MỘT CẢI TIẾN THUẬT TOÁN DỰ BÁO HỌC LỰC HỌC SINH PHỔ THÔNG DỰA TRÊN PHƯƠNG PHÁP BAYES SỬ DỤNG MÔ HÌNH MAPREDUCE

Nguyễn Tu Trung[*], Đào Đức Anh[*], Vũ Văn Thỏa[#]
[*]Đại học Thủy Lợi
[#]Học 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. Ngoài ra, việc xây dựng các ứng dụng quản lý học sinh tập trung của toàn tỉnh, thành phố cũng như toàn quốc làm nảy sinh một khối lượng dữ liệu khổng lồ. Mô hình MapReduce hiện đang được sử dụng hiệu quả trong phân tích dữ liệu lớn. Bài báo này ứng dụng thuật toán Bayes và mô hình MapReduce 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, MapReduce, 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 khữ và điều kiện cho trước…

Phân lớp Bayes cũng được sử dụng một cách trong phân lớp chủ đề văn bản [7]. Trong [13], các tác giả đã sử dụng Deep learning để phân lớp chủ đề văn bản. 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 [3], Jialin và cộng sự đã thảo luận, đánh giá về phương pháp lọc SMS rác sử dụng SVM và MTM (message topic model). Trong [5], 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 [6], Tianda và cộng sự đã trình bày một so sánh giữa bộ phân loại thư rác chỉ sử dụng kĩ thuật Naïve Bayes và bộ phân loại thư rác sử dụng bộ phân loại thư rác kĩ thuật và luật kết hợp. Trong [4], các tác giả đã đánh giá một số cách thức tính xác suất SPAM của token trong phân loại thư rác.

Hiện nay, với sự phát triển của công nghệ thông tin, cuộc cách mạng công nghiệp 4.0 dẫn đến sự bùng nổ về dữ liệu (Big Data). Dữ liệu lớn và phân tích của nó đóng một vai trò quan trọng trong thế giới Công nghệ thông tin với các ứng dụng của Công nghệ đám mây, Khai thác dữ liệu, Hadoop và MapReduce [10]. Các công nghệ truyền thống chỉ áp dụng cho dữ liệu có cấu trúc trong khi dữ liệu lớn bao gồm cả dữ liệu có cấu trúc, bán cấu trúc và không có cấu trúc. Làm thế nào để xử lý hiệu quả dữ liệu lớn đã trở thành thức lớn trong thời đại mới và cần có những phương pháp xử lý mới. MapReduce là mô hình xử lý dữ liệu phân tán rất hiệu quả, đã và đang được ứng dụng rộng rãi trong xử lý dữ liệu lớn [2].

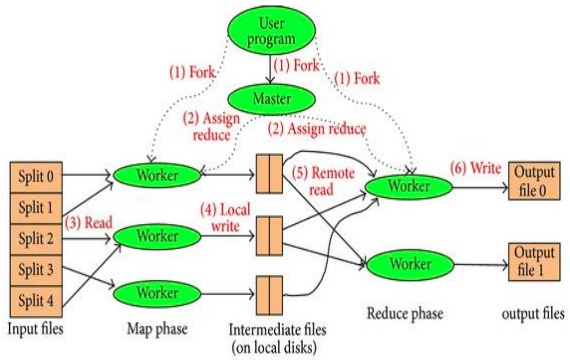
Hạnh kiểm và học lực là hai yếu tố rất quan trọng của mỗi học sinh khi tham gia học tập tại trường. Trong đó, kết quả xếp loại học lực của học sinh sẽ được sử dụng để đánh giá và xét cho học sinh lên lớp và để đánh giá xếp loại khen thưởng [8]. Căn cứ vào điểm trung bình các môn học kỳ và cả năm, xếp 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. Hiện nay, do nhu cầu của việc kết nối, chia sẽ, quản lý tập trung, dữ liệu của các trường, các cấp giáo dục được lưu trữ trên các máy chủ của một tỉnh, một quốc gia sẽ làm phát sinh một khối lượng dữ liệu khổng lồ. Vì vậy, các phương pháp khai thác, tính toán trên dữ liệu truyền thống sẽ gặp khó khăn và thiếu hiệu quả. Nếu như có thể áp dụng những mô hình tính toán mới trên dữ liệu này sẽ mang lại hiệu quả to lớn.

