

ỨNG DỤNG THUẬT TOÁN BAYES TRONG VẤN ĐỀ DỰ BÁO HỌC LỰC CỦA HỌC SINH PHỔ THÔNG

Đào Đức Anh¹, Nguyễn Tu Trung¹, Vũ Văn Thỏa²

¹Đại học Thủy Lợi

²Học viện Công nghệ Bưu chính Viễn thông

Tóm tắt: Đánh giá học lực là vấn đề quan trọng trong việc đánh giá học sinh phổ thông. Việc đánh giá dựa trên điểm các môn học của học sinh trong suốt quá trình học. Từ lâu, các thuật toán học máy nói chung, thuật toán phân lớp Bayes nói riêng đã được ứng dụng để giải quyết các bài toán phân lớp, dự báo một cách hiệu quả. Bài báo này ứng dụng thuật toán Bayes trong việc dự báo học lực của học sinh để hỗ trợ cho việc quản lý cũng như đánh giá học sinh trong trường phổ thông.

Từ khóa: Học lực, điểm trung bình, Bayes, học máy thông kê, dự báo.

I. MỞ ĐẦU

Dự báo là một khoa học và nghệ thuật tiên đoán những sự việc sẽ xảy ra trong tương lai, trên cơ sở phân tích khoa học về các dữ liệu đã thu thập được. Khi tiến hành dự báo cần căn cứ vào việc thu thập, xử lý số liệu trong quá khứ và hiện tại để xác định xu hướng vận động của các hiện tượng trong tương lai nhờ vào một số mô hình toán học (định lượng). Tuy nhiên, dự báo cũng có thể là một dự đoán chủ quan hoặc trực giác về tương lai (định tính) và để dự báo định tính được chính xác hơn, người ta cố loại trừ những tính chủ quan của người dự báo.

Có nhiều phương pháp dự báo khác nhau. Hiện nay, việc sử dụng các phương pháp học máy ứng dụng cho các bài toán dự báo trở nên rất phổ biến. Trong đó, dự báo sử dụng phân lớp Bayes được ứng dụng rất rộng rãi... Ví dụ, dự báo giá cả các loại mặt hàng, dự báo tỉ lệ tăng dân số... khi biết các thông tin trong quá khứ và điều kiện cho trước...

Phân lớp Bayes cũng được sử dụng một cách trong phân lớp chủ đề văn bản [6]. Một trong những ứng dụng rất phổ biến của phân lớp Bayes là phân loại thư rác. Trong [1], Awad đã trình bày một đánh giá, so sánh một số phương pháp học máy (Bayesian classification, k-NN, ANNs, SVMs...) cho vấn đề lọc thư rác. Trong [2], Jialin và cộng sự đã thảo luận, đánh giá về phương pháp lọc SMS rác sử dụng SVM và MTM (message topic model). Trong [4], Phan Hữu Tiếp cùng các cộng sự trình bày quy trình lọc thư rác tiếng Việt dựa trên thuật toán Naïve Bayes và việc xử lý tách câu tiếng Việt. Trong [5],

Tianda và cộng sự đã trình bày một so sánh giữa bộ phân loại thư rác chỉ sử dụng kỹ thuật Naïve Bayes và bộ phân loại thư rác sử dụng bộ phân loại thư rác kỹ thuật và luật kết hợp. Trong [3], các tác giả đã đánh giá một số cách thức tính xác suất SPAM của token trong phân loại thư rác.

Hạnh kiểm và học lực là hai yếu tố rất quan trọng của mỗi học sinh khi tham gia học tập tại trường. Trong đó, kết quả xếp loại học lực của học sinh sẽ được sử dụng để đánh giá và xét cho học sinh lên lớp và để đánh giá xếp loại khen thưởng [7]. Căn cứ vào điểm trung bình các môn học kỳ và cả năm, xếp loại học tập được chia thành 5 loại là: Giỏi, Khá, Trung bình, Yếu, Kém. Do đó, việc đánh giá xếp loại học lực học sinh được thực hiện rất chặt chẽ.

