Phương pháp trích trọn thuộc tính hiệu quả cho dữ liệu có số chiều lớn

Hà Văn Sang[1]
[1]Học Viện Tài chính,

Tóm tắt

Một thách thức của bài toán phân lớp là số lượng thuộc tính thường rất lớn, việc phân lớp sao cho chính xác và hiệu quả hiện vẫn là một nghiên cứu thú vị cho các nhà khoa học trong lĩnh vực khoa học máy tính. Báo cáo đi sâu vào nghiên cứu giải thuật phân lớp thuộc tính random forest (RF). Đây là một giải thuật đã được nhiều nghiên cứu chứng minh là rất hiệu quả trong phân lớp thuộc tính đối với bộ dữ liệu có số lượng thuộc tính lớn. Trên cơ sở đó bài báo đề xuất một phương pháp học máy cho giải thuật phân lớp này nhằm tăng hiệu quả phân lớp của thuật toán. Cách tiếp cận này về cơ bản đã làm tăng khả năng phân lớp của giải thuật RF, phương pháp đề xuất còn cho thấy khả năng phân lớp tốt hơn một số phương pháp trích chọn đã được công bố. Như vậy, hướng cải tiến mà báo cáo đề xuất là có khả thi và thu được kết quả tương đối cao.

Giới thiệu

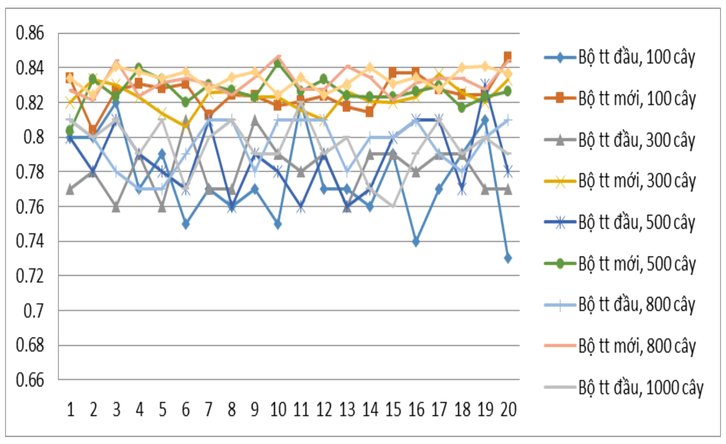
Trong xu hướng hội nhập quốc tế, thời đại thông tin bùng nổ, chúng ta đang "ngập lụt" trong dữ liệu nhưng lại "đói" về tri thức, cho nên một trong các vấn đề cấp thiết đó là làm sao phân tích và xử lý một khối lượng thông tin khổng lồ liên tục được cập nhật để đáp ứng các yêu cầu về phát triển mọi mặt văn hoá, kinh tế, chính trị, xã hội của đất nước. Hiện nay phần lớn các thuật toán phân lớp đã phát triển chỉ có thể giải quyết được một lượng số liệu giới hạn cũng như một độ phức tạp dữ liệu biết trước.. Các nghiên cứu cho thấy có rất nhiều hướng cải tiến các thuật toán phân lớp như áp dụng các thuật toán lai ghép (ensemble method), các thuật toán dựa vào phương pháp nhân (kernel-based method), hoặc áp dụng các phương pháp trích chọn đặc trưng (feature extraction/selection method). Với các phương pháp kể trên phương pháp trích chọn đặc trưng trở nên nổi trội và có một số ưu điểm phù hợp trong việc xử lý dữ liệu có số lượng thuộc tính lớn (vài nghìn đến vài trăm nghìn thuộc tính) nhưng đồng thời chỉ có một số lượng khá nhỏ các mẫu phân tích (vài chục hoặc vài trăm). Trong khai phá dữ liệu thì phương pháp trích chọn đóng một vai trò quan trọng để trích chọn và chuẩn bị dữ liệu. Hướng tiếp cận này làm tăng hiệu năng thu nhận tri thức trong các ngành như tin sinh, xử lý tiếng nói, hình ảnh,... Phương pháp này có ảnh hưởng ngay lập tức đến các ứng dụng như tăng tốc độ của các thuật toán khai phá dữ liệu, cải thiệu chất lượng dữ liệu và vì vậy tăng hiệu suất khai phá dữ liệu, kiểm soát được các kết quả của thuật toán.
Báo cáo này trình bày một đề xuất mới để dựa vào đó xây dựng mô hình trích chọn đặc trưng tối ưu giúp giảm kích cỡ của dữ liệu theo hướng chỉ giữ lại các thuộc tính đặc trưng, loại bỏ những thuộc tính không liên quan và những thuộc tính nhiễu nhằm tăng tốc độ các thuật toán phân lớp cải thiện chất lượng dữ liệu và vì vậy sẽ tăng hiệu suất của việc khai phá dữ liệu. Cụ thể, phương pháp đề xuất sẽ chọn ra những thuộc tính tốt nhất để làm tăng năng suất của thuật toán phân lớp Random Forest.

Kiến thức cơ sở

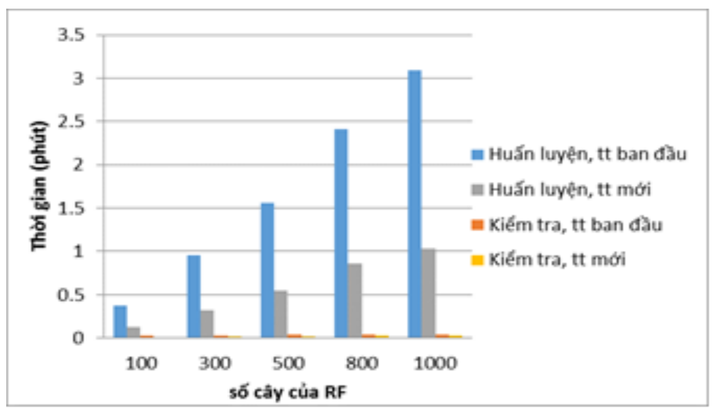
Trích chọn thuộc tính: là một bước cơ bản nhất trong việc tiền xử lý dữ liệu, nó làm giảm bớt số chiều của mẫu. Lựa chọn thuộc tính có thể là một phần vốn có của trích chọn thuộc tính ví dụ như phương pháp PCA hoặc thậm chí là một thiết kế xử lý thuật toán ví dụ như trong thiết kế cây quyết định. Tuy nhiên, lựa chọn thuộc tính thường là một bước cô lập riêng biệt trong một chuỗi xử lý
Lựa chọn thuộc tính có thể dựa vào các mô hình, các chiến lược tìm kiếm, thước đo chất lượng thuộc tính và ước lượng. Có ba loại mô hình như Filter, Wrapper, Embedded. Các chiến lược tìm kiếm bao gồm: forward, backward, floating, branch and bound, randomized
RF là một phương pháp phân lớp tốt do: (1) Trong RF các sai số (variance) được giảm thiểu do kết quả của RF được tổng hợp thông qua nhiều bộ học (learner), (2) Việc chọn ngẫu nhiên tại mỗi bước trong RF sẽ làm giảm mối tương quan (correlation) giữa các bộ học trong việc tổng hợp các kết quả

Kết quả

Môi trường thực nghiệm: Laptop với bộ xử lý Intel (R) Core i7 -2620 M CPU @ 2.70 GHz 2.69