TẠP CHÍ Khoa học & Công nghệ
Trường Đại học Công nghiệp Hà Nội
P-ISSN 1859-3585
E-ISSN 2615-9619

Tòa soạn: Tầng 3 - Nhà A1- Đại học Công nghiệp Hà Nội
Địa chỉ: 298 Đường Cầu Diễn, Q. Bắc Từ Liêm, TP. Hà Nội
Điện thoại: 024 37655121 (Số máy lẻ 8251) I Fax: 024 37655261
E-mail: tapchikhcn@haui.edu.vn I Website: http://www.tapchikhcn.haui.edu.vn

NGHIÊN CỨU MỘT SỐ THUẬT TOÁN ADABOOST VÀ ỨNG DỤNG CHO BÀI TOÁN PHÂN LỚP

RESEARCH VARIATIONS OF ADABOOST ALGORITHM AND ITS APPLICATION IN CLASSIFICATION PROBLEMS

Nguyễn Trung Đức[1], Đinh Minh Đại[1]

[1]Khoa Công nghệ thông tin, Trường Đại học Công nghiệp Hà Nội

[*]Email: ducnt.bn.2002@gmail.com

TÓM TẮT

Học kết hợp (Ensemble Learning) là phương pháp huấn luyện nhiều mô hình học máy và kết hợp đầu ra của các mô hình để tạo ra một mô hình phức tạp hơn. Boosting là kỹ thuật thực hiện trong quy trình trên và AdaBoost là thuật toán tiêu biểu cho kỹ thuật Boosting nhằm cải thiện độ chính xác của mô hình dự đoán. Bài báo này tập trung vào việc phân tích và làm rõ một vài biến thể phổ biến của thuật toán AdaBoost cũng như ứng dụng của từng biến thể vào các bài toán cụ thể. Bài báo trình bày các biến thể của AdaBoost như AdaBoost.M1, AdaBoost.MH và AdaBoost.MR cùng với mã giả của từng thuật toán.

Tiếp theo, bài báo ứng dụng các biến thể của AdaBoost trên các bài toán phân lớp với các bộ dữ liệu khác nhau, bao gồm bộ dữ liệu Marketing and Sales, bộ dữ liệu đậu khô, bộ dữ liệu Yelp và bộ dữ liệu Yeast. Kết quả cho thấy các thuật toán AdaBoost đều đạt được kết quả tốt hơn so với các mô hình khác như SVM, Logistic Regression, Decision Tree và Random Forest.

Từ khóa: *Ensemble Learning, AdaBoost, bài toán phân lớp.*

ABSTRACT

Ensemble Learning refers to the processes used to train multiple machine learning models and combine the output of the models to create a more complex model. Boosting is a technique to implement in the above process and AdaBoost is a popular algorithm for the Boosting technique to improve the accuracy of the prediction model. This article will focus on analyzing and clarifying a few common variations of the AdaBoost algorithm as well as the application of each variation to specific problems. The article presents variations of AdaBoost such as AdaBoost.M1, AdaBoost.MH and AdaBoost.MR along with the pseudocode of each algorithm.

Furthermore, the article applies variations of AdaBoost to classification problems with different datasets, including the Marketing and Sales dataset, the dry bean dataset, the Yelp dataset and the Yeast dataset. The results show that AdaBoost algorithms all achieve better results than other models such as SVM, Logistic Regression, Decision Tree and Random Forest.

