Tạp chí Khoa học Trường Đại học Cần Thơ
website: sj.ctu.edu.vn

PHÂN LOẠI VĂN BẢN: MÔ HÌNH TÚI TỪ VÀ TẬP HỢP MÔ HÌNH MÁY HỌC TỰ ĐỘNG

Đỗ Thanh Nghị[1] và Phạm Nguyên Khang[1]
[1] Khoa Công nghệ Thông tin và Truyền thông, Trường Đại học Cần Thơ

Thông tin chung:
Ngày nhận: 17/04/2013
Ngày chấp nhận: 29/10/2013

Title:
Text classification: Bag-of-words and ensemble-based learning methods

Từ khóa:
Phân loại văn bản, Mô hình túi từ, Phương pháp tập hợp mô hình máy học, Phân loại có giám sát

Keywords:
Text classification, Bag-of-Words, Ensemble-based Learning Model, Supervised Classification

ABSTRACT

This paper presents an approach to classify text documents usingthe Bag-of-Word (BoW) model and ensemble-based learning algorithms. The ensemble-based learning algorithms include random multinomial naive Bayes (rMNB) and random oblique decision stump (rODS) models. The bag-of-word model is used to look for the sparse vectors of occurrence counts of words in text documents. The pre-processing step using the bag-of-word model brings out a dataset with a very large number of dimensions. Thus, we propose the new algorithms, called boosting of random multinomial naive Bayes and oblique decision stump models,whichare usually suited for classifying very-high-dimensional datasets. The results of the experiment on a real dataset show that our proposed algorithms have a high performance compared with other algorithms. The new approach has achieved an accuracy of 94.8%.

TÓM TẮT

Trong bài này, chúng tôi giới thiệu tiếp cận phân lớp văn bản với độ chính xác cao. Nghiên cứu của chúng tôi dựa trên sự kết hợp giữa phương pháp biểu diễn văn bản bằng mô hình túi từ và các giải thuật xây dựng tập hợp các mô hình học tự động như Bayes thơ ngây ngẫu nhiên (random multinomial naive Bayes (rMNB)), cây xiên phân ngẫu nhiên đơn giản (random oblique decision stump (rODS)). Bước tiền xử lý, bao gồm phân tích từ vựng, xây dựng mô hình túi từ để biểu diễn văn bản dưới dạng véc tơ tần số xuất hiện của từ trong văn bản, số chiều rất lớn. Chúng tôi đề xuất các giải thuật boosting mới dựa trên mô hình cơ bản như cây ngẫu nhiên xiên phân đơn giản (rODS), Bayes thơ ngây ngẫu nhiên (rMNB), cho phép phân lớp hiệu quả tập dữ liệu này. Kết quả thực nghiệm với tập dữ liệu thực cho thấy rằng phương pháp của chúng tôi đề xuất phân lớp rất hiệu quả khi so sánh với các giải thuật hiện có, đạt được chính xác 94.8%.

1 GIỚI THIỆU

Trong kỷ nguyên công nghệ thông tin, chúng ta nhận ngày càng nhiều nguồn thông tin dưới dạng văn bản. Nguồn thông tin này đến từ các thư viện điện tử, thư điện tử, trang web. Việc khám phá tri thức tiềm ẩn từ kho dữ liệu văn bản là cần thiết cho việc quản lý, khai thác triệt để nguồn thông tin văn bản khổng lồ này. Các tri thức có thể là mô hình gom cụm hay phân lớp văn bản, mà ở đó mô hình phân lớp được sử dụng phổ biến trong ứng dụng như: gán nhãn tự động một bản tin, phân tích nội dung để phát hiện nhóm khủng bố, nhận dạng thư rác. Phân lớp tự động văn bản có thể được mô tả ngắn gọn như sau. Phân loại văn bản là gán nhãn cho từng văn bản theo chủ đề đã được định nghĩa trước dựa vào nội dung của văn bản.

Phân lớp văn bản thường được dựa trên mô hình ngữ nghĩa hoặc máy học. Tuy nhiên như bài phỏng vấn được thực hiện bởi M. Lucas (Tạp chí Mappa Mundi) năm 1999, M. Hearst (Giáo sư đầu ngành về phân tích dữ liệu của Đại học California, Berkeley) cho rằng tiếp cận ngữ nghĩa là vấn đề rất khó, phức tạp. Thay vì vậy, tiếp cận dựa trên máy học tự động lại đơn giản và cho nhiều kết quả tốt trong thực tiễn. Hầu hết các phương pháp phân loại văn bản dựa trên mô hình thống kê từ và các giải thuật học tự động Theo (Sebastiani, 99). Theo mô hình túi từ, dữ liệu văn bản không có cấu trúc (độ dài khác nhau) được biểu diễn dưới dạng véc tơ tần số xuất hiện của từ trong văn bản. Tập từ vựng của chúng ta có thể lên đến hàng chục ngàn. Tập các dữ liệu văn bản được chuyển về dạng một bảng có số cột (chiều, từ vựng) rất lớn. Bước tiếp theo là huấn luyện mô hình học tự động từ bảng dữ liệu này. Các mô hình máy học thường sử dụng như giải thuật k láng giềng (kNN (Fix & Hodges, 52)), Bayes thơ ngây (NB (Good, 65)), cây quyết định (Quinlan, 93), (Breiman et al., 84), máy học véc tơ hỗ trợ (SVM (Vapnik, 95)), giải thuật tập hợp mô hình bao gồm Boosting (Freund & Schapire, 95), (Breiman, 98) và rừng ngẫu nhiên (Breiman, 01). Do dữ liệu có số chiều lớn, chỉ có máy học SVM và phương pháp tập hợp mô hình xử lý hiệu quả.