Trong bài báo này, chúng tôi đề xuất giải pháp ứng dụng thuật toán Bayes trong vấn đề dự báo học lực học sinh dựa trên điểm số các môn của học sinh.

II. ĐÁNH GIÁ HỌC LỰC

Theo quy chế đánh giá xếp loại học lực [7], kết quả học lực của học sinh được tổng hợp, tính toán và đánh giá qua các bài kiểm tra.

A. Dữ liệu phục vụ cho việc đánh giá

Các hình thức kiểm tra bao gồm: Kiểm tra miệng (kiểm tra bằng hỏi đáp), kiểm tra viết, kiểm tra thực hành.

Các loại bài kiểm tra bao gồm:

- ✓ Kiểm tra thường xuyên: Kiểm tra miệng; kiểm tra viết dưới 1 tiết, kiểm tra thực hành dưới 1 tiết.
- ✓ Kiểm tra định kỳ: Kiểm tra viết từ 1 tiết trở lên; kiểm tra thực hành từ 1 tiết trở lên, kiểm tra học kỳ.

Hệ số các loại bài kiểm tra:

- ✓ Đối với các môn học đánh giá bằng cho điểm: Điểm kiểm tra thường xuyên tính hệ số 1, điểm kiểm tra viết và kiểm tra thực hành từ 1 tiết trở lên tính hệ số 2, điểm kiểm tra học kỳ tính hệ số 3.
- ✓ Đối với các môn đánh giá bằng nhận xét: Kết quả nhận xét của các bài kiểm tra đều tính 1 lần khi xếp loại môn học sau mỗi học kỳ.

Tác giả liên hệ: Nguyễn Tu Trung

Email: Trungnt.sremis@gmail.com

Đến tòa soạn: 2/2020, chỉnh sửa 4/2020, chấp nhận đăng 4/2020

Điểm trung bình môn học kỳ (ĐTB_{mhk}) là trung bình cộng của điểm các bài KT_{tx} , $\text{KT}_{\text{đk}}$ và KT_{hk} với các hệ số quy định tại Điểm a, Khoản 3, Điều 7 Quy chế này:

$$\text{ĐTB}_{\text{mhk}} = \frac{\text{TĐKT}_{\text{tx}} + 2 \times \text{TĐKT}_{\text{đk}} + 3 \times \text{TĐKT}_{\text{hk}}}{\text{Số bài KT}_{\text{tx}} + 2 \times \text{Số bài KT}_{\text{đk}} + 3} \quad (1)$$

Trong đó:

- ✓ TĐKT_{tx} : Tổng điểm của các bài KT_{tx} .
- ✓ $\text{TĐKT}_{\text{đk}}$: Tổng điểm của các bài $\text{KT}_{\text{đk}}$.
- ✓ TĐKT_{hk} : Điểm bài KT_{hk} .

Điểm trung bình môn cả năm (ĐTB_{mcn}) là trung bình cộng của ĐTB_{mhkI} với $\text{ĐTB}_{\text{mhkII}}$, trong đó $\text{ĐTB}_{\text{mhkII}}$ tính hệ số 2:

$$\text{ĐTB}_{\text{mcn}} = \frac{\text{ĐTB}_{\text{mhkI}} + 2 \times \text{ĐTB}_{\text{mhkII}}}{3} \quad (2)$$

ĐTB_{mhk} và ĐTB_{mcn} là số nguyên hoặc số thập phân được lấy đến chữ số thập phân thứ nhất sau khi làm tròn số.

