SỬ DỤNG PHẦN MỀM MS EXCEL DỰ BÁO THÔNG TIN THEO PHÂN LỚP NAÏVE BAYES

USING MS EXCEL TO FORECAST INFORMATION ACCORDING TO NAÏVE BAYES CLASSIFICATION

HÀ ĐỒNG HƯNG(*)

TÓM TẮT: *Dự báo thông tin có vai trò quan trọng trong việc hoạch định của tất cả các lĩnh vực ngành nghề. Việc dự báo thông tin chính xác sẽ đem lại nhiều lợi ích cho các cá nhân và tổ chức. Để dự báo, chúng ta có thể sử dụng các phần mềm chuyên dụng với các chi phí về bản quyền và đào tạo. Vấn đề đặt ra là tìm một giải pháp áp dụng công cụ thông dụng để dự báo thông tin. Bài viết này đề xuất giải pháp áp dụng phần mềm MS Excel, một phần mềm rất phổ biến và dễ sử dụng, để khai thác dữ liệu, dự báo thông tin theo phân lớp Naïve Bayes. Kết quả thử nghiệm với dữ liệu cho thấy: chúng ta có thể dự báo được thông tin dựa vào sự phân lớp dữ liệu; Dữ liệu huấn luyện được bổ sung một cách dễ dàng bằng cách nhập thêm vào tập tin MS Excel; Các công thức đã viết sẽ tự động cập nhật kết quả khi có bất kỳ sự thay đổi nào của tập huấn luyện làm tăng độ tin cậy của thông tin dự báo.*

Từ khóa: *dự báo; phân lớp; khai thác dữ liệu; Naïve Bayes.*

ABSTRACT: *Information forecasting plays an important role in industry planning. Accurate forecasted information will bring many benefits to individuals and organizations. Specialized software can be used with beneficial licensing and training costs. The problem is to find a solution applying popular tools to forecast information. This paper proposes applying MS Excel, a very popular and easy-to-use software in mining data and forecasting information according to the Naïve Bayes classification. Experimental results with data show that: Forecast information is produced based on data classification; The training data can be easily added by appending into the MS Excel file; That written formulas automatically update the results in any change in the training set increases forecasted information reliability.*

Key words: *forecasting; classification; data mining; Naïve Bayes.*

1. ĐẶT VẤN ĐỀ

Ngày nay, dữ liệu là hạt nhân của mọi hoạt động trong các lĩnh vực ngành nghề, từ dữ liệu chúng ta có thể có được các thông tin hữu ích. Tuy nhiên, làm sao để chuyển từ các dữ liệu đó thành các thông tin hữu ích phục vụ con người là một vấn đề đã đang và sẽ tiếp tục được nghiên cứu. Xử lý dữ liệu có nhiều phương pháp gồm các phương pháp thủ công và các phương pháp tự động. Các phương pháp thủ công tốn rất nhiều công sức, xử lý chậm, dễ sai và không phù hợp với xử lý dữ liệu lớn. Các phương pháp tự động nhanh chóng, chính xác, hiệu quả với xử lý dữ liệu lớn.

Một nghiên cứu của Fatimetou Zahra Mohamed Mahmoud đã kết luận rằng: *"Thật vậy, phân tích dự báo là hệ thống đã được trong các ngành nghề khác nhau cho các mục đích khác nhau, một số thu được kết quả mong muốn và số khác thì không. Trong khi hầu hết*

(*) ThS. Trường Đại học Văn Lang, *hung.hd@vlu.edu.vn, Mã số: TCKH25-03-2021*

các nghiên cứu tập trung vào việc phát triển và tạo ra các mô hình. Nhưng điều này có đủ không?"[3]. Câu hỏi trong kết luận cho chúng ta ý tưởng cần có thêm những nghiên cứu mang tính chất thực nghiệm ứng dụng về dự báo thông tin như bài viết này.

Trong một nghiên cứu của Vaibhav Kumar và M. L. Garg đã kết luận: *"dựa trên các tham số đầu vào, đầu ra hay tương lai của bất kỳ giá trị nào có thể được dự đoán"*[4]. Vì vậy, để dự báo thông tin, chúng ta cần một lượng các dữ liệu đầu vào làm cơ sở cho dự báo.