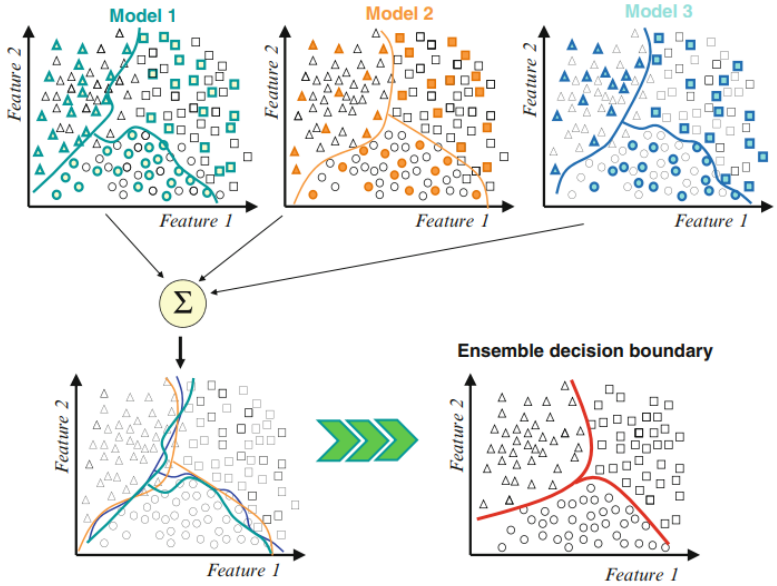
Keywords**:** *Ensemble Learning, AdaBoost, classificaiton problems.*

1. GIỚI THIỆU

Ensemble Learning đề cập đến các quy trình được sử dụng để huấn luyện nhiều mô hình học máy và kết hợp đầu ra của các mô hình đó, coi những mô hình thuộc một nhóm gồm những mô hình đưa ra quyết định. Với nguyên tắc là quyết định của nhóm phải có trung bình độ chính xác cao hơn so với các mô hình riêng lẻ nếu được kết hợp theo một cách thích hợp [1]. Ý tưởng về việc kết hợp, thu thập những ý kiến, quyết định cho một vấn đề xuất hiện rất nhiều ở cuộc

TẠP CHÍ Khoa học & Công nghệ
Trường Đại học Công nghiệp Hà Nội
P-ISSN 1859-3585
E-ISSN 2615-9619
Tòa soạn: Tầng 3 - Nhà A1- Đại học Công nghiệp Hà Nội
Địa chỉ: 298 Đường Cầu Diễn, Q. Bắc Từ Liêm, TP. Hà Nội
Điện thoại: 024 37655121 (Số máy lẻ 8251) I Fax: 024 37655261
E-mail: tapchikhcn@haui.edu.vn I Website: http://www.tapchikhcn.haui.edu.vn

sống hằng ngày, ví dụ như: Bản chất của nền dân chủ nơi một nhóm người bỏ phiếu để đưa ra quyết định, chọn một quan chức được bầu hay quyết định một luật mới, trên thực tế là dựa trên việc ra quyết định dựa trên tập thể, v.v. Hơn thế, nhiều nghiên cứu thực nghiệm và lý thuyết đã chứng minh rằng các mô hình tập hợp thường đạt được độ chính xác cao hơn các mô hình đơn lẻ [8]. Một ví dụ cho điều này được thể hiện ở Hình 1.



Hình 1. Ví dụ minh họa cho Ensemble Learning

Trong hầu hết các trường hợp, các thuật toán thuộc ensemble có thể được phân vào hai loại [8]: Lựa chọn bộ phân lớp (Classifier selection) và hợp nhất bộ phân lớp (Classifier fusion). Trong đó, việc lựa chọn bộ phân lớp nghĩa là mỗi bộ phân lớp được huấn luyện để có kế quả tốt nhất với một phần của toàn bộ không gian đặc trưng (local expert). Khi có một mẫu dữ liệu mới, bộ phân lớp mà có dữ liệu huấn luyện có sự tương đồng lớn nhất với mẫu mới theo một thang đo cụ thể sẽ được lựa chọn để đưa dự đoán, hoặc sẽ được đánh trọng số lớn nhất trong quá trình dự đoán. Đối với hợp nhất bộ phân lớp, tất cả bộ phân lớp sẽ được huấn luyện trên toàn bộ không gian đặc trưng, sau đó được kết hợp thành một bộ phân lớp có phương sai thấp hơn (như vậy lỗi sẽ thấp hơn). Bagging, random forets, arc-x4 và boosting/Adaboost là những thuật toán phổ biến đại diện cho loại hợp nhất bộ phân lớp [8].

