấn luyện California Housing (Pace, & Barry, 1997) gồm 20.640 điểm dữ liệu với 9 thuộc tính liên tục (kể cả mục tiêu). Tập dữ liệu được tiền xử lí để đưa giá trị các thuộc tính liên tục và mục tiêu về khoảng [0,1]. Mỗi phương pháp được đánh giá bằng kĩ thuật 10-fold Cross Validation dùng trung bình sai số tuyệt đối (MAE).

4.1. Đánh giá MAE theo tham số riêng tư

Vì chưa có công trình nào dùng riêng tư sai biệt trên cây hồi quy nên chúng tôi dùng 3 phương pháp sau làm cơ sở đánh giá: cây hồi quy trong thư viện Python scikit-learn (Pedregosa et al., 2011), giá trị trung bình của mục tiêu của tất cả các điểm dữ liệu hỗ trợ

riêng tư sai biệt theo cơ chế Laplace (Định lí 4) và rừng hồi quy dùng các cây hồi quy Python scikit-learn. Các phương pháp này được đặt tên lần lượt là Python_Tree (1), DPMean (2) và Python_Forest (3).

Các phương pháp chúng tôi xây dựng được đánh giá gồm phương pháp một cây hồi quy tham lam và phương pháp rừng hồi quy phân vùng theo các tiêu chí chia nhánh nút nội và chọn giá trị tương ứng của nút lá là trung bình và trung vị. Các phương pháp này được đặt tên lần lượt là DPMean_Tree (4), DPMean_Forest (5), DPMedian_Tree (6), DPMedian_Forest (7) và được đánh giá qua các tham số riêng tư ε là 0.25, 0.5, 1, 2, 4, 8, 16, 32, 64. Kết quả đánh giá được trình bày ở Hình 1 và chi tiết ở Bảng 3 với độ lệch chuẩn ghi trong ngoặc. Các giá trị chọn cho thông số là $N_{range} = 40, N_{split} = 20, N_{leaf} = 10$, với một cây tham lam thì $l_{max} = 15$ và với rừng cây thì $l_{max} = 5$ và số cây trong rừng là $\tau = 25$.



Hình 1. *Kết quả đánh giá MAE theo tham số riêng tư của các mô hình cây hồi quy*

Bảng 3. *Kết quả chi tiết MAE theo tham số riêng tư của các mô hình cây hồi quy*

ε	(1)	(2)	(3)	(4)	(5)	(6)	(7)
0.25	0.1378 (0.031)	0.1915 (0.0309)	0.1185 (0.0188)	0.497 (0.0525)	0.2244 (0.0465)	0.3097 (0.0426)	0.2186 (0.0423)
0.5	0.1361 (0.0273)	0.1916 (0.0309)	0.1185 (0.0188)	0.4708 (0.0537)	0.2169 (0.0364)	0.3275 (0.0405)	0.2219 (0.0369)
1	0.1342 (0.0265)	0.1916 (0.0309)	0.1202 (0.0185)	0.4485 (0.0341)	0.2073 (0.0384)	0.322 (0.0329)	0.2177 (0.0235)
2	0.136 (0.0275)	0.1916 (0.0309)	0.1185 (0.0183)	0.4222 (0.0193)	0.1764 (0.0229)	0.3179 (0.0228)	0.215 (0.0308)
4	0.1329 (0.0252)	0.1916 (0.0309)	0.1191 (0.0178)	0.365 (0.0285)	0.1615 (0.0292)	0.314 (0.0306)	0.2075 (0.0288)

8	0.1327 (0.0264)	0.1916 (0.0309)	0.1192 (0.0189)	0.3022 (0.0233)	0.1402 (0.0223)	0.2993 (0.0184)	0.1858 (0.0286)
16	0.1358 (0.031)	0.1916 (0.0309)	0.1185 (0.0195)	0.2249 (0.0191)	0.1343 (0.025)	0.2866 (0.0213)	0.1492 (0.0299)
32	0.1317 (0.0246)	0.1916 (0.0309)	0.1178 (0.0192)	0.1652 (0.0118)	0.1284 (0.0233)	0.2422 (0.0242)	0.1219 (0.0348)
64	0.1343 (0.0262)	0.1916 (0.0309)	0.1178 (0.0188)	0.1437 (0.0225)	0.1226 (0.0218)	0.1701 (0.0205)	0.1151 (0.031)