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

Tác giả liên hệ: Nguyễn Tú Trung,
Email: trungnt.sremis@gmail.com
Đến tòa soạn: 7/2021, chỉnh sửa: 10/2021, chấp nhận đăng: 11/2021.

II. CÁC NGHIÊN CỨU LIÊN QUAN

A. *Tổng quan về MapReduce*



Hình 1: Sơ đồ mô hình MapReduce [2].

MapReduce là mô hình xử lý tính toán song song và phân tán do google đề xuất (hình 1). Nó bao gồm hai chức năng cơ bản: "Map" và "Reduce" được xác định bởi người dùng [2]. Thông qua thư viện MapReduce ứng dụng với từng ngôn ngữ, chương trình có nhiệm vụ phân mảnh tệp dữ liệu đầu vào. Các máy gồm có: master và worker. Trong đó máy master làm nhiệm vụ điều phối sự hoạt động của quá trình thực hiện MapReduce trên các máy worker, các máy worker làm nhiệm vụ thực hiện Map và Reduce với dữ liệu mà nó nhận được. Dữ liệu được cấu trúc theo dạng key, value.

Biểu diễn hình thức mô hình MapReduce

Theo [11][12], ta có biểu diễn hình thức của mô hình MapReduce như sau:

$$\text{map: (K1 k1, V1 v1)} \rightarrow \text{list(K2 k2, V2 v2)} \quad (1)$$

$$\text{reduce: (K2 k2, list(V2 v2))} \rightarrow \text{list(K3 k3, V3 v3)} \quad (2)$$

Trong đó:

- K1, V1 là kiểu khóa và giá trị đầu vào của hàm map; k1, v1 là các đối tượng tương ứng có kiểu K1, V1
- K2, V2 là kiểu khóa và giá trị đầu ra của hàm map, cũng là kiểu khóa và giá trị đầu vào của hàm reduce; k2, v2 là các đối tượng tương ứng có kiểu K2, V2
- K3, V3 là kiểu khóa và giá trị đầu ra của hàm reduce; k3, v3 là các đối tượng tương ứng có kiểu K3, V3

Nói cách khác, ta thấy:

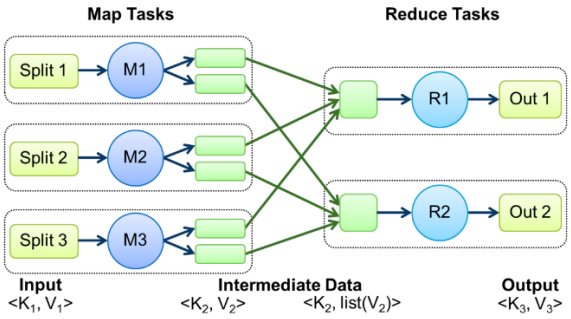
- Nếu xác định được k1, v1, k2, v2, ta có input, output của hàm map. Thông thường, với dữ liệu text, k1 là giá trị offset của dòng dữ liệu, v1 là nội dung dòng dữ liệu.
- Nếu xác định được k2, v2, k3, v3, ta có input, output của hàm reduce.

Biểu diễn hình thức có thể viết lại chỉ có k1, v1, k2, v2, k3, v3 như sau:

$$\text{map: (k1, v1)} \rightarrow \text{list(k2, v2)} \quad (3)$$

$$\text{reduce: (k2, list(v2))} \rightarrow \text{list(k3, v3)} \quad (4)$$

Hình 2 thể hiện sơ đồ quá trình thực thi job MapReduce và chuyển đổi dữ liệu từ kiểu (K1, V1) sang (K2, V2) và từ kiểu (K2, V2) sang (K2, V3).



Hình 2: Thực thi công việc MapReduce [12].