Trong bài báo này, chúng tôi đề xuất giải thuật học boosting của Bayes thơ ngây ngẫu nhiên (rMNB) và cây xiên phân ngẫu nhiên đơn giản (rODS) cho phân lớp hiệu quả dữ liệu có số chiều lớn thu được từ biểu diễn văn bản với mô hình túi từ. Giải thuật boosting để xây dựng tuần tự k mô hình cơ sở rMNB hay rODS, mỗi mô hình tập trung hầu hết các lỗi được tạo ra bởi các mô hình trước đó. Ngoài ra, chúng tôi đề nghị sử dụng các tập con chiều ngẫu nhiên khi xây dựng các bộ phân lớp cơ sở (rMNB, rODS), ý tưởng này nhằm tăng khả năng chịu đựng nhiễu (số chiều lớn, mỗi chiều chỉ chứa đựng một lượng nhỏ thông tin cho phân lớp, đây là trường hợp biểu diễn văn bản bằng mô hình túi từ). Vì vậy, giải thuật boosting của chúng tôi có thể xử lý hiệu quả tập dữ liệu với số chiều lớn. Chúng tôi làm thực nghiệm trên tập dữ liệu văn bản thu thập bởi (Trần & Phạm, 12), gồm 10 chủ đề văn bản của trang báo điện tử vnexpress.net. Kết quả cho thấy rằng phương pháp của chúng tôi đề xuất phân lớp rất hiệu quả khi so sánh với các giải thuật hiện có, đạt được chính xác 94.8%.

Phần tiếp theo của bài viết được trình bày như sau: phần 2 trình bày ngắn gọn về biểu diễn văn bản bằng mô hình túi từ; phần 3 trình bày giải thuật boosting của rMNB, rODS; phần 4 trình bày các kết quả thực nghiệm tiếp theo sau đó là kết luận và hướng phát triển.

2 BIỂU DIỄN VĂN BẢN BẰNG MÔ HÌNH TÚI TỪ

Theo tiếp cận phân lớp tự động văn bản bằng mô hình máy học (Sebastiani, 99), việc phân loại văn bản bao gồm hai bước chính: biểu diễn dữ liệu văn bản, huấn luyện mô hình phân lớp. Do dữ liệu văn bản ở đầu vào ở dạng không cấu trúc, trong khi các giải thuật máy học ở giai đoạn tiếp theo sau thường chỉ có thể xử lý được dữ liệu dạng cấu trúc bảng (mỗi dòng là một phần tử dữ liệu, cột là chiều hay thuộc tính). Để giải quyết vấn đề này, mô hình túi từ cho phép chúng ta biểu diễn tập dữ liệu văn bản về cấu trúc bảng.

Bước tiền xử lý này bao gồm việc phân tích từ vựng và tách các từ trong nội dung của tập văn bản, sau đó chọn tập hợp các từ có ý nghĩa quan trọng dùng để phân loại, biểu diễn dữ liệu văn bản về dạng bảng để từ đó các giải thuật máy học có thể học để phân loại. Ở bước phân tích từ vựng, công việc có thể là quy về từ gốc của các biến thể từ, có thể xóa bỏ các từ không có ý nghĩa cho việc phân lớp như các mạo từ, từ nối,... Tiếp đến là tách các từ, đưa vào tự điển. Một văn bản được biểu diễn dạng véc tơ (có n thành phần, chiều) mà giá trị thành phần thứ j là tần số xuất hiện từ thứ j trong văn bản. Nếu xét tập T gồm m văn bản và tự điển có n từ vựng, thì T có thể được biểu diễn thành bảng D kích thước m x n, dòng thứ i của bảng là véc tơ biểu diễn văn bản thứ i tương ứng. Xem ví dụ trong Bảng 1 và 2.