Hiện nay, trên thế giới đã có các phần mềm chuyên dụng hay những tính năng thêm vào (Plug-in) vào MS Excel để dự báo thông tin. Tuy nhiên, việc sử dụng chúng đòi hỏi nhiều về chi phí bản quyền và chi phí đào tạo.

Chúng ta có một tập các dòng dữ liệu, trong đó mỗi dòng dữ liệu bao gồm các thuộc tính điều kiện và một thuộc tính kết quả. Tập dữ liệu này được gọi là tập dữ liệu huấn luyện (tập học). Vậy, khi chúng ta có thêm những dòng dữ liệu mới đã xác định được các giá trị ở các thuộc tính điều kiện thì thuộc tính kết quả được dự báo sẽ có kết quả như thế nào?

Phương pháp phân lớp Naïve Bayes được sử dụng để giải quyết vấn đề này. Tuy nhiên, khi thực hiện thủ công, tập dữ liệu lớn sẽ tốn rất nhiều thời gian, công sức và dễ sai sót. Mỗi khi có biến động về dữ liệu trong tập huấn luyện thì phải làm lại từ đầu. Nếu dùng phần mềm Excel, chúng ta chỉ cần viết các hàm thực thi trên tập dữ liệu huấn luyện sẽ cho ra kết quả dự báo tức thì, không sai sót; Hoặc, khi có biến động trong tập dữ liệu huấn luyện, Excel sẽ lập tức cập nhật, cho kết quả dự báo tốt nhất. Đặc biệt, phần mềm Excel rất phổ biến, linh hoạt tùy biến và dễ sử dụng. Bài viết này sẽ trình bày cách dùng phần mềm Excel để dự báo thông tin theo phân lớp Naïve Bayes.

2. NỘI DUNG

Naïve Bayes là một kỹ thuật để xây dựng bộ phân lớp: Gán nhãn lớp cho các trường hợp vấn đề, trong đó các nhãn lớp được rút ra từ một số tập hữu hạn của các giá trị thuộc tính kết quả. Một lợi thế của Native Bayes là chỉ cần một lượng nhỏ dữ liệu huấn luyện để tính các tham số cần thiết cho việc phân lớp.

Cho V_1, V_2,…, V_m là phân hoạch không gian mẫu V, mỗi V_i là một lớp. Không gian các thể hiện X gồm các thể hiện được mô tả bởi tập thuộc tính A_1, A_2,…, A_n. Không gian các thể hiện X là tập học. Khi có thể hiện mới với giá trị <a_1, a_2,…, a_n>, bộ phân lớp sẽ xuất giá trị hàm phân lớp f(x) là một trong các V_i.

Tiếp cận Bayes lấy giá trị có xác suất cao nhất V_{MAP} cho thể hiện mới. Chữ MAP viết tắt của cụm từ *Maximum A Posterior*.

$$V_{MAP} = \max P(v_j)P(a_1, a_2, \dots, a_n, |v_j)$$

Trong công thức trên có hai số hạng cần quan tâm là $P(v_j)$ và $P(a_1, a_2, \dots, a_n)$. Ta tính $P(v_j)$ bằng cách đếm số lần xuất hiện của giá trị đích v_j trong tập học. Để tính $P(a_1, a_2,\dots, a_n)$ ta giả thiết ban đầu các thuộc tính là độc lập nhau. Nói cách khác, xác suất của một thể hiện quan sát được < a_1, a_2,…, a_n > trên mỗi lớp v_j là tích các khả năng của từng thuộc tính riêng biệt trên v_j.

$$P(a_1, a_2, \dots, a_n|v_j) = \prod_i P(a_i|v_j)$$

Do vậy, công thức được viết lại là:

$$V_{NB} = \max_{v_j \in V} P(v_j) \prod_i P(a_i|v_j)$$

Với NB là viết tắt của cụm từ Naïve Bayes"[1]

2.1. Dữ liệu huấn luyện

Giả sử chúng ta có tập dữ liệu gọi là tập dữ liệu huấn luyện bao gồm các thuộc tính điều kiện: Tuổi (Già, Trẻ, Trung niên), Thu nhập (Cao, Thấp, Trung bình), Sinh viên (Không, Phải), Hạng tín dụng (Bình thường, Tốt); thuộc tính kết quả: Mua máy tính (Có, Không).