B. Tiêu chuẩn xếp loại học lực dựa trên điểm số

- Loại Giỏi:
 - ✓ Điểm trung bình các môn học từ 8.0 trở lên, trong đó điểm trung bình của 1 trong 2 môn Toán, Ngữ văn từ 8.0 trở lên.
 - ✓ Không có môn học nào điểm trung bình dưới 6.5.
 - ✓ Các môn học đánh giá bằng nhận xét đạt loại Đ.
- Loại Khá:
 - ✓ Điểm trung bình các môn học từ 6.5 trở lên, trong đó điểm trung bình của 1 trong 2 môn Toán, Ngữ văn từ 6.4 trở lên.
 - ✓ Không có môn học nào điểm trung bình dưới 5.0.
 - ✓ Các môn học đánh giá bằng nhận xét đạt loại Đ.
- Loại Trung bình
 - ✓ Điểm trung bình các môn học từ 5.0 trở lên, trong đó điểm trung bình của 1 trong 2 môn Toán, Ngữ văn từ 5.0 trở lên.
 - ✓ Không có môn học nào điểm trung bình dưới 3.5.
 - ✓ Các môn học đánh giá bằng nhận xét đạt loại Đ.
- Loại Yếu:
 - ✓ Điểm trung bình các môn học từ 3.5 trở lên.
 - ✓ Không có môn học nào điểm trung bình dưới 2.0.
- Loại Kém: Các trường học còn lại.

III. THUẬT TOÁN NAÏVE BAYES

Theo[8], có thể mô tả bài toán cần giải quyết như sau:

Dữ kiện cần có:

- D : tập dữ liệu huấn luyện, được vector hoá dưới dạng $\vec{x} = (x_1, x_2, \dots, x_n)$.
- C_i : tập các tài liệu của D thuộc lớp C_i với $i = \{1, 2, 3, \dots\}$.
- Các $x_1, x_2, \dots, x_n$ độc lập xác suất đôi một với nhau.

Thuật toán Naïve Bayes cơ bản:

- Bước 1: Huấn luyện Naïve Bayes (dựa vào tập dữ liệu)

- ✓ Tính xác suất $P(C_i)$.
- ✓ Tính xác suất $P(x_k|C_i)$.

- Bước 2: Phân lớp X_{new}

- ✓ Tính $F(X_{\text{new}}, C_i) = P(C_i) \prod_{k=1}^n P(x_k|C_i)$
- ✓ X_{new} được gán vào lớp C_q sao cho

$$F(X_{\text{new}}, C_q) = \max(F(X_{\text{new}}, C_i)) \quad (3)$$

$P(x_k|C_i)$ được tính như sau:

$$P(x_k|C_i) = \frac{c_{i,D}\{x_k\}}{|C_{i,D}|} \quad (4)$$

Trong đó:

- ✓ $C_{i,D}$ số mẫu của tập dữ liệu huấn luyện D thuộc về lớp C_i .
- ✓ $C_{i,D}\{x_k\}$ số mẫu trong tập $C_{i,D}$ mà có nhân giá trị là x_k .

IV. ĐỀ XUẤT PHƯƠNG PHÁP DỰ BÁO HỌC LỰC HỌC SINH

A. Sử dụng thuật toán Bayes để dự báo học lực

Dữ liệu đầu vào là thông tin điểm các môn học của học sinh: Toán, Vật lý, Hóa, Sinh, Tin học, Ngữ văn, Lịch sử, Địa lý, Tiếng Anh, GD&CD, KTNN, Thể dục, GDQP như hình 1.

Đầu ra là thông tin dự báo xếp loại học lực: Giỏi, Khá, Trung bình, Yếu, Kém.

Toán	VLY	Hóa	Sinh	Tin	NVAN	LSu	DLy	Anh	GD&CD	KTNN	TDuc	GDQP	ĐTB	XUHL
3.9	5.4	5.1	7.7	6.8	6.1	7.6	6.9	5.5	6.7	7.2	Đ	7.4	6.4	Tb
5.6	5.4	4.5	5.4	6.6	5.7	8.4	5.1	4.9	5.5	6.3	Đ	7.4	5.9	Tb
5.9	6.6	4.5	7.2	7.5	5.6	6.8	5.7	5.4	6.6	6.4	Đ	7.7	6.3	Tb
4.2	6.6	6.6	6	6.2	5.8	8.3	7.1	4.6	6.5	6.2	Đ	6.7	6.2	Tb
8.7	7.1	7.5	7.2	7.8	6.7	7.9	7.5	5.5	6.1	7	Đ	6.9	7.2	K
4.7	6.8	5.7	5.3	6.9	5.6	8.1	6.4	4.5	5.1	7.8	Đ	7.3	6.2	Tb
8.8	7.6	7.4	7	7.7	7.1	8.6	7.1	7.5	6.6	8.1	Đ	8.3	7.7	K
4.2	5.3	4.4	5.3	7.3	4.5	4.4	7	3.7	5.2	6.6	Đ	5.9	5.3	Y

Hình 1: Ví dụ về dữ liệu đầu vào.