Bảng 1: Ví dụ về tập dữ liệu văn bản

STT	Nội dung	Chủ đề
1	Brazil - đối thủ khắc tinh của Italy	Thể thao
2	Mưa đá dữ dội, hàng trăm nhà dân bị thiệt hại	Xã hội
…	…	…
M	Đột nhập nhà đại gia trộm 2 kg vàng	Pháp luật

Bảng 2: Biểu diễn tập dữ liệu văn bản bằng mô hình túi từ

STT	1 (bị)	2 (brazil)	…	n (tinh)	Chủ đề
1	0	1	…	1	Thể thao
2	1	0	…	0	Xã hội
…	…	…	…	…	…
m	0	0	…	0	Pháp luật

Chúng ta có thể thấy rằng, khi tập dữ liệu vài trăm văn bản, tự điển có thể lên đến khoảng vài chục ngàn từ. Do đó bảng D có số cột n rất lớn.

Trong khi các mô hình máy học như k láng giềng (kNN), Bayes thơ ngây (NB) hay cây quyết định xử lý kém hiệu quả. Để khắc phục, người ta thường thực hiện việc rút gọn chiều dữ liệu. Phương pháp rút gọn có thể là lựa chọn những từ quan trọng nhất để có thể phân biệt văn bản này với văn bản khác, hay phương pháp giảm chiều. Các phương pháp để lựa chọn các từ có thể dựa vào ngưỡng tần số xuất hiện, độ lợi thông tin (information gain), thông tin tương quan (mutual information). Bước rút gọn này thường gây mất thông tin, làm giảm độ chính xác của bộ phân lớp sau này. Tuy nhiên, nếu không thực hiện bước rút gọn chiều, chúng ta cần xây dựng giải thuật máy có thể xử lý được bảng có số chiều lớn. Thường thì các mô hình máy học SVM và phương pháp tập hợp mô hình xử lý hiệu quả trên dữ liệu có số chiều lớn.

Chúng tôi đề xuất giải thuật học boosting của Bayes thơ ngây ngẫu nhiên (rMNB) và cây xiên phân ngẫu nhiên đơn giản (rODS) cho phân lớp hiệu quả dữ liệu có số chiều lớn thu được từ biểu diễn văn bản với mô hình túi từ.

3 GIẢI THUẬT BOOSTING CỦA RMNB VÀ RODS

Tập dữ liệu văn bản được biểu diễn theo mô hình túi từ. Khi không qua bất kỳ xử lý đặc biệt nào cho việc rút gọn chiều, bảng dữ liệu thu được có số chiều lên đến vài chục ngàn, mỗi chiều chỉ chứa đựng một lượng nhỏ thông tin cho phân lớp, tập dữ liệu được xem là nhiễu. Dựa theo đề xuất của (Breiman, 01), chúng tôi xây dựng giải thuật Bayes thơ ngây ngẫu nhiên (rMNB) và cây xiên phân ngẫu nhiên đơn giản (rODS). Thay vì giải thuật MNB và ODS sử dụng toàn bộ tập các thuộc tính (chiều) để huấn luyện mô hình phân lớp thì rMNB và rODS chỉ sử dụng tập con các thuộc tính được lấy ngẫu nhiên từ tập thuộc tính ban đầu.

3.1 Giải thuật Bayes thơ ngây ngẫu nhiên (rMNB)

Phương pháp ước lượng xác suất khi phân lớp một văn bản của mô hình MNB (Lewis & Gale, 94) được trình bày tóm tắt như sau. Giả sử C là tập hợp các lớp của văn bản. Tập các từ vựng của văn bản có kích thước là N. Khi có một văn bản mới đến là ti thì mô hình MNB gán lớp cho ti sao cho ước lượng xác suất để ti thuộc vào một lớp ci là lớn nhất hay là tìm giá trị lớn nhất của Pr(c|ti). Ước lượng xác suất Pr(c|ti) được tính như sau:

$$\Pr(c|t_i)=\frac{\Pr(c)\Pr(t_i|c)}{\Pr(t_i)} \qquad c\in C \qquad (1)$$

Trong công thức (1), xác suất *Pr(c)* được tính bằng tổng số văn bản của lớp *c* chia cho tổng số văn bản của tất cả các lớp. Trong tính toán tìm giá trị lớn nhất của $Pr(c|t_i)$, người ta có thể bỏ qua $Pr(t_i)$ do nó không đổi khi ước lượng xác suất của từng lớp.

Xác suất $Pr(t_i|c)$ được tính bằng công thức (2) như sau:

$$\Pr(t_i|c)=(\sum_n f_{ni})!\prod_n \frac{\Pr(w_n|c)^{f_{ni}}}{f_{ni}!} \qquad (2)$$