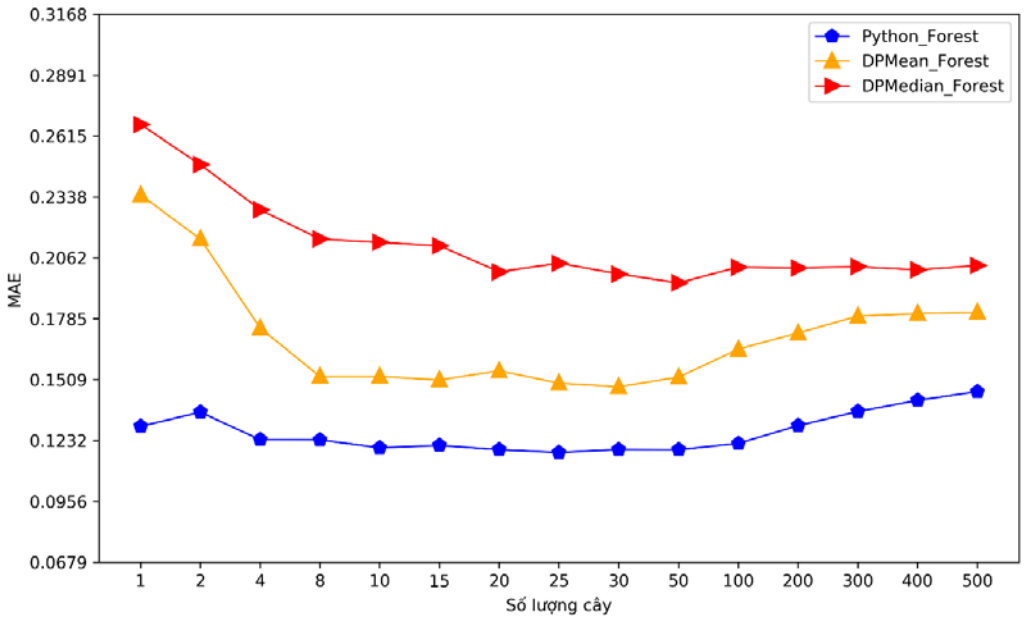
Từ kết quả Bảng 3, chúng tôi nhận thấy, việc dùng cây hồi quy riêng tư sai biệt cho tham số riêng tư $\varepsilon \leq 1$ là không có ý nghĩa vì phương pháp đơn giản DPMean cho kết quả tốt hơn. Cũng lưu ý, DPMean cho kết quả tốt vì số điểm dữ liệu trong tập dữ liệu là lớn ($n = 20640$) nên độ nhạy của truy vấn trung bình $\frac{1}{n}$ là rất nhỏ. Trường hợp tập dữ liệu nhỏ (có ít điểm dữ liệu) thì phương pháp đơn giản như DPMean sẽ cho kết quả không tốt. Phương pháp dùng rừng cây cho kết quả tốt hơn 1 cây vì khắc phục được nhược điểm quá khớp dữ liệu và nhạy cảm của cây tham lam.

Với tham số riêng tư trong khoảng $1 \leq \varepsilon \leq 10$, phương pháp dùng tập cây phân vùng với giá trị trung bình mục tiêu (DPMean_Forest) có kết quả rất tốt, thậm chí gần đạt được kết quả của cây hồi quy thông thường không hỗ trợ riêng tư. Trường hợp $\varepsilon > 10$, DPMean_Forest thậm chí cho kết quả tốt hơn 1 cây tham lam.

Đặc biệt, khi $\varepsilon > 50$ thì tập cây phân vùng có dùng riêng tư sai biệt lại cho kết quả tốt hơn rừng cây thông thường. Điều này có thể được lí giải bởi việc dùng ngẫu nhiên cho cây hồi quy riêng tư giúp hạn chế việc quá khớp dữ liệu huấn luyện do đó giúp tăng khả năng tổng quát hóa cho dữ liệu mới. Dĩ nhiên, khi ε nhỏ thì nhiễu quá lớn nên dẫn đến kết quả không thể tốt như cây hồi quy thông thường.

4.2. Đánh giá MAE theo số lượng cây trong rừng hồi quy

Chúng tôi cũng chạy thực nghiệm để đánh giá MAE theo số lượng cây trong rừng cây phân vùng. Rừng cây phân vùng được đánh giá với các cây được dùng là cây thông thường (Python_Forest), cây riêng tư với giá trị trung bình (DPMean_Forest) và giá trị trung vị (DPMedian_Forest). Số lượng cây τ được đánh giá là 1, 2, 4, 8, 10, 15, 20, 25, 30, 50, 100, 200, 300, 400, 500 với tham số riêng tư là $\varepsilon = 5$. Kết quả đánh giá được trình bày ở Hình 2.



Hình 2. *Kết quả đánh giá MAE theo số lượng cây trong rừng hồi quy*

Trong trường hợp tập dữ liệu này, kết quả cho thấy số cây tối ưu là khoảng 25 cây. Dĩ nhiên, số cây phân vùng tốt phụ thuộc vào kích thước dữ liệu. Tập dữ liệu California Housing gồm 20.640 điểm dữ liệu. Như vậy, số cây nên được chọn để có khoảng 700-1000 điểm dữ liệu cho 1 cây.

5. Kết luận

Kết quả thực nghiệm cho thấy việc xây dựng cây hồi quy hỗ trợ riêng tư sai biệt là khả thi, giúp tạo mô hình hồi quy có tính giải thích cao, khả năng dự đoán tốt mà vẫn đảm bảo tính riêng tư cho tập dữ liệu huấn luyện.

Phương pháp dùng nhiều cây phân vùng cho kết quả rất tốt, nhất là trên tập dữ liệu huấn luyện lớn (có nhiều điểm dữ liệu), vì vừa giúp khắc phục nhược điểm dễ quá khớp dữ liệu và nhạy cảm lại vừa hạn chế việc dùng quỹ riêng tư do được huấn luyện trên các phân vùng dữ liệu rời nhau.