B. *Thuật toán Naïve Bayes*

Theo [9], 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,…}
- 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, như minh họa trong hình 3)
  - ✓ Tính xác suất $P(C_i)$
  - ✓ Tính xác suất $P(x_k|C_i)$
- Bước 2: Phân lớp X_{new}
  - ✓ Tính $F(X_{new}, C_i) = P(C_i) \prod_{k=1}^{n} P(x_k|C_i)$
  - ✓ X_{new} được gán vào lớp C_q sao cho

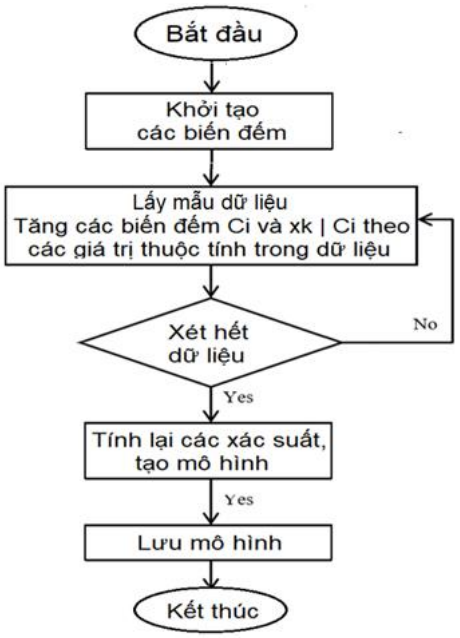
$$F(X_{new}, C_q) = \max (F(X_{new}, C_i)) \quad (5)$$

$P(x_i|C_i)$ được tính như sau:

$$P(x_k|C_i) = \frac{C_{i,D}\{x_k\}}{|C_{i,D}|} \quad (6)$$

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 nhận giá trị là x_k



Hình 3: Lưu đồ thuật toán huấn luyện theo Bayes.

C. *Đánh giá học lực sử dụng Hồi quy tuyến tính*

Theo quy chế đánh giá xếp loại học lực [8], kết quả học lực của học sinh được tổng hợp, tính toán và đánh giá qua các bài kiểm tra.

❖ **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 tư 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.

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

$$ĐTB_{mhk} = \frac{TĐKT_{tx} + 2 \times TĐKT_{đk} + 3 \times ĐKT_{hk}}{Số\ bài\ KT_{tx} + 2 \times Số\ bài\ KT_{đk} + 3} \quad (1)$$

Trong đó:

- ✓ TĐKTtx: Tổng điểm của các bài KTtx.
- ✓ TĐKTđk: Tổng điểm của các bài KTđk.
- ✓ ĐKThk: Điểm bài KT_{hk}.

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

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

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

❖ ***Tiêu chuẩn xếp loại hoc 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.

D. *Dự báo học lực dựa trên phân lớp Bayes*

Trong [14], chúng tôi đã đề xuất thuật toán dự báo học lực của học sinh phổ thông dựa trên phương pháp Bayes.

❖ ***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, GDCD, KTNN, Thể dục, GDQP như hình 1.

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

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

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

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

- Phương án 1: Gồm 3 kĩ thuật, không sử dụng trực tiếp điểm môn học làm dữ liệu thành phần của vector $\vec{x}$:
  - ✓ Kỹ thuật 1: Phân điểm thành G, K, TB, Y, K.
  - ✓ 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 trơn Laplace như sau:

$$P(x_i|C_i) = \frac{C_{i,D}\{x_k\}+1}{|C_{i,D}|+r} \quad (5)$$

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

Luật quyết định học lực dựa trên Bayes bao gồm: Luật quyết định loại Giỏi, Luật quyết định loại Khá, Luật quyết định loại Trung bình, Luật quyết định loại Kém [9].