Trong công thức (2), f_{ni} là tần suất từ thứ *n* trong t_i và $Pr(w_n|c)$ là xác suất của từ thứ *n* khi cho trước lớp *c*. $Pr(w_n|c)$ có thể được ước lượng bằng cách lấy tần suất từ thứ *n* trong tất cả các văn bản của lớp *c* chia cho tổng số tần suất của các từ vựng trong các văn bản của lớp *c*. Hơn nữa, $(\sum_n f_{ni})!$ và $\prod_n f_{ni}!$ trong công thức (2) có thể thay bằng hằng số chuẩn hóa α mà không làm thay đổi kết quả. Việc ước lượng xác suất $Pr(t_i|c)$ của công thức (2) được tính bằng công thức (3) như sau:

$$\Pr(t_i|c)=\alpha\prod_n \Pr(w_n|c)^{f_{ni}} \qquad (3)$$

Khác với thực hiện ước lượng xác suất $Pr(t_i|c)$ trong công thức (3) của MNB, giải thuật rMNB tính bằng công thức (4), tương tự như (3) nhưng thay thế *n* từ vựng bởi *n'* từ vựng lấy ngẫu nhiên từ *n* từ vựng.

$$\Pr(t_i|c)=\alpha\prod_{n'} \Pr(w_{n'}|c)^{f_{n'i}} \qquad (4)$$

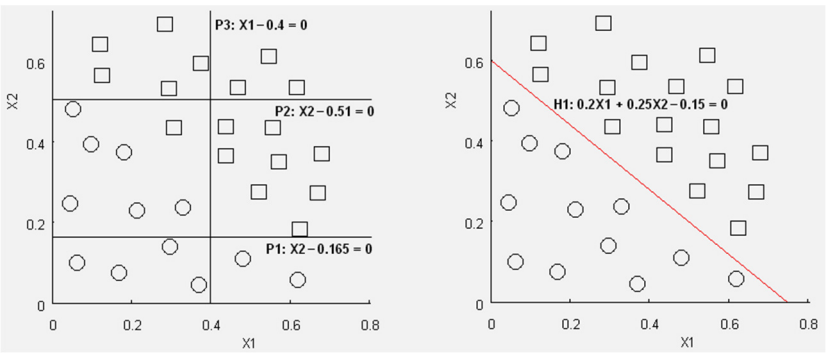
3.2 Giải thuật cây xiên phân ngẫu nhiên (rODS)

Mô hình cây quyết định có cấu trúc dạng cây mà ở đó nút lá được gán nhãn tương ứng với lớp của dữ liệu và nút trong được tích hợp với điều kiện kiểm tra để rẽ nhánh. Có hai giải thuật học tự động là CART (Breiman *et al*., 84) và C4.5 (Quinlan, 93).

Mô hình cây quyết định đơn giản (decision stump) được đề xuất trong (Freund & Schapire, 95) là cây có số nút lá bằng với số lớp của dữ liệu. Với vấn đề phân lớp nhị phân (dữ liệu có 2 lớp dương và âm), thì cây quyết định đơn giản chỉ có 1 nút gốc và 2 nút lá (tương ứng với 2 nhãn hay lớp dự đoán của dữ liệu). Quá trình xây dựng cây quyết

định đơn giản của giải thuật học chỉ chọn một thuộc tính tốt nhất cho việc phân hoạch dữ liệu tại nút gốc tạo thành 2 nút lá (mỗi nút tương ứng một lớp).

Chúng ta có thể thấy rằng xây dựng cây đơn giản rất nhanh vì chỉ sử dụng duy nhất 1 thuộc tính để phân hoạch và kết thúc ngay. Do đó, độ chính xác của mô hình cây quyết định đơn giản bị giảm khi làm việc với các tập dữ liệu có số chiều lớn, mỗi chiều cung cấp ít thông tin cho phân lớp và các chiều còn phụ thuộc lẫn nhau, chẳng hạn như dữ liệu văn bản thu được mà chúng ta xử lý ở đây. Một ví dụ trong Hình 1, bất kỳ việc phân hoạch đơn thuộc tính nào (song song với trục tọa độ) đều không thể tách dữ liệu một lần duy nhất thành hai lớp một cách hoàn toàn mà phải thực hiện nhiều lần phân hoạch, nhưng việc phân hoạch đa chiều (xiên phân, kết hợp 2 thuộc tính) có thể thực hiện một cách hoàn hảo với duy nhất một lần. Tức là, cây quyết định đơn giản không hiệu quả bằng cây quyết định xiên phân đơn giản.



Hình 1: Phân hoạch đơn thuộc tính (trái), phân hoạch đa thuộc tính (phải)

Để khắc phục nhược điểm trên, nhiều giải thuật xây dựng cây quyết định sử dụng phân hoạch đa thuộc tính (xiên phân) tại các nút được đề nghị. Vấn đề xây dựng cây quyết định xiên tối ưu đã được biết như là một vấn đề có độ phức tạp NP-hard. Nghiên cứu tiên phong của Murthy và các cộng sự trong (Murthy et al., 93) đã đưa ra giải thuật OC1, một hệ thống dùng để xây dựng các cây quyết định xiên trong đó dùng thuật toán leo đồi để tìm một phân hoạch xiên tốt dưới dạng một siêu phẳng. Rừng ngẫu nhiên xiên phân RF-ODT của (Do et al., 09) xây dựng các cây xiên phân ngẫu nhiên dựa trên siêu phẳng tối ưu (phân hoạch hiệu quả cao, khả năng chịu đựng nhiễu tốt) thu được từ huấn luyện SVM (Vapnik, 95).