Kết quả chưa tốt cho trường hợp tham số riêng tư nhỏ ($\varepsilon < 1$) cho thấy việc cần phải tiếp tục nghiên cứu, thử nghiệm, cải tiến phương pháp. Các hướng phát triển có thể là:

- Thử nghiệm các chiến lược cấp phát quỹ riêng tư không đều. Chẳng hạn, chia quỹ nhiều hơn cho các nút lá vì chúng mang giá trị dự đoán.
- Thử nghiệm cây ngẫu nhiên vì việc chọn ngẫu nhiên thuộc tính và giá trị chia nhánh không truy cập dữ liệu nên giúp tiết kiệm quỹ riêng tư. Dĩ nhiên, các cây này phải được kết hợp thành rừng cây vì mỗi cây thường không cho kết quả dự đoán tốt.

❖ ***Tuyên bố về quyền lợi**: Các tác giả xác nhận hoàn toàn không có xung đột về quyền lợi*

❖ ***Lời cảm ơn:** Nghiên cứu này được tài trợ bởi Đại học Quốc gia Thành phố Hồ Chí Minh (VNU-HCM) trong dự án NCM2019-18-01.*

TÀI LIỆU THAM KHẢO

Blum, A., Dwork, C., McSherry, F., & Nissim, K. (2005). Practical privacy: The SuLQ framework. *PODS '05*.

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (2017). *Classification and Regression Trees*.

Dwork, C., McSherry, F., Nissim, K., & Smith, A. (2006). Calibrating Noise to Sensitivity in Private Data Analysis. *J. Priv. Confidentiality, 7*, 17-51.

Dwork, C., & Roth, A. (2014). The Algorithmic Foundations of Differential Privacy. *Foundations and Trends in Theoretical Computer Science, 9*, 211-407.

Fletcher, S., & Islam, M. Z. (2015). A Differentially Private Decision Forest. *AusDM*.

Fletcher, S., & Islam, M. Z. (2016). Decision Tree Classification with Differential Privacy: A Survey. *ACM Comput. Surv., 52*, 83:1-83:33.

Fletcher, S., & Islam, M. Z. (2017). Differentially Private Random Decision Forests using Smooth Sensitivity. *ArXiv, abs/1606.03572*.

Friedman, A., & Schuster, A. (2010). Data mining with differential privacy. *KDD '10*.

Han, J., Kamber, M., & Pei, J. (2012). *Data mining concepts and techniques*, third edition Morgan Kaufmann Publishers.

Jagannathan, G., Pillaipakkamnatt, K., & Wright, R. N. (2012). A Practical Differentially Private Random Decision Tree Classifier. *2012 IEEE International Conference on Data Mining Workshops*, 114-121.

McSherry, F., & Talwar, K. (2007). Mechanism Design via Differential Privacy. *48th Annual IEEE Symposium on Foundations of Computer Science (FOCS'07)*, 94-103.

McSherry, F. (2009). Privacy integrated queries: an extensible platform for privacy-preserving data analysis. *SIGMOD Conference.*

Pace, R. K. & Barry, R. (1997). Sparse spatial autoregressions. *Statistics & Probability Letters, 33*, 291-297.

Patil, A., & Singh, S. (2014). Differential private random forest. *2014 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, 2623-2630.

Pedregosa et al. (2011). Scikit-learn: Machine Learning in Python. *JMLR 12*, 2825-2830.

Rana, S., Gupta, S. K., & Venkatesh, S. (2015). Differentially Private Random Forest with High Utility. *2015 IEEE International Conference on Data Mining*, 955-960.

Sarwate, A. D., & Chaudhuri, K. (2013). Signal Processing and Machine Learning with Differential Privacy: Algorithms and Challenges for Continuous Data. *IEEE Signal Processing Magazine, 30*, 86-94.

Xin, B., Yang, W., Wang, S., & Huang, L. (2019). Differentially Private Greedy Decision Forest. *ICASSP 2019 - 2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2672-2676.

Vu Quoc Hoang*, Nguyen Dinh Thuc
University of Science, Vietnam National University Ho Chi Minh City, Vietnam
Corresponding author: Vu Quoc Hoang – Email: vqhoang@fit.hcmus.edu.vn
Received: April 28, 2020; Revised: May 22, 2020; Accepted: December 28, 2020

ABSTRACT

Data modeling is an important problem in data analysis as well as machine learning. There exist many different data modeling solutions, of which regression tree is a method which has many advantages compared to other regression methods. In addition to the accuracy and interpretability of the result model, the issue of ensuring the privacy of the training dataset is also very important and urgent, especially with sensitive and personal data. This paper proposes basic methods and algorithms to build privacy-preserving regression trees based on the differential privacy techniques and algorithms. The experimental results indicate the feasibility of the proposed methods, while also raise challenges which could be further studied.

Keywords: differential privacy; privacy-preserving data analysis; regression; regression tree