2.2. Áp dụng phương pháp phân lớp Naïve Bayes

Ước lượng $P(v_j)$ với v_1 = "Có", v_2 = "Không", và $P(a_i \mid v_j)$. Ta thu được $P(v_j)$:

$P(v_1)$ = P(Mua máy tính = Có) = 6/10

$P(v_2)$ = P(Mua máy tính = Không) = 4/10

Và

Bảng 1. Tập dữ liệu huấn luyện

Tuổi	Thu nhập	Sinh viên	Hạng tín dụng	Mua máy tính
Trung niên	Cao	Không	Bình thường	Không
Trẻ	Cao	Không	Tốt	Không
Trung niên	Cao	Không	Bình thường	Có
Già	Trung bình	Không	Bình thường	Có
Già	Thấp	Phải	Bình thường	Có
Già	Thấp	Phải	Tốt	Không
Trung niên	Thấp	Phải	Tốt	Có
Trẻ	Trung bình	Không	Bình thường	Không
Trẻ	Thấp	Phải	Bình thường	Có
Già	Trung bình	Phải	Bình thường	Có

Nguồn: Dịch từ [2]

Dự đoán cho dữ liệu mới:

Bảng 2. Dữ liệu dự báo

Tuổi	Thu nhập	Sinh viên	Hạng tín dụng	Mua máy tính
Trung niên	Trung bình	Phải	Bình thường	?

Nguồn: Dịch từ [2]

Bảng 3. Xác suất theo thuộc tính và phân lớp

Tuổi			
P(Tuổi = Già | Mua máy tính = Có)	3/6	P(Tuổi = Già | Mua máy tính = Không)	1/4
P(Tuổi = Trẻ | Mua máy tính = Có)	1/6	P(Tuổi = Trẻ | Mua máy tính = Không)	2/4
P(Tuổi = Trung niên | Mua máy tính = Có)	2/6	P(Tuổi = Trung niên | Mua máy tính = Không)	1/4
Thu nhập			
P(Thu nhập = Cao | Mua máy tính = Có)	1/6	P(Thu nhập = Cao | Mua máy tính = Không)	2/4
P(Thu nhập = Thấp | Mua máy tính = Có)	3/6	P(Thu nhập = Thấp | Mua máy tính = Không)	1/4
P(Thu nhập = Trung bình | Mua máy tính = Có)	2/6	P(Thu nhập = Trung bình | Mua máy tính = Không)	1/4
Sinh viên			
P(Sinh viên = Không | Mua máy tính = Có)	2/6	P(Sinh viên = Không | Mua máy tính = Không)	3/4
P(Sinh viên = Phải | Mua máy tính = Có)	4/6	P(Sinh viên = Phải | Mua máy tính = Không)	1/4
Hạng tín dụng			
P(Hạng tín dụng = Bình thường | Mua máy tính = Có)	5/6	P(Hạng tín dụng = Bình thường | Mua máy tính = Không)	2/4
P(Hạng tín dụng = Tốt | Mua máy tính = Có)	1/6	P(Hạng tín dụng = Tốt | Mua máy tính = Không)	2/4

Phân lớp:

X^{new} = (Tuổi = Trung niên, Thu nhập = Trung bình, Sinh viên = Phải, Tín nhiệm = Bình thường)

Ta cần tính:

P(Mua máy tính = Có) P(X^{new} | Mua máy tính = Có) = 6/10 * 2/6 * 2/6 * 4/6 * 5/6 = 0.037

P(Mua máy tính = Không) P(X^{new} | Mua máy tính = Không) = 4/10 * 1/4 * 1/4 * 1/4 * 2/4 = 0.003

Vậy X^{new} = (Tuổi = Trung niên, Thu nhập = Trung bình, Sinh viên = Phải, Tín nhiệm = Bình thường) thuộc phân lớp Mua máy tính = Có.

Trong cách xác định xác suất trên, ta hoàn toàn có thể tính được bằng cách nhẩm đếm vì tập dữ liệu huấn luyện có số lượng dòng dữ liệu ít. Trong suy luận Naïve Bayes, chỉ cần số lượng nhỏ dữ liệu để cho ra được thông tin dự đoán. Tuy nhiên, để thông tin dự đoán đạt độ tin cậy cao, ta cần một lượng dữ liệu đủ lớn. Khi có dữ liệu lớn, ta nên dùng công cụ để hỗ trợ cho hiệu quả (thời gian nhanh, tốn ít công, hạn chế tối đa sai sót,…). Một công cụ phổ biến và dễ sử dụng là phần mềm Microsoft Excel. Với việc tổ chức dữ liệu trên

Excel cùng với việc áp dụng các hàm của Excel theo phương pháp phân lớp Naïve Bayes sẽ cho ra được kết quả dự báo thông tin rất hiệu quả.