Để có thể sử dụng phân lớp Bayes, ta xác định nhãn lớp C_i , $\vec{x}$ như sau:

- ✓ Nhân C_i là: Giỏi, Khá, Trung bình, Yếu, Kém.
- ✓ $\vec{x}$ là vector thông tin điểm các môn học của học sinh.

Có một vấn đề nảy sinh từ công thức (4) như sau: nếu sử dụng trực tiếp các giá trị điểm từng môn làm giá trị của $x_1, x_2, \dots, x_n$ thì trong trường hợp x_k không tồn tại trong dữ liệu huấn luyện. Khi này, giá trị của $P(x_k|C_i)$ là 0 $\forall i$. Từ đó, $F(X_{new}, C_i) = 0 \forall i$. Điều này nghĩa là, ta sẽ không chọn được nhân học lực phù hợp nhất.

Để tránh trường hợp $P(x_k|C_i) = 0$ do không có mẫu nào trong dữ liệu huấn luyện thỏa mãn từ số trong công thức (4), ta sử dụng một trong 2 phương án sau:

- Phương án 1: Không sử dụng trực tiếp điểm môn học làm dữ liệu thành phần của vector $\vec{x}$. Bài báo đề xuất 3 kỹ thuật chuyển đổi dữ liệu cho phương án này như sau:
 - ✓ Kỹ thuật 1: Phân điểm thành G (điểm ≥ 8), K ($6.5 \leq \text{điểm} < 8$), TB ($5.0 \leq \text{điểm} < 6.5$), Y ($3.5 \leq \text{điểm} < 5.0$), K (điểm < 3.5).
 - ✓ Kỹ thuật 2: Phân điểm thành số nguyên: 0, 1, 2, 3, 4, 5, 6, 7, 8, 9, 10.
 - ✓ Kỹ thuật 3: Phân mỗi điểm nguyên thành mốc A và B được phân chia bởi 0.5: 0A, 0B, 1A, 1B, 2A, 2B, 3A, 3B, 4A, 4B, 5A, 5B, 6A, 6B, 7A, 7B, 8A, 8B, 9A, 9B và 10.
- Phương án 2: Sử dụng công thức làm tròn Laplace như sau:

$$P(x_i|C_i) = \frac{c_{iD}\{x_k\}+1}{|c_{iD}|+r} \quad (5)$$

Trong đó, r là số giá trị rời rạc của thuộc tính.

B. Luật quyết định học lực dựa trên Bayes

- Luật quyết định loại Giỏi:
 - ✓ $F(X_{new}, \text{Giỏi}) = \max(F(X_{new}, C_i))$.
 - ✓ $\text{Or}(\text{ĐTB}_{\text{mhk}}(\text{Toán}), \text{ĐTB}_{\text{mhk}}(\text{Văn})) \geq 8.0$
 - ✓ $\text{ĐTB}_{\text{mhk}}(\text{Môn}[i]) < 6.5 \forall i$.
 - ✓ $\text{Điểm}(\text{Môn_Đánh_Giá}[i]) = \text{ĐV}i$.
- Luật quyết định loại Khá:
 - ✓ $F(X_{new}, \text{Khá}) = \max(F(X_{new}, C_i))$ hoặc không phải loại Giỏi.
 - ✓ $\text{Or}(\text{ĐTB}_{\text{mhk}}(\text{Toán}), \text{ĐTB}_{\text{mhk}}(\text{Văn})) \geq 6.4$.
 - ✓ $\text{ĐTB}_{\text{mhk}}(\text{Môn}[i]) < 5.0 \forall i$.
 - ✓ $\text{Điểm}(\text{Môn_Đánh_Giá}[i]) = \text{ĐV}i$.
- Luật quyết định loại Trung bình
 - ✓ $F(X_{new}, \text{Trung Bình}) = \max(F(X_{new}, C_i))$ hoặc không phải loại Giỏi, Khá.
 - ✓ $\text{Or}(\text{ĐTB}_{\text{mhk}}(\text{Toán}), \text{ĐTB}_{\text{mhk}}(\text{Văn})) \geq 5.0$.
 - ✓ $\text{ĐTB}_{\text{mhk}}(\text{Môn}[i]) < 3.5 \forall i$.
 - ✓ $\text{Điểm}(\text{Môn_Đánh_Giá}[i]) = \text{ĐV}i$.
- Luật quyết định loại Yếu:

- ✓ $F(X_{new}, \text{Yếu}) = \max(F(X_{new}, C_i))$ hoặc không phải loại Giỏi, Khá, Trung bình.

- ✓ $\text{ĐTB}_{\text{mhk}}(\text{Môn}[i]) < 2.0 \forall i$.

- Luật quyết định loại Kém: $F(X_{new}, \text{Kém}) = \max(F(X_{new}, C_i))$ hoặc không phải loại Giỏi, Khá, Trung bình, Yếu.

V. THỬ NGHIỆM

Tập dữ liệu là điểm số của học sinh của một số trường cấp 3 được thu thập trên internet (xin phép không chia sẻ vì lý do bảo mật).

Dữ liệu huấn luyện: đọc từ file Excel, bao gồm 1784 bản ghi.

Tập dữ liệu kiểm thử: lấy từ file Excel, bao gồm 199 bản ghi.

Bảng 1 thống kê thời gian huấn luyện mô hình với tập dữ liệu huấn luyện theo từng phương án và kỹ thuật cụ thể. Từ kết quả trong bảng 2, ta thấy thời gian huấn luyện mô hình của Phương án 1-Kỹ thuật 1 là nhỏ nhất còn Phương án 1-Kỹ thuật 3 là lớn nhất. Với tất cả các phương án kỹ thuật sử dụng, chúng ta thấy tốc độ huấn luyện là rất nhanh. Điều này sẽ rất thuận lợi nếu như cần thay đổi quy mô dữ liệu huấn luyện và huấn luyện lại để tăng cường độ chính xác.

Bảng 1: Tốc độ huấn luyện.

Phương án/ Kỹ thuật	Phương án 1			Phương án 2
	Kỹ thuật 1	Kỹ thuật 2	Kỹ thuật 3	
Thời gian huấn luyện	190 ms	246 ms	307 ms	263ms

Bảng 2 thống kê độ chính xác với dữ liệu test theo từng phương án và kỹ thuật cụ thể.

Bảng 2: Độ chính xác.

Phương án/ Kỹ thuật	Phương án 1			Phương án 2
	Kỹ thuật 1	Kỹ thuật 2	Kỹ thuật 3	
Độ chính xác test	95.48 %	96 %	95.48 %	95.5 %

Từ kết quả trong bảng 3, ta thấy độ chính xác trên dữ liệu test của Phương án 1-Kỹ thuật 1 là nhỏ nhất còn Phương án 1-Kỹ thuật 2 là lớn nhất. Độ chính xác chỉ đạt được như vậy có thể do tập dữ liệu huấn luyện chưa đủ lớn và bao quát miền dữ liệu điểm.

VI. KẾT LUẬN

Trong bài báo này, nhóm tác giả đã đề xuất phương pháp dự báo học lực sử dụng thuật toán phân lớp Bayes. Kết quả thử nghiệm cho thấy tốc độ huấn luyện rất nhanh và độ chính xác cao. Tuy nhiên, nếu dữ liệu huấn luyện rất lớn, đặc biệt khi thu thập dữ liệu đủ lớn để có thể phủ tất cả điểm thập phân (một hoặc hai chữ số sau dấu phẩy), thời gian huấn luyện chắc chắn sẽ lâu hơn... Vì vậy, trong nghiên cứu tiếp theo, nhóm tác giả dự định

nghiên cứu các phương án song song, phân tán để giải quyết vấn đề này.