Để giải quyết 2 vấn đề chính là độ phức tạp và hiệu quả của bộ phân lớp yếu của kỹ thuật boosting, chúng tôi đề xuất chỉ xây dựng cây ngẫu nhiên xiên phân đơn giản (rODS). Giải thuật rODS xây dựng cây như mô tả trong Hình 2 cho vấn đề phân lớp nhị phân (2 lớp dương và âm). Cây xiên phân 3 nút, bắt đầu với toàn bộ dữ liệu nằm ở nút gốc, chọn ngẫu nhiên n' thuộc tính từ tập n thuộc tính ban đầu của dữ liệu là tìm ra siêu phẳng tối ưu n' chiều (SVM) để phân hoạch dữ liệu.

Siêu phẳng cần xác định có dạng:

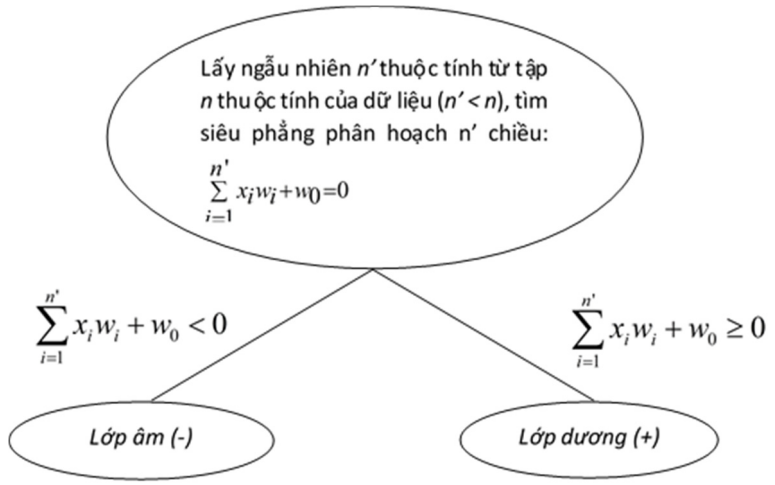
$$\sum_{i=1}^{n'} x_i w_i + w_0 = 0$$

Trong đó x_i là thuộc tính thứ i (chiều) của dữ liệu, w_i là trọng số véctơ pháp tuyến của siêu phẳng, w_0 là độ lệch của siêu phẳng.

Dựa vào dấu của biểu thức $\sum_{i=1}^{n'} x_i w_i + w_0$ mà dữ liệu sẽ được phân hoạch qua trái hay qua phải để dự báo nhãn.

Cây xiên phân ngẫu nhiên đơn giản có thể làm việc hiệu quả trên tập dữ liệu có số chiều lớn do nó đảm bảo được 2 yếu tố cơ bản là thời gian xây dựng nhanh và hiệu quả phân lớp cao. Do đơn giản chỉ có 3 nút, việc xây dựng cây xiên phân ngẫu nhiên rất nhanh khi chỉ tìm một siêu phẳng tối ưu trong không gian n' chiều ($n' < n$). Việc kết hợp nhiều thuộc tính để tạo phân hoạch xiên phân giúp phân lớp hiệu quả dữ liệu có số chiều lớn.

So với mô hình MNB, ODS với tập đầy đủ các thuộc tính thì giải thuật rMNB, rODS đơn giản, nhanh hơn, hiệu quả phân lớp tốt hơn do khả năng chịu đựng nhiễu cao hơn. Mặc dù mô hình đơn của rMNB và rODS thì không mạnh do quá đơn giản, nhưng khi áp dụng kỹ thuật boosting (Freund & Schapire, 95), (Breiman, 98) để xây dựng tập hợp các mô hình rMNB, rODS thì hiệu quả của giải thuật được cải thiện rất nhiều.



Hình 2: Cây ngẫu nhiên xiên phân đơn giản

3.3 Giải thuật boosting của rMNB, rODS

Breiman đã nghiên cứu phân tích hiệu quả giải thuật học dựa trên cơ sở của hai thành phần lỗi là bias và variance mà ở đó, thành phần lỗi bias là lỗi của mô hình học và variance là lỗi do tính biến thiên của mô hình so với tính ngẫu nhiên của các mẫu dữ liệu học (Breiman, 01). Trong nghiên cứu kết hợp nhiều mô hình phân loại yếu thành tập hợp các mô hình phân loại để cho tính chính xác cao hơn so với chỉ một mô hình đơn.

Boosting, AdaBoost (Freund & Schapire, 95), ArcX4 (Breiman, 98) là kỹ thuật áp dụng một tập các bộ phân lớp yếu (weak learner) để nâng cao hiệu quả của các bộ phân lớp này bằng cách giảm bias và variance. Giải thuật ArcX4 cho kết quả tương tự như AdaBoost nhưng đơn giản và dễ cài đặt.