2.3. Sử dụng Microsoft Excel

Tạo tập tin Microsoft Excel đặt tên NaiveBayes.xlsx (tên này đặt theo tùy ý) bao gồm các sheet. Các sheet này được đặt tên lần lượt theo thứ tự: "Dữ liệu", "Phân lớp", "Tuổi", "Thu nhập", "Sinh viên", "Hạng tín dụng", và "Dự báo".

2.3.1. Sheet "Dữ liệu"

Sheet này chứa dữ liệu cho việc suy luận. Dữ liệu trong sheet này là các dòng với các giá trị thuộc tính điều kiện xác định và giá trị dự báo đã được kiểm nghiệm thực tế. Dữ liệu càng nhiều, dự báo cho kết quả càng đáng tin cậy. Đặt tên các tên biến tham chiếu đến địa chỉ của Sheet "Dữ liệu": CotTuoi = 'Dữ liệu'!$A:$A, CotThuNhap = 'Dữ liệu'!$B:$B, CotSinhVien = 'Dữ liệu'!$C:$C, CotHangTinDung = 'Dữ liệu'!$D:$D, CotMuaMayTinh = 'Dữ liệu'!$E:$E

	A	B	C	D	E
1	**Tuổi**	**Thu nhập**	**Sinh viên**	**Hạng tín dụng**	**Mua máy tính**
2	Trung niên	Cao	Không	Bình thường	Không
3	Trẻ	Cao	Không	Tốt	Không
4	Trung niên	Cao	Không	Bình thường	Có
5	Già	Trung bình	Không	Bình thường	Có
6	Già	Thấp	Phải	Bình thường	Có
7	Già	Thấp	Phải	Tốt	Không
8	Trung niên	Thấp	Phải	Tốt	Có
9	Trẻ	Trung bình	Không	Bình thường	Không
10	Trẻ	Thấp	Phải	Bình thường	Có
11	Già	Trung bình	Phải	Bình thường	Có

Hình 1. Sheet "Dữ liệu"

2.3.2. Sheet "Phân lớp"

Sheet này chứa xác suất cho mỗi phân lớp dựa trên sheet dữ liệu. Trong tập huấn luyện gồm có 2 phân lớp cho 2 dự đoán "Có" hoặc "Không" trong dự đoán thông tin có mua máy tính hay không mua máy tính. Các giá trị xác suất được tính dựa vào các hàm thống kê của Microsoft Excel:

B2 = COUNTIF (CotMuaMayTinh, A2) / (COUNTA (CotMuaMayTinh) - 1);

B3 = COUNTIF (CotMuaMayTinh, A3) / (COUNTA (CotMuaMayTinh) - 1).

Đặt tên các tên biến tham chiếu đến địa chỉ: XSLopCo='Phân lớp'!$B$2, XSLopKhong='Phân lớp'!$B$3.

	A	B	C
1	**Phân lớp**	**Xác suất**	
2	Có	0.6000	
3	Không	0.4000	

Hình 2. Sheet "Phân lớp"

2.3.3. Sheet "Tuổi"

Sheet này chứa các giá trị của thuộc tính tuổi: già, trẻ, trung niên (được sắp xếp tăng dần) cùng với các xác suất phân lớp tương ứng. Công thức tính các xác suất như sau:

B2 = COUNTIFS (CotTuoi, A2, CotMuaMayTinh, "Có") / COUNTIF (CotMuaMayTinh, "Có")

B3 = COUNTIFS (CotTuoi, A3, CotMuaMayTinh, "Có") / COUNTIF (CotMuaMayTinh, "Có")

B4 = COUNTIFS (CotTuoi, A4, CotMuaMayTinh, "Có") / COUNTIF (CotMuaMayTinh, "Có")

C2 = COUNTIFS (CotTuoi, A2, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không")

C3 = COUNTIFS (CotTuoi, A3, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không")