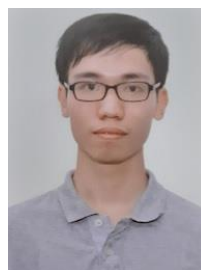
TÀI LIỆU THAM KHẢO

- [1] Awad W.A. and ELseuofi S.M., *Machine learning methods for spam e-mail classification*, International Journal of Computer Science & Information Technology (IJCSIT), Vol 3, No 1, Feb 2011, pp.173-184.
- [2] Jialin ma, Yongjun zhang, Jinling liu, *Intelligent SMS spam filtering using topic model*, iee international conference on intelligent networking and collaborative systems (incos), 2016.
- [3] Nguyễn Tu Trung, Nguyễn Ngọc Hưng, Phạm Thanh Giang, *Đánh giá một số cách thức tính xác suất SPAM của Token ứng dụng trong phân loại thư rác*, Tạp chí Học viện Bưu chính, số 3, 2018.
- [4] Phan Hữu Tiếp, Vũ Đức Lung, Cao Nguyễn Thủy Tiên, Lâm Thành Hiền, *Phương pháp lọc thư rác tiếng việt dựa trên từ ghép và theo vết người sử dụng*, Hội thảo “Một số vấn đề chọn lọc của Công nghệ thông tin và truyền thông”, Cần Thơ, 2011.
- [5] Tianda Yang, Kai Qian, Dan Chia-Tien Lo, *Spamfiltering using Association Rules and Naïve Bayes Classifier*, IEEE International Conference on Progress in Informatics and Computing (PIC), 2015.
- [6] Zhang Haiyi, Li Di, *Naïve Bayes Text Classifier*, Proceedings - 2007 IEEE International Conference on Granular Computing, 2007.
- [7] <https://thuvienphapluat.vn/van-ban/giao-duc/Thong-tu-58-2011-TT-BGDDT-Quy-che-danh-gia-xep-loai-hoc-sinh-trung-hoc-co-so-133268.aspx>
- [8] Duan Li-guo, Di peng*, Li Ai-ping, *A New Naive Bayes Text Classification Algorithm*, Indonesian Journal of Electrical Engineering, 2014.

USING BAYESIAN CLASSIFICATION IN PREDICTING LEARNING ABILITY OF HIGH SCHOOL STUDENTS

Abstract: Learning ability assessment is an important issue in assessing high school students. The assessment is based on a student's subject grades throughout the learning process. For a long time, machine learning algorithms in general and Bayes classification algorithms in particular have been applied to solve classification and prediction problems effectively. This paper applies the Bayes algorithm in predicting student performance to support the management and assessment of students in high school.

Keyword: Learning ability, Bayes, Statistical machine learning, Predicting.



Đào Đức Anh, sinh viên năm cuối trường Đại học Thủy Lợi. Bắt đầu nghiên cứu về học máy...



Nguyễn Tu Trung, tốt nghiệp Đại học Sư phạm Hà Nội 2 năm 2007, hoàn thành luận văn Thạc sĩ tại trường ĐHCông Nghệ, ĐHQGHN năm 2011, luận án Tiến sĩ, Học viện Công nghệ Bưu chính Viễn thông năm 2018. Hiện tại làm việc tại trường Đại học Thủy Lợi. Lĩnh vực nghiên cứu: Xử lý ảnh, xử lý tiếng nói, hệ thống thông tin, hệ thống nhúng.



Vũ Văn Thỏa, Tốt nghiệp Đại học Sư phạm Vinh năm 1975, Tiến Sĩ 1990 Viện Điều khiển tại Liên Xô cũ. Hiện công tác tại Khoa Quốc tế và Đào tạo Sau Đại học, Học viện Công nghệ Bưu chính Viễn thông.. Lĩnh vực nghiên cứu: Lý thuyết thuật toán, tối ưu hóa, hệ thống tin địa lý, mạng viễn thông