Ý tưởng chính của giải thuật ArcX4 (như mô tả trong giải thuật 1) lặp lại quá trình học của một bộ phân lớp yếu nhiều lần. Sau mỗi bước lặp, bộ phân lớp yếu (ví dụ như: Bayes thơ ngây ngẫu nhiên rMNB hay cây xiên phân ngẫu nhiên đơn rODS) sẽ tập trung học trên các phần tử bị phân lớp sai trong các lần trước. Để làm được điều này, cần gán cho mỗi phần tử một trọng số. Khởi tạo, trọng số của các phần tử bằng nhau trong lần lặp đầu tiên. Sau mỗi bước học, các trọng số này sẽ được cập nhật lại (tăng trọng số cho các phần tử bị phân lớp sai). Ở bước thứ i, ta lấy tập mẫu S_i trên tập dữ liệu và xây dựng mô hình h_i từ tập mẫu S_i. Lặp lại quá trình này sau T bước, ta sẽ được T mô hình cơ sở, kết hợp các mô hình cơ sở này lại ta sẽ có được một bộ phân lớp mạnh.

ArcX4 của khiêuMNB, rODS rất dễ cài đặt, đơn giản, nhanh hơn, hiệu quả phân lớp tốt do khả năng chịu đựng nhiễu cao hơn.

Giải thuật 1: ArcX4 của rMNB, rODS

Đầu vào:

- m phần tử dữ liệu : $\{(x_i, y_i)\}i=1,m$ với $x_i \in R^n$ và $y_i \in R^1$
- số bước lặp T

Huấn luyện:

- khởi động trọng số của m phần tử dữ liệu $Dist_1(j)$
  - cho $j = 1$ tới m thực hiện
    - $Dist_1(j) = 1/m$
- cho i = 1 tới T thực hiện (lặp T bước)
  - lấy mẫu S_i phần tử dựa trên trọng số $Dist_i$
  - học mô hình học cơ sở h_i từ tập mẫu S_i
    - $h_i = \{rMNB, rODS\}(S_i)$
  - tính lại lỗi dự đoán của từng phần tử x_j khi sử dụng các bộ phân lớp được xây dựng trước đó
    - $\varepsilon_i = \sum_{h_i(x_j) \neq y_j}^{i}$

4 KẾT QUẢ THỰC NGHIỆM

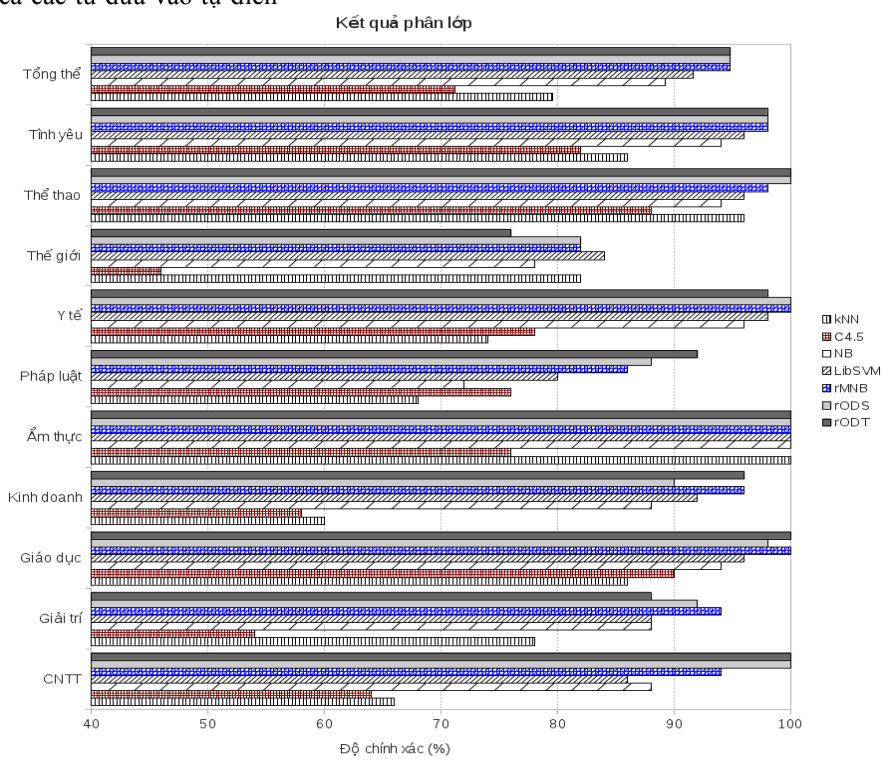
Để đánh giá hiệu quả của phương pháp đề xuất (mô hình túi từ và giải thuật boosting của rMNB, rODS) cho phân loại văn bản, chúng tôi đã tiến hành cài đặt giải thuật boosting của rMNB, rODS bằng C/C++. Chúng tôi muốn so sánh hiệu quả của giải thuật boosting của rMNB, rODS với các giải thuật học khác, bao gồm k láng giềng (kNN), Bayes thơ ngây (NB), máy học SVM, cây quyết định C4.5 và rừng ngẫu nhiên xiên (RF-ODT). Chúng tôi tiến hành cài đặt giải 2 giải thuật NB và kNN bằng ngôn ngữ lập trình C/C++. Giải thuật SVM chuẩn đã có trong các thư viện phần mềm miễn phí LibSVM (Chang & Lin, 01). Tất cả thực nghiệm được thực hiện trên PC (Intel Dual Core, 2.2 GHz, 2GB RAM), hệ điều hành LINUX (Mandriva 2010).