C4 = COUNTIFS (CotTuoi, A4, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không")

Đặt tên các tên biến tham chiếu đến địa chỉ:

VungGTTuoi=Tuổi!$A$2:$A$4

VungXSTuoiLopCo=Tuổi!$B$2:$B$4

VungXSTuoiLopKhong=Tuổi!$C$2:$C$4

	A	B	C
1	**Giá trị thuộc tính "Tuổi"**	**Xác suất phân lớp "Có"**	**Xác suất phân lớp "Không"**
2	Già	0.5000	0.2500
3	Trẻ	0.1667	0.5000
4	Trung niên	0.3333	0.2500
5			

Hình 3. Sheet "Tuổi"

2.3.4. Sheet "Thu nhập"

Sheet này chứa các giá trị của thuộc tính thu nhập: cao, thấp, trung bình (được sắp xếp tăng dần) cùng với các xác suất phân lớp tương ứng. Công thức tính các xác suất như sau: B2 = COUNTIFS (CotThuNhap, A2, CotMuaMayTinh, "Có") / COUNTIF (CotMuaMayTinh, "Có"); B3 = COUNTIFS (CotThuNhap, A3, CotMuaMayTinh, "Có") / COUNTIF (CotMuaMayTinh, "Có"); B4 = COUNTIFS (CotThuNhap, A4, CotMuaMayTinh,

"Có") / COUNTIF (CotMuaMayTinh, "Có"); C2 = COUNTIFS (CotThuNhap, A2, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không"); C3 = COUNTIFS (CotThuNhap, A3, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không"); C4 = COUNTIFS (CotThuNhap, A4, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không").

Đặt tên các tên biến tham chiếu đến địa chỉ:

VungGTThuNhap = 'Thu nhập'!$A$2:$A$4,

VungXSThuNhapLopCo = 'Thu nhập'!$B$2:$B$4,

VungXSThuNhapLopKhong = 'Thu nhập'!$C$2:$C$4.

	A	B	C
1	**Giá trị thuộc tính "Thu nhập"**	**Xác suất phân lớp "Có"**	**Xác suất phân lớp "Không"**
2	Cao	0.1667	0.5000
3	Thấp	0.5000	0.2500
4	Trung bình	0.3333	0.2500
5			

Hình 4. Sheet "Thu nhập"

2.3.5. Sheet "Sinh viên"

Sheet này chứa các giá trị của thuộc tính sinh viên: không, phải (được sắp xếp tăng dần) cùng với các xác suất phân lớp tương ứng. Công thức tính các xác suất như sau: B2 = COUNTIFS (CotSinhVien, A2, CotMuaMayTinh, "Có") / COUNTIF (CotMuaMayTinh, "Có"); B3 = COUNTIFS (CotSinhVien, A3, CotMuaMayTinh, "Có") / COUNTIF (CotMuaMayTinh, "Có"); C2 = COUNTIFS (CotSinhVien, A2, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không"); C3 = COUNTIFS (CotSinhVien, A3, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không").

Đặt tên các tên biến tham chiếu đến địa chỉ:

VungGTSinhVien = 'Sinh viên'!$A$2:$A$3,

VungXSSinhVienLopCo = 'Sinh viên'!$B$2:$B$3,

VungXSSinhVienLopKhong = 'Sinh viên'!$C$2:$C$3.

	A	B	C
1	**Giá trị thuộc tính "Sinh viên"**	**Xác suất phân lớp "Có"**	**Xác suất phân lớp "Không"**
2	Không	0.3333	0.7500
3	Phải	0.6667	0.2500

Hình 5. Sheet "Sinh viên"

2.3.6. Sheet "Hạng tín dụng"

Sheet này chứa các giá trị của thuộc tính hạng tín dụng: bình thường, tốt (được sắp xếp tăng dần) cùng với các xác suất phân lớp tương ứng. Công thức tính các xác suất như sau: B2 = COUNTIFS (CotHangTinDung, A2, CotMuaMayTinh, "Có") / COUNTIF (CotMuaMayTinh, "Có"); B3 = COUNTIFS (CotHangTinDung, A3, CotMuaMayTinh, "Có") / COUNTIF (CotMuaMayTinh, "Có"); C2 = COUNTIFS (CotHangTinDung, A2, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không"); C3 = COUNTIFS (CotHangTinDung, A3, CotMuaMayTinh, "Không") / COUNTIF (CotMuaMayTinh, "Không").