2. THUẬT TOÁN ADABOOST VÀ CÁC BIẾN THỂ

AdaBoost

Một trong những ý tưởng chính của thuật toán là duy trì phân phối D của các trọng số trên tập huấn luyện $\{(x_1,y_1),\ldots,(x_M,y_M); x_i \in X, y_i \in Y\}$. Trọng số của phân phối này trên mẫu thứ i trong vòng lặp t được kí hiệu là $D_t(i)$. Ở mỗi vòng, trọng số của các mẫu được phân loại không chính xác sẽ tăng lên khiến cho mô hình học yếu buộc phải tập trung vào các mẫu "khó" trong quá trình huấn luyện. Mục tiêu của mô hình học yếu là tìm ra giả thuyết h_t phù hợp cho phân bố D_t.

Thuật toán Adaboost sẽ được trình bày theo bài báo ban đầu [2] như sau:

TẠP CHÍ Khoa học & Công nghệ
Trường Đại học Công nghiệp Hà Nội
P-ISSN 1859-3585
E-ISSN 2615-9619
Tòa soạn: Tầng 3 - Nhà A1- Đại học Công nghiệp Hà Nội
Địa chỉ: 298 Đường Cầu Diễn, Q. Bắc Từ Liêm, TP. Hà Nội
Điện thoại: 024 37655121 (Số máy lẻ 8251) I Fax: 024 37655261
E-mail: tapchikhcn@haui.edu.vn I Website: http://www.tapchikhcn.haui.edu.vn

- Cho tập dữ liệu huấn luyện $(x_1,\ y_1), \ldots, (x_m,\ y_m)$. Trong đó $x_i\ \in\ X$ và y_i là nhãn của lớp.
- Khởi tạo trọng số cho tất cả các mẫu huấn luyện với giá trị bằng nhau:

$$D_1(i) = \frac{1}{m} \qquad \text{với } i = 1,\ldots, \text{m.}$$

- Lặp lại quá trình sau T lần, với t = 1, …, T:
  (1) Huấn luyện mô hình yếu với phân phối D_t
  (2) Trích xuất nhãn dự đoán h_t
  (3) Chọn mô hình yếu h_t có tỷ lệ lỗi nhỏ nhất, với:

$$\epsilon_t = Pr_{i\sim D_t}[h_t(x_i)\ \neq\ y_i]$$

  (4) Tính toán trọng số cho mô hình yếu thứ t:

$$\alpha_t = \frac{1}{2}\log\left(\frac{1-\epsilon_t}{\epsilon_t}\right)$$

  (5) Cập nhật trọng số cho tất cả mẫu dữ liệu huấn luyện theo mô hình yếu được chọn:

$$D_{t+1}(i) = \frac{D_t(i)}{E_t} \times \begin{cases} e^{-\alpha_t} & \text{nếu } h_t(x_i)\ =\ y_i \\ e^{\alpha_t} & \text{nếu } h_t(x_i)\ \neq\ y_i \end{cases}$$

  Trong đó, việc chia cho E_t có tác dụng để đảm bảo rằng phân phối D_{t+1} là một phân phối hợp lệ (có tổng bằng 1).

$$E_t = \sum_{i=1}^{N} D_t(i) e^{-y_i h_t(i)\alpha_t}$$

- Sau khi đã có được T mô hình yếu, đầu ra sẽ được tính bằng cách:

$$H(x) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

Adaboost có thể được sử dụng để giải quyết các bài toán phân loại đa lớp. Với điều kiện là các mô hình học yếu có độ chính xác không dưới 50%, nếu không mô hình có thể không hoạt động hiệu quả. Vì vậy, những bài toán phân loại đa lớp thường được đơn giản hóa bằng việc giải quyết nhiều bài toán phân loại hai lớp.