III. ĐỀ XUẤT CẢI TIẾN PHƯƠNG PHÁP DỰ BÁO HỌC LỰC HỌC SINH SỬ DỤNG MÔ HÌNH MAPREDUCE

A. *Phân tích thuật toán huấn luyện mô hình dự báo học lực Naïve Bayes*

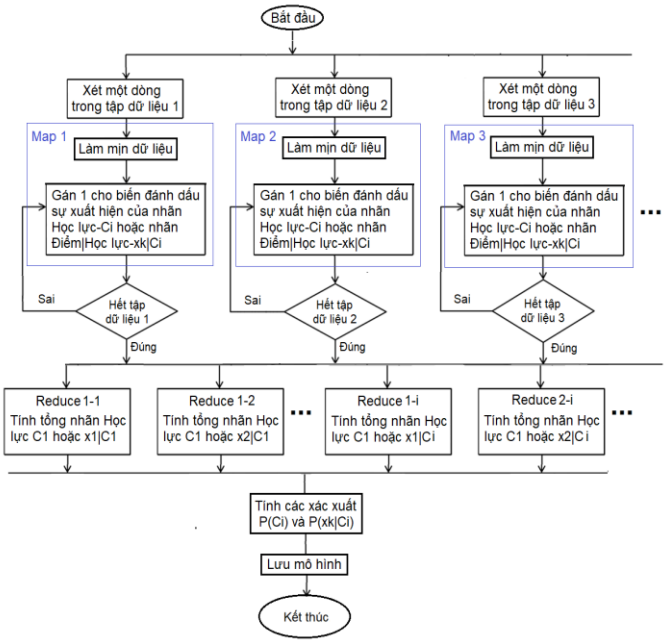
Ta có một số nhận xét như sau:

- Trong quá trình học (trainning), nếu số lượng dữ liệu quá lớn, sẽ phát sinh các vấn đề như thiếu bộ nhớ, tốn thời gian thực thi.
- Hầu hết thời gian huấn luyện Bayes là đếm số lần xuất hiện của các biến nhãn Học lực C_i hoặc các nhãn Điểm|Học lực: $x_k|C_i$.
- ❖ Việc tính các xác xuất, đếm các số lần xuất hiện của từng nhãn là độc lập => chia dữ liệu thành nhiều phần nhỏ và thực hiện song song.

B. *MapReduce hoá thuật toán huấn luyện mô hình dự báo học lực Naïve Bayes*

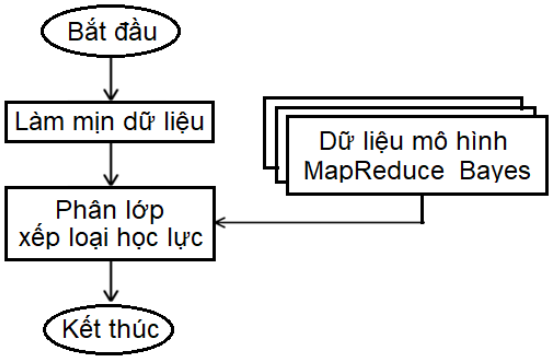
Ý tưởng:

- Dữ liệu huấn luyện là danh sách điểm-học lực của học sinh được chia thành nhiều phần nhỏ bởi hệ thống.
- Hàm Map: Tích lũy 1 cho mỗi lần xuất hiện nhãn Học lực C_i và nhãn Điểm|Học lực $x_k|C_i$ (với nhãn kèm theo)
- Hệ thống tự động gom nhóm các số 1 có cùng nhãn Học lực C_i hoặc nhãn Điểm|Học lực $x_k|C_i$
- Hàm Reduce: Tính tổng các số 1 theo từng nhãn Học lực và nhãn Điểm|Học lực
- Tính các xác suất của các nhãn Học lực $P(C_i)$ và nhãn Điểm|Học lực $P(x_k|C_i)$
- Lưu file mô hình gồm thông tin xác suất của các nhãn Học lực $P(C_i)$ và nhãn Điểm|Học lực $P(x_k|C_i)$



Hình 4: Lưu đồ thuật toán huấn luyện mô hình dự báo học lực dựa trên Bayes cải tiến sử dụng mô hình MapReduce.

Thuật toán dự báo học lực học sinh dựa trên luật quyết định Bayes và mô hình MapReduce gọi là DBHL_MapReduce_Bayes có các bước giống với thuật toán DBHL_Bayes chỉ thay file mô hình với kết quả từ thuật toán huấn luyện trong hình 8.