Đặt tên các tên biến tham chiếu đến địa chỉ:

VungGTHangTinDung = 'Sinh viên'!$A$2:$A$3,

VungXSHangTinDungLopCo = 'Sinh viên'!$B$2:$B$3,

VungXSHangTinDungLopKhong = 'Sinh viên'!$C$2:$C$3.

	A	B	C
1	**Giá trị thuộc tính "Hạng tín dụng"**	**Xác suất phân lớp "Có"**	**Xác suất phân lớp "Không"**
2	Bình thường	0.8333	0.5000
3	Tốt	0.1667	0.5000

Hình 6. Sheet "Hạng tín dụng"

2.3.7. Sheet "Dự báo"

Sheet này chứa giá trị các thuộc tính điều kiện (tuổi, thu nhập, sinh viên, hạng tín dụng) của dòng dữ liệu cần dự báo để cho ra thông tin kết quả dự báo (mua máy tính). Ta nhập thông tin cho dòng dữ liệu cần dự báo: Tuổi = Trung niên, Thu nhập = Trung bình, Hạng tín dụng = Bình thường. Sau đó ta viết hàm cho E2, F2 và G2 như sau:

E2=XSLopCo*LOOKUP(A2,VungGTTuoi, VungXSTuoiLopCo)*LOOKUP(B2,VungGTThuNhap, VungXSThuNhapLopCo)*LOOKUP(C2,VungGTSinhVien,VungXSSinhVienLopCo)*LOOKUP(D2,VungGTHangTinDung,VungXSHangTinDungLopCo);

F2=XSLopKhong*LOOKUP(A2,VungGTTuoi,VungXSTuoiLopKhong)*LOOKUP(B2, VungGTThuNhap,VungXSThuNhapLopKhong)*LOOKUP(C2,VungGTSinhVien,VungXSSinh

VienLopKhong)*LOOKUP(D2,VungGTHang TinDung, VungXSHangTinDungLopKhong);

G2 = IF (OR (MIN (E2:F2) = 0, F2 = E2), "Chưa thể dự đoán", IF (E2>F2, "Có", "Không")).

Vậy X^{new} = (Tuổi = Trung niên, Thu nhập = Trung bình, Sinh viên = Phải, Tín nhiệm = Bình thường) thuộc phân lớp Mua máy tính = Có.

	A	B	C	D	E	F	G
1	Tuổi	Thu nhập	Sinh viên	Hạng tín dụng	Xác suất Phân lớp Mua	Xác suất Phân lớp Không mua	Dự báo Mua máy tính
2	Trung niên	Trung bình	Phải	Bình thường	0.0370	0.0031	Có

Hình 7. Sheet "Dự báo"

3. KẾT LUẬN

Với một tập dữ liệu được huấn luyện, áp dụng phương pháp Naïve Bayes và sử dụng phần mềm Excel, ta hoàn toàn có thể dự báo được thông tin dựa vào sự phân lớp dữ liệu. Dữ liệu huấn luyện được bổ sung một cách dễ dàng chỉ bằng cách nhập thêm dữ liệu vào tập tin Excel. Các công thức đã viết sẽ cập nhật tự động kết quả khi có sự thay đổi của tập huấn luyện làm cho độ tin cậy của thông tin dự báo ngày càng cao. Qua cách thức phân tích dữ liệu và dự báo trên, rõ ràng thấy được rằng, ta có thể áp dụng phương pháp này cho các dữ liệu tương tự khác mà ta cần để dự báo kết quả chỉ bằng cách thay đổi dữ liệu được huấn luyện cho phù hợp.

TÀI LIỆU THAM KHẢO

[1] Đỗ Phúc (2009), *Giáo trình khai thác dữ liệu*, Nxb Đại học Quốc gia Thành phố Hồ Chí Minh.

[2] Jing Gao (Fall 2013), *Data Mining and Bioinformatics*, https://cse.buffalo.edu/~jing/cse601/fa13/materials/classification_methods.pdf, ngày truy cập: 26-08-2020.

Ngày nhận bài: 22-8-2020. Ngày biên tập xong: 06-01-2021. Duyệt đăng: 25-3-2021