AdaBoost.M1

Tiếp đó, Freund và Schapire đã đề xuất thuật toán AdaBoost.M1, đây là một dạng tổng quát hóa đơn giản của AdaBoost cho bài toán phân loại đa lớp bằng cách sử dụng các bộ phân loại đa lớp. Thuật toán AdaBoost.M1 chỉ khác với thuật toán AdaBoost ban đầu ở bước cuối cùng sau khi đã thực hiện quá trình huấn luyện. Cụ thể hơn, thay vì sử dụng hàm sign thì dùng hàm argmax để lấy ra lớp có xác suất vào cao nhất. Cụ thể như sau [6]:

$$H(x) = \arg\max_{y\in Y} f(x,y) = \arg\max_{y\in Y}\left(\sum_{t=1}^{T} \alpha_t I(h_t(x) = y)\right)$$

TẠP CHÍ Khoa học & Công nghệ
Trường Đại học Công nghiệp Hà Nội
P-ISSN 1859-3585
E-ISSN 2615-9619
Tòa soạn: Tầng 3 - Nhà A1- Đại học Công nghiệp Hà Nội
Địa chỉ: 298 Đường Cầu Diễn, Q. Bắc Từ Liêm, TP. Hà Nội
Điện thoại: 024 37655121 (Số máy lẻ 8251) I Fax: 024 37655261
E-mail: tapchikhcn@haui.edu.vn I Website: http://www.tapchikhcn.haui.edu.vn

AdaBoost.MH

AdaBoost.MH – M có nghĩa là multi-class và H có nghĩa là Hamming (Hamming loss) là một biến thể của AdaBoost sử dụng hàm mất mát Hamming để giải quyết bài toán phân loại đa lớp.

Dựa trên việc quy giản bài toán về phân loại nhị phân, việc sử dụng boosting để giảm thiểu hamming loss trở nên khá đơn giản. Ý tưởng chính ở đây là thay thế mỗi mẫu huấn luyện (x_i, g_i) thành K mẫu $\big((x_i, l), g_i[l]\big)$ với $l \in Y$. Nói cách khác, mỗi mẫu thực chất là một cặp dữ liệu-nhãn có dạng (x_i, l) với nhãn nhị phân là +1 nếu $l \in g_i$ và nhãn là -1 trong trường hợp còn lại.

Mã giả của thuật toán AdaBoost.MH sẽ có dạng như sau [6]:

- Cho tập dữ liệu huấn luyện: $(x_1, g_1), \dots, (x_m, g_m)$ với $x_i \in X, g_i \subseteq Y$.
- Khởi tạo trọng số cho từng mẫu dữ liệu với từng nhãn:

$$D_1(i, l) = \frac{1}{mK} \qquad \text{với } i = 1, \dots, m;\ l \in Y; K = |Y|$$

- Lặp lại quá trình sau T lần, với t = 1, …, T:
  - (1) Huấn luyện mô hình yếu với phân phối D_t.
  - (2) Trích xuất nhãn dự đoán $h_t: X \times Y \to \mathbb{R}$.
  - (3) Chọn $\alpha_t \in \mathbb{R}$.
  - (4) Chọn h_t và α_t sao cho cực tiểu hóa vector chuẩn hóa

$$Z_t = \sum_{i=1}^{m} \sum_{l \in Y} D_t(i, l) e^{-\alpha_t g_i[l] h_t(x_i, l)}$$

  - (5) Cập nhật trong số cho mỗi nhãn $l \in Y$ và cho từng mẫu dữ liệu $i = 1, \dots, m$:

$$D_{t+1}(i, l) = \frac{D_t(i, l) e^{-\alpha_t g_i[l] h_t(x_i, l)}}{Z_t}$$

- Sau khi đã có được T mô hình yếu, đầu ra của mô hình dự đoán sẽ được tính bằng cách:

$$H(x, l) = \text{sign}\left(\sum_{t=1}^{T} \alpha_t h_t(x, l)\right)$$

AdaBoost.MR

Giống như đã trình bày trong phần trước AdaBoost.MH. Mỗi mẫu huấn luyện là một tập hợp (x_i, g_i), trong đó $x_i \in X$ và $g_i \subseteq Y$. Giả định học yếu sẽ có dạng $f: X \times Y \to \mathbb{R}$. Giá trị của hàm $f(x, l)$ sẽ thể hiện thứ hạng của nhãn l với bản ghi x. Có nghĩa là nhãn l_0 sẽ được xếp hạng cao hơn nhãn l_1 đối với mẫu x nếu $f(x, l_0) > f(x, l_1)$. Đối với mỗi mẫu dữ liệu (x, g), chỉ cần quan tâm đến thứ hạng của các cặp nhãn quan trọng l_0, l_1 mà $l_1 \in g$ và $l_0 \notin g$ .Giả định sẽ dự đoán sai nếu $f(x, l_0) \geq f(x, l_1)$, có nghĩa là f dự đoán sai thứ hạng của l_0 và l_1 (thứ hạng của l_1 phải cao hơn của l_0 mới đúng). Mỗi mẫu huấn luyện này được xem như một

TẠP CHÍ Khoa học & Công nghệ
Trường Đại học Công nghiệp Hà Nội
P-ISSN 1859-3585
E-ISSN 2615-9619
Tòa soạn: Tầng 3 - Nhà A1- Đại học Công nghiệp Hà Nội
Địa chỉ: 298 Đường Cầu Diễn, Q. Bắc Từ Liêm, TP. Hà Nội
Điện thoại: 024 37655121 (Số máy lẻ 8251) I Fax: 024 37655261
E-mail: tapchikhcn@haui.edu.vn I Website: http://www.tapchikhcn.haui.edu.vn

Quasi-bipartite layered feedback, thể hiện rằng mỗi nhãn trong g_i sẽ được xếp hạng cao hơn tất cả các nhãn còn lại trong $Y - g_i$. [6]

Mục tiêu ở đây là giảm thiểu tỷ lệ trung bình các cặp quan trọng bị xếp hạng sai. Giá trị này được gọi là lỗi xếp hạng (ranking loss), được biểu riễn như sau:

$$rloss_D f = E_{(x,g)\sim D}\left[\frac{|\{(l_0,l_1) \in (Y-g)\times g : f(x,l_0) \geq f(x,l_1)\}|}{|g||Y-g|}\right]$$

Từ đây, ta có thuật toán AdaBoost.MR, trong đó M thể hiện cho Multi-Class, R thể hiện cho Ranking Loss như sau [6]:

- Cho tập dữ liệu huấn luyện: $(x_1, g_1), \ldots, (x_m, g_m)$ với $x_i \in X, g_i \subseteq Y$.
- Khởi tạo trọng số cho từng mẫu dữ liệu với từng nhãn:

$$D_1(i, l_0, l_1) = \begin{cases} \dfrac{1}{m|g_i||Y-g_i|} & \text{Nếu } l_0 \notin g_i \text{ và } l_1 \in g_i \\ 0 & \text{Còn lại} \end{cases}$$

- Lặp lại quá trình sau T lần, với t = 1, …, T:

  (1) Huấn luyện mô hình yếu với phân phối D_t.

  (2) Trích xuất nhãn dự đoán $h_t: X \times Y \rightarrow \mathbb{R}$.

  (3) Chọn $\alpha_t \in \mathbb{R}$.

  (4) Cập nhật trọng số cho mỗi nhãn cặp (l_0, l_1) và cho từng mẫu dữ liệu $i = 1, \ldots, m$:

$$D_{t+1}(i, l_0, l_1) = \frac{D_t(i, l_0, l_1) e^{\frac{1}{2}\alpha_t (h_t(x_i, l_0) - h_t(x_i, l_1))}}{Z_t}$$

  Với Z_t là vector chuẩn hóa.