Chúng tôi sử dụng tập dữ liệu được sưu tập bởi (Trần & Phạm, 12). Đây là tập dữ liệu văn bản thu thập từ trang báo điện tử vnexpress.net, gồm có 10 chủ đề như công nghệ thông tin (cntt), giải trí, giáo dục, kinh doanh, ẩm thực, pháp luật, y tế, thế giới, thể thao, tình yêu. Mỗi chủ đề có 200 văn bản khác nhau tạo thành tập dữ liệu văn bản có 2000 bản tin. Chúng tôi chia tập dữ liệu ra thành 2 tập, một tập học có 1500 bản tin và tập kiểm thử có 500 bản tin. Các chủ đề có cùng số lượng bản tin trong cả tập học và kiểm thử. Giai đoạn tiền xử lý, chúng tôi phân tích và rút trích tất cả các từ đưa vào tự điển với số từ lên đến 12182. Chúng tôi không dùng bất kỳ xử lý đặc biệt nào khác như đã làm trong (Phạm et al., 06), (Trần & Phạm, 12). Chúng tôi thu được tập học là bảng có 1500 dòng (bản tin) và tập kiểm thử 500 dòng (bản tin), với 12182 cột (chiều, từ), trong 10 chủ đề (lớp).

Tập dữ liệu học dùng để huấn luyện mô hình phân lớp, bao gồm cả việc điều chỉnh các tham số cho các giải thuật học. Cuối cùng, kết quả kiểm thử thu được trên tập kiểm thử dùng để so sánh hiệu quả phân lớp.

Với các mô hình đơn, giải thuật Bayes thơ ngây (NB) và cây quyết định C4.5 không cần điều chỉnh tham số. Riêng với *k* láng giềng (kNN), chúng tôi thử tất cả các giá trị *k* từ 1 đến 10, kết quả vẫn không thay đổi. Nên chúng tôi báo cáo kết quả thực nghiệm của 1NN. Với máy học SVM, chúng tôi cố gắng sử dụng các hàm nhân (kernel function) của giải thuật SVM gồm hàm đa thức bậc *d*, Radial Basis Function (hàm nhân RBF), tuyến tính, cuối cùng kết quả thu được tốt như nhau. Chính lý do đó, chúng tôi huấn luyện SVM sử dụng hàm nhân tuyến tính cho nhanh.

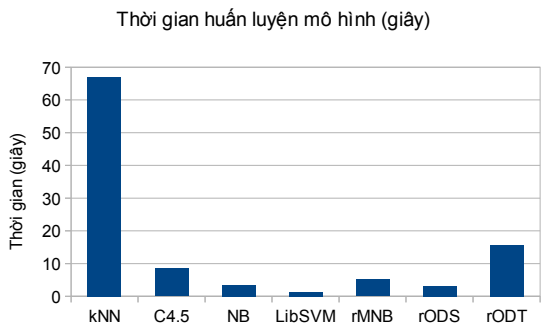
Với các phương pháp tập hợp mô hình như Boosting và ngẫu nhiên xiên (RF-ODT), chúng tôi đều xây dựng 50 mô hình cơ sở (rMNB, rODS, ODT sử dụng 1000 chiều ngẫu nhiên từ 12182 chiều).



Hình 3: Kết quả phân lớp trên tập dữ liệu văn bản 10 chủ đề

Kết quả thu được từ các giải thuật được trình bày trong Hình 3. Quan sát kết quả thu được, không có gì ngạc nhiên khi các mô hình học kNN, NB, C4.5 cho kết quả thấp khi so sánh với các các giải thuật khác. Điều này hoàn toàn phù hợp do dữ liệu có số chiều lớn, các mô hình đơn giản không còn phân lớp hiệu quả. Trong khi đó, giải thuật máy học SVM cho kết quả tốt hơn nhóm giải thuật đơn giản trước. Nhóm tập hợp mô hình, gồm 2 giải thuật boosting của rMNB, rODS và rừng ngẫu nhiên xiên RF-ODT cho kết quả phân lớp chính xác nhất.