Hình 5: Lưu đồ thuật toán phân lớp học lực dựa trên MapReduce_Bayes.

C. *Biểu diễn hình thức cho hàm Map và Reduce trong mô hình dự báo học lực sử dụng MapReduce*

Input: Mỗi dòng dữ liệu row_i là một bộ gồm: danh sách điểm các môn học, kết quả học lực: $(list(x_k),C_i)$.

Output: Các cặp nhãn Học lực C_i và nhãn Điểm|Học lực $x_k|C_i$ và tổng số lần xuất hiện của nhãn tương ứng: $list(Or(C_i,x_k|C_i),count)$.

Khi này, ta xác định được các cặp k1, v1 và k3, v3 như sau:

- k1 là offset, v1 là nội dung dòng dữ liệu (list(x_k),C_i)
- k3 là nhãn Học lực C_i hoặc nhãn Điểm|Học lực $x_k|C_i$, v3 là tổng số lần xuất hiện tượng ứng của nhãn được lưu bởi k3

Hàm Map thực hiện việc tích lũy 1 cho mỗi lần xuất hiện nhãn Học lực C_i hoặc nhãn Điểm|Học lực $x_k|C_i$ (với nhãn kèm theo) nên suy ra k2, v2 như sau:

- k2 là nhãn Học lực C_i hoặc nhãn Điểm|Học lực $x_k|C_i$, v2 là 1.

Khi này, ta có biểu diễn hình thức của các thủ tục Map và Reduce như sau:

$$\text{map2D: (offset, row}_i) \rightarrow \text{list(Or}(C_i,x_k|C_i),1) \quad (10)$$

$$\text{reduce2D: (Or}(C_i,x_k|C_i), \text{list(1))} \rightarrow \text{list(Or}(C_i,x_k|C_i), \text{sum(list(1))} \quad (11)$$

D. *Thuật toán của thủ tục map_DBHL*

Bảng 2 mô tả thuật toán cho thủ tục map_DBHL. Nhiệm vụ thuật toán map_ DBHL là tách dữ liệu đầu vào thành từng nhãn Điểm từng môn học và nhãn Học lực. Tiếp theo, tích lũy 1 cho mỗi lần xuất hiện nhãn Học lực C_i hoặc nhãn Điểm|Học lực $x_k|C_i$ (với nhãn kèm theo).

Bảng 2: Thuật toán cho hàm map_DBHL

Input: key k1 là giá trị offset, value v1 là thông tin row_i: (list(x_k),C_i)
Output: Danh sách lstk2v2 các cặp (k2,v2): k2 là nhãn Học lực C_i hoặc Điểm|Học lực $x_k|C_i$, v2 là 1
B1: Tách các nhãn Điểm x_k và nhãn Học lực C_i từ v1
B2: Khởi tạo danh sách lstk2v2 để lưu các cặp (k2,v2)
B3: Thêm cặp (C_i,1) vào danh sách lstk2v2
B3: Duyệt các nhãn điểm x_k
B3.1: Thêm cặp (xk|Ci,1) vào danh sách lstk2v2

E. *Thuật toán của thủ tục reduce_DBHL*

Bảng 3 mô tả thuật toán cho thủ tục reduce_DBHL. Nhiệm vụ thuật toán reduce_DBHL là tính tổng các giá trị 1 ứng với nhãn đầu vào là Học lực C_i hoặc nhãn Điểm|Học lực $x_k|C_i$.

Bảng 3: Thuật toán cho hàm reduce_DBHL.