- Sau khi đã có được T mô hình yếu, đầu ra của mô hình dự đoán sẽ được tính bằng cách:

$$f(x, l) = \sum_{t=1}^{T} \alpha_t h_t(x, l)$$

3. ỨNG DỤNG THUẬT TOÁN ADABOOST CHO BÀI TOÁN PHÂN LỚP

3.1. Các bộ dữ liệu được sử dụng

Các bộ dữ liệu được sử dụng để đánh giá các thuật toán AdaBoost bao gồm: Bộ dữ liệu Marketing and Sales [4], bộ dữ liệu đậu khô [3], bộ dữ liệu Yelp [9], bộ dữ liệu Yeast [5] . Mô tả các bộ dữ liệu được thể hiện ở bảng dưới đây:

TẠP CHÍ Khoa học & Công nghệ
Trường Đại học Công nghiệp Hà Nội
P-ISSN 1859-3585
E-ISSN 2615-9619
Tòa soạn: Tầng 3 - Nhà A1- Đại học Công nghiệp Hà Nội
Địa chỉ: 298 Đường Cầu Diễn, Q. Bắc Từ Liêm, TP. Hà Nội
Điện thoại: 024 37655121 (Số máy lẻ 8251) I Fax: 024 37655261
E-mail: tapchikhcn@haui.edu.vn I Website: http://www.tapchikhcn.haui.edu.vn

Bảng 1. Bảng mô tả bộ dữ liệu được sử dụng để đánh giá các thuật toán AdaBoost

Bộ dữ liệu	Tổng số bản ghi	Số thuộc tính	Kiểu phân lớp	Số lớp
Marketing and Sales	45,211	16	Nhị phân	2
Đậu khô	13,166	16	Đa lớp	7
Yelp	10,806	676	Đa nhãn	5
Yeast	2,417	103	Đa nhãn	14

3.2. Kết quả phân lớp trên các bộ dữ liệu

Kết quả phân lớp trên các bộ dữ liệu được thực hiện trên nhiều thuật toán, mô hình khác nhau như SVM, Logistic Regression, Random Forest, v.v. để so sánh với các thuật toán AdaBoost. Bộ phân lớp cở sở được sử dụng cho thuật toán AdaBoost là Decesion Tree với các tham số phù hợp bằng cách thử lần lượt các giá trị.

Cuối cùng, kết quả thử nghiệm được thực hiện với quy trình K-fold với k=5. Quy trình K-fold là một quy trình đơn giản mà hiệu quả để có thể đánh giá được khả năng khái quát của mô hình. Trong mỗi fold, tính các thang đo accuracy và f1-score trên các tập train, validate và test. Và kết quả cuối cùng sẽ được lấy trung bình từ kết quả của các fold.

Phân loại nhị phân với thuật toán AdaBoost cho bộ dữ liệu Marketing and Sales

Trong bộ dữ liệu Marketing and Sales chứa 45.211 bản ghi, trong đó 39.922 bản ghi có nhãn là “no” và 5.289 bản ghi có nhãn là “yes”. Bộ dữ liệu bị mất cân bằng, do đó thực hiện đánh trọng số cho các lớp (“no”, “yes”) là (0.56, 4.27).

Dưới đây là bảng kết quả thực hiện phân lớp trên bộ dữ liệu Marketing and Sales của nhiều mô hình khác nhau với quy trình K-fold, k=5 và tham số max_depth=7 cho thuật toán Decesion Tree và bộ phân lớp cơ sở là Decesion Tree của Random Forest, AdaBoost.

Bảng 2 Bảng kết quả trên bộ dữ liệu Market and Sales trên nhiều mô hình khác nhau

Dữ liệu / **Mô hình**	Tập train		Tập validate		Tập test	
	mAcc	mF1-score	mAcc	mF1-score	mAcc	mF1-score
Logistic Regression	0.793	0.474	0.792	0.472	0.790	0.472
SVM	0.778	0.421	0.777	0.417	0.781	0.431
Decision Tree	0.842	0.555	0.833	0.524	0.835	0.535
Random Forest	0.827	0.544	0.821	0.528	0.824	0.539
AdaBoost	**0.925**	**0.753**	**0.873**	**0.586**	**0.877**	**0.588**

TẠP CHÍ Khoa học & Công nghệ
Trường Đại học Công nghiệp Hà Nội
P-ISSN 1859-3585
E-ISSN 2615-9619
Tòa soạn: Tầng 3 - Nhà A1- Đại học Công nghiệp Hà Nội
Địa chỉ: 298 Đường Cầu Diễn, Q. Bắc Từ Liêm, TP. Hà Nội
Điện thoại: 024 37655121 (Số máy lẻ 8251) I Fax: 024 37655261
E-mail: tapchikhcn@haui.edu.vn I Website: http://www.tapchikhcn.haui.edu.vn

Phân loại đa lớp với thuật oán AdaBoost.M1 cho bộ dữ liệu phân loại đậu khô.