Nếu quan sát thời gian cần thiết để huấn luyện mô hình học, mặc dù giải thuật kNN không có huấn luyện nhưng lại mất thời gian khi phân lớp lâu nhất. Kế đến là RF-ODT mặc dù nhanh hơn kNN đến 4 lần nhưng vẫn chậm hơn các giải thuật khác đến 4 hoặc 10 lần. Giải thuật SVM có thời gian huấn luyện nhanh, cho kết quả cũng rất khả quan. Hai giải thuật chúng tôi đề xuất là boosting của rMNB, rODS có thời gian huấn luyện nhanh và cho kết quả chính xác nhất.



Hình 4: Thời gian huấn luyện mô hình

Kết quả thu được từ thực nghiệm này cho phép chúng tôi tin rằng giải thuật đề xuất rMNB, rODS phân loại tốt dữ liệu văn bản, được biểu diễn theo mô hình túi từ (rất đơn giản, nhanh, không cần xử lý phức tạp nào).

5 KẾT LUẬN VÀ HƯỚNG PHÁT TRIỂN

Chúng tôi vừa trình bày một tiếp cận phân lớp văn bản với độ chính xác cao. Nghiên cứu của chúng tôi dựa trên sự kết hợp giữa phương pháp biểu diễn văn bản bằng mô hình túi từ và các giải thuật boosting, xây dựng tập hợp các mô hình học tự động như rMNB, rODS. Mô hình túi từ được xây dựng đơn giản, nhanh, để biểu diễn văn bản dưới dạng véc tơ tần số xuất hiện của từ trong văn bản, số chiều rất lớn. Thay vì cần các xử lý đặc thù để rút gọn chiều, chúng tôi đề xuất các giải thuật boosting mới dựa trên mô hình cơ bản ngẫu nhiên rMNB, rODS cho phép phân lớp hiệu quả tập dữ liệu này. Kết quả thực nghiệm với tập dữ liệu thực cho thấy rằng phương pháp của chúng tôi đề xuất phân lớp rất hiệu quả khi so sánh với các giải thuật hiện có, đạt được chính xác 94.8%.

Trong tương lai, chúng tôi dự định mở rộng giải thuật để xử lý vấn đề tương tự như phân lớp ảnh, video, sử dụng mô hình biểu diễn túi từ. Bên cạnh đó, chúng tôi cũng muốn tăng tốc quá trình xây dựng mô hình học của rMNB, rODS bằng việc xây dựng giải thuật song song.

TÀI LIỆU THAM KHẢO

1. Breiman, L.: Arcing classifiers. The annals of statistics 26(3), 801–849 (1998).
2. Breiman, L.: Random forests. Machine Learning 45(1), 5–32 (2001).
3. Chang, C.C., Lin, C.J.: LIBSVM – a library for support vector machines (2001). http://www.csie.ntu.edu.tw/~cjlin/libsvm
4. Do, T-N., Lenca, P., Lallich, S. and Pham, N-K.: Classifying Very-high-dimensional Data with Random Oblique Decision Trees. in *Advances in Knowledge Discovery and Management*, Springer-Verlag, pp. 39-55 (2009).
5. Fix, E and Hodges J.: Discriminatoiry Analysis: Small Sample Performance. Technical Report 21-49-004, USAF School of Aviation Medicine, Randolph Field, USA (1952).
6. Freund, Y., and Schapire, R.: A decision-theoretic generalization of on-line learning and an application to boosting. In: Computational Learning Theory: Proceedings of the Second EuropeanConference, pp. 23–37 (1995).
7. Good, I.: The Estimation of Probabilities: An Essay on Modern Bayesian Methods. *MIT Press* (1965).
8. Grove, A.J. and Schuurmans, D.: Boosting in the limit: Maximizing the margin of learned ensembles. In Proceedings of the Fifteenth National Conference on Artificial Intelligence (AAAI-98), pp. 692–699 (1998).
9. Lewis, D., Gale, W.: A sequential algorithm for training text classifiers. In: Proceedings of SIGIR (1994).

10. Phạm N.K., Đỗ T.N. và Poulet F.: Phân loại văn bản với BPSVM. Kỷ yếu hội nghị @CNTT, pp. 269-278 (2006).
11. Quinlan, J.R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann, San Mateo, CA (1993).
12. Sebastiani, F.: Machine learning in automated text categorization. *ACM Computing Surveys* 34(1), 1–47 (1999).
13. Trần, C.Đ và Phạm N.K.: Phân loại văn bản với máy học véc tơ hỗ trợ và cây quyết định. *Tạp chí Khoa học Trường Đại học Cần Thơ* số (21a):52-63 (2012).
14. Vapnik, V.: *The* Nature *of Statistical Learning Theory*. Springer-Verlag (1995).
15. Witten, I., Frank, E.: DataMining: *Practical Machine Learning Tools and Techniques*. Morgan Kaufmann (2005).