Input: key là Or(C_i,$x_k|C_i$), value là danh sách các số 1 ứng với nhãn được lưu trong key, tức là list(1)
Output: Cặp (k3,v3): k3 là Or(C_i,$x_k|C_i$), v3 là tổng các phần tử của list(1)
B1: Khởi tạo sum = 0
B2: Duyệt list(1)
B2.1: Tăng sum = sum + 1
B3: Gán k3 = Or(C_i,$x_k|C_i$)
B4: Gán v3 = sum

F. *Chứng minh độ chính xác dự báo hoc lực dựa trên Bayes và MapReduce_Bayes là giống nhau*

Từ các phần trên, ta có một số nhận xét như sau:

- Nhận xét 1: Điểm khác biệt giữa 2 thuật toán huấn luyện mô hình dự báo học lực dựa trên Bayes được trình bày trong hình 3 và hình 4 là:
  - ✓ Trong hình 3, việc tính số lần xuất hiện của các nhãn Học lực C_i hay nhãn Điểm|Học lực $x_k|C_i$ được thực hiện tuần tự.
  - ✓ Trong hình 4, việc tính số lần xuất hiện của các nhãn Học lực C_i hay nhãn Điểm|Học lực $x_k|C_i$ được thực hiện song song và phân tán sử dụng mô hình MapReduce.
- Nhận xét 2: Theo nhận xét 1, tổng số lần xuất hiện của các nhãn Học lực C_i hay nhãn Điểm|Học lực $x_k|C_i$ chỉ là được thực hiện theo 2 cách khác nhau trong 2 thuật toán nên sẽ có cùng kết quả.
- Nhận xét 3: Từ nhận xét 2 suy ra giá trị các xác xuất các nhãn Học lực $P(C_i)$ và nhãn Điểm|Học lực $P(x_k|C_i)$ là giống nhau với cả hai thuật toán.
- Nhận xét 4: Từ nhận xét 3, khi phân lớp, giá trị hàm $F(X_{new}, C_i)$ là giống nhau với cả hai thuật toán phân lớp. Như vậy, kết quả luật quyết định học lực là giống nhau với cả hai thuật toán DBHL_Bayes và DBHL_MapReduce_ Bayes. Đây là ***điều phải chứng minh***.

IV. THỬ NGHIỆM

Tập dữ liệu thử nghiệm là điểm số của học sinh và kết quả xếp loại học lực của một số trường cấp 3 được thu thập trên internet của một trường THPT ở Hà Nội (xin phép không chia sẻ vì lý do bảo mật). Dữ liệu huấn luyện được lưu trong the excel file, bao gồm 7962 bản ghi. Tập dữ liệu kiểm thử được lưu trong the excel file, bao gồm 1162 bản ghi.

Bảng 4 thống kê thời gian huấn luyện mô hình và độ chính xác dự báo theo từng phương án và kĩ thuật cụ thể với thuật toán DBHL_Bayes.

Bảng 4: Thuật toán DBHL_Bayes.

Phương án/ Kĩ thuật	Phương án 1			Phương án 2
	KT 1	KT 2	KT 3	
Thời gian huấn luyện	188 ms	332 ms	322 ms	285 ms
Độ chính xác test	99.14 %	100%	100%	99.48 %

Bảng 5 thống kê thời gian huấn luyện mô hình và độ chính xác dự báo theo từng phương án và kĩ thuật cụ thể với thuật toán DBHL_MapReduce_Bayes.

Bảng 5: Thuật toán DBHL_MapReduce_ Bayes.

Phương án/ Kĩ thuật	Phương án 1			Phương án 2
	KT 1	KT 2	KT 3	
Thời gian huấn luyện	100 ms	119 ms	111 ms	112 ms

Độ chính xác test	99.14 %	100%	100%	99.48%

Từ các kết quả trong bảng 4 và 5, ta thấy:

- Về thời gian huấn luyện:
  - ✓ Thời gian huấn luyện của Phương án 1-Kĩ thuật 1 là nhỏ nhất còn Phương án 1-Kĩ thuật 2 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 huấn luyện lại để tăng cường độ chính xác trong trường hợp quy mô dữ liệu huấn luyện bị thay đổi.
  - ✓ Thời gian huấn luyện của thuật toán DBHL_MapReduce_Bayes là nhỏ hơn rất nhiều so với thuật toán DBHL_Bayes. Điều này thể hiện sự ưu việt của mô hình xử lý song song và phân tán MapReduce.
- Về độ chính xác:
  - ✓ Độ chính xác trên dữ liệu test của Phương án 1-Kĩ thuật 1 là nhỏ nhất với 99.14% còn Phương án 1-Kĩ thuật 2 và 3 là lớn nhất với 100% độ chính xác. Điều này cho thấy việc sử dụng phương pháp học máy Bayes là rất phù hợp cho vấn đề dự báo học lực.
  - ✓ Độ chính xác dự báo sử dụng thuật toán Bayes và MapReduce_Bayes là giống nhau.