Bộ dữ liệu bao gồm 13,611 bản ghi và đã được xử lý thành dạng số, có các trường thuộc tính như hình dạng, kích thước, v.v. Bộ dữ liệu này cũng bị mất cân bằng, trọng số cho các lớp (SEKER, BARBUNYA, BOMBAY, CALI, HOROZ, SIRA, DERMASON) là (0.95, 1.47, 3.72, 1.19, 1.00, 0.73, 0.54). Ngoài ra, trước khi được sử dụng để phân lớp, bộ dữ liệu được chuẩn hóa theo Z-score.

Dưới đây là bảng kết quả trên bộ dữ liệu phân loại đậu khô của một số mô hình với quy trình K-fold, k=5 và tham số max_depth=7:

Bảng 3 Bảng kết quả trên bộ dữ liệu phân loại đậu khô trên nhiều mô hình khác nhau

Dữ liệu / **Mô hình**	Tập train		Tập validate		Tập test	
	mAcc	mF1-score	mAcc	mF1-score	mAcc	mF1-score
SVM	0.930	0.930	**0.926**	**0.926**	**0.931**	**0.931**
Decision Tree	0.918	0.918	0.906	0.906	0.913	0.913
Random Forest	0.922	0.922	0.908	0.908	0.913	0.913
AdaBoost.M1	**0.998**	**0.998**	0.919	0.919	0.919	0.919

Phân loại đa nhãn với AdaBoost.MH cho bộ dữ liệu Yelp

Yelp là một bộ dữ liệu liên quan đến đánh giá của khách hàng cho nhiều lĩnh vực khác nhau dưới dạng văn bản. Trong báo cáo này, bộ dữ liệu Yelp liên quan đến việc đánh giá của khách hàng cho nhà hàng với các nhãn liên quan đến chất lượng đồ ăn, dịch vụ, bầu không khí, giao dịch và giá cả. Bộ dữ liệu đã được xử lý sẵn từ dạng văn bản thành dạng token.

Việc phân lớp đa nhãn có trong giới hạn báo cáo này được thực hiện theo phương pháp biến đổi bài toán (Problem transformation methods). Cụ thể hơn, phương pháp được sử dụng là phương pháp PT5 [7] biến đổi bài toán thành dạng bài toán nhị phân để thực hiện với thuật toán AdaBoost.MH. Nhãn ban đầu sẽ sở thành thuộc tính mới trong bản ghi, nhãn mới là nhãn có giá trị nhị phân.

Bảng 4 Bảng kết quả trên bộ dữ liệu Yelp trên nhiều mô hình khác nhau

Dữ liệu / **Mô hình**	Tập train		Tập validate		Tập test	
	mAcc	mF1-score	mAcc	mF1-score	mAcc	mF1-score
SVM	0.930	0.930	0.926	0.926	0.931	0.931
Decision Tree	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**
Random Forest	0.677	0.677	0.675	0.677	0.651	0.651
AdaBoost.MH	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**	**1.000**

TẠP CHÍ Khoa học & Công nghệ
Trường Đại học Công nghiệp Hà Nội
P-ISSN 1859-3585
E-ISSN 2615-9619

Tòa soạn: Tầng 3 - Nhà A1- Đại học Công nghiệp Hà Nội
Địa chỉ: 298 Đường Cầu Diễn, Q. Bắc Từ Liêm, TP. Hà Nội
Điện thoại: 024 37655121 (Số máy lẻ 8251) I Fax: 024 37655261
E-mail: tapchikhcn@haui.edu.vn I Website: http://www.tapchikhcn.haui.edu.vn

Phân loại đa nhãn với AdaBoost.MR cho bộ dữ liệu Yeast

Cũng thực hiện biến đổi bài toán đa nhãn theo phương pháp PT5, đưa bài toán về dạng bài bài nhị phân và thực hiện phân lớp với quy trình K-Fold, k=5. Dưới đây là kết quả thực nghiệm tìm kiếm tham số chiều sâu sao cho thuật toán Decision Tree đạt kết quả thích hợp nhất:

Bảng 5 Bảng kết quả trên bộ dữ liệu Yeast trên nhiều mô hình khác nhau

Dữ liệu / Mô hình	Tập train		Tập validate		Tập test	
	mAcc	mF1-score	mAcc	mF1-score	mAcc	mF1-score
SVM	0.698	0.0	0.698	0.0	0.696	0.0
Decision Tree	0.981	0.969	0.719	0.529	**0.817**	0.692
Random Forest	0.973	0.954	0.772	0.557	0.771	0.492
AdaBoost.MR	**1.000**	**1.000**	**0.802**	**0.802**	0.800	**0.800**

Từ các bảng kết quả trên, có thể nhận thấy các kết quả của các mô hình AdaBoost nhìn chung mang lại kết quả cao hơn khi so với các mô hình khác như SVM, Random Forest, Decision Tree.

5. KẾT LUẬN

Từ các kết quả ứng dụng đã nêu, có thể thấy rằng độ hiệu quả của thuật toán AdaBoost cho bài toán phân lớp vẫn còn hữu hiệu. Và việc ứng dụng vào các bài toán cụ thể cần được thử nghiệm sao cho đạt được kết quả cao nhất.

Tuy nhiên, kết quả trên bộ dữ liệu Marketing and Sales thể hiện rằng việc mất cân bằng dữ liệu vẫn là một vấn đề mở chưa có giải pháp toàn diện. Và AdaBoost cũng không phải ngoại lệ cho vấn đề đó.

TÀI LIỆU THAM KHẢO

1. Brown G (2010) Ensemble Learning. In: Sammut C, Webb GI (eds) Encyclopedia of Machine Learning. Springer US, Boston, MA, pp 312–320

2. Freund Y, Schapire RE (1997) A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. J Comput Syst Sci 55:119–139. doi: https://doi.org/10.1006/jcss.1997.1504

3. Koklu M, Ozkan IA (2020) Multiclass classification of dry beans using computer vision and machine learning techniques. Comput Electron Agric 174:105507. doi: https://doi.org/10.1016/j.compag.2020.105507

4. Moro S, Rita, P, Cortez P (2012) Bank Marketing

TẠP CHÍ Khoa học & Công nghệ
Trường Đại học Công nghiệp Hà Nội
P-ISSN 1859-3585
E-ISSN 2615-9619

Tòa soạn: Tầng 3 - Nhà A1- Đại học Công nghiệp Hà Nội
Địa chỉ: 298 Đường Cầu Diễn, Q. Bắc Từ Liêm, TP. Hà Nội
Điện thoại: 024 37655121 (Số máy lẻ 8251) I Fax: 024 37655261
E-mail: tapchikhcn@haui.edu.vn I Website: http://www.tapchikhcn.haui.edu.vn

5. Nakai K (1996) Yeast

6. Schapire R, Freund Y (2012) Boosting: Foundations and Algorithms

7. Tsoumakas G, Katakis I (2007) Multi-label classification: An overview. Int J Data Warehous Min 3:1–13. doi: 10.4018/jdwm.2007070101

8. Zhang C, Ma Y (2012) Ensemble machine learning: Methods and applications

9. Yelp Dataset Challenge. http://www.yelp.com/dataset_challenge/. Accessed 5 Jan 2024

AUTHORS INFORMATION

Nguyen Trung Duc[1], Dinh Minh Dai[1]

[1]Faculty of Information Technology, Hanoi University of Industry