V. 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 và một cải tiến của thuật toán này sử dụng mô hình MapReduce nhằm tăng tốc độ thực thi của thuật toán. Trong đó, chúng tôi cũng đề xuất các kĩ thuật làm mịn dữ liệu thô (giá trị điểm ban đầu) trước khi thực hiện huấn luyện hay phân lớp Bayes. Kết quả thử nghiệm cho thấy tốc độ huấn luyện rất nhanh và độ chính xác rất cao, đều vượt 99% với cả 4 phương án-kĩ thuật trương ứng. Đặc biệt, thoạt toán DBHL_MapReduce_Bayes cho tốc độ thực thi nhanh hơn rất nhiều so với DBHL_Bayes mà không làm suy giảm độ chính xác so với thuật toán DBHL_Bayes.

Trong nghiên cứu tiếp theo, nhóm tác giả dự định tiếp tục áp dụng mô hình MapReduce cho những thuật toán khác để tăng cường hiệu quả trong các thuật toán trong bối cảnh dữ liệu ngày càng lớn và phức tạp.

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] **Jeffrey Dean and Sanjay Ghemawat,** *MapReduce: Simplified Data Processing on Large Clusters*, USENIX Association OSDI '04: 6th Symposium on Operating Systems Design and Implementation, 2004.

[3] Jialin ma, Yongjun zhang, Jinling liu, *Intelligent SMS spam filtering using topic model*, ieee international conference on intelligent networking and collaborative systems (incos), 2016.

[4] 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.

[5] Phan Hữu Tiếp, Vũ Đức Lung, Cao Nguyễn Thủy Tiên, Lâm Thành Hiền, *Phương pháp lọc thư rác tiếng việt dựa trện 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.

[6] Tianda Yang, Kai Qian, Dan Chia-Tien Lo, *Spam filtering using Association Rules and Naïve Bayes Classifier*, IEEE International Conference on Progress in Informatics and Computing (PIC), 2015.

[7] http://viet.jnlp.org/kien-thuc-co-ban-ve-xu-ly-ngon-ngu-tu-nhien/machine-learning-trong-nlp/phan-loai-van-ban-bang-dinh-ly-bayes

[8] 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

[9] https://towardsdatascience.com/introduction-to-naive-bayes-classification-4cffabb1ae54

[10] Nandhini.P, *A Research on Big Data Analytics Security and Privacy in Cloud, Data Mining, Hadoop and Mapreduce*, Int. Journal of Engineering Research and Application, 2018.

[11] Herodotos Herodotou, *Business Intelligence and Analytics: Big Systems for Big Data*, Cyprus University of Technology, 2016.

[12] Tom White**,** *Hadoop: The Definitive Guide : The Definitive Guide*, 2009.

[13] Piotr S. and Henryk M., *Deep learning methods for Subject Text Classification of Articles*, Proceedings of the Federated Conference on Computer Science and Information Systems, 2017.

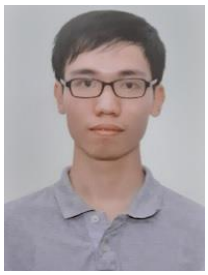
[14] Đào Đức Anh, Nguyễn Tu Trung, Vũ Văn Thỏa, *Ứng dụng thuật toán Bayes trong vấn đề dự báo học lực của học sinh phổ thông*, số 1, năm 2020, 46-49.

USING BAYESIAN CLASSIFICATION AND MAPREDUCE MODEL 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 and MapReduce model in predicting student performance to support the management and assessment of students in high school.

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

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



Đào Đức Anh, sinh viên vừa tốt nghiệp trường Đại học Thủy Lợi. Bắt đầu nghiên cứu về học máy...

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 thông tin địa lý, mạng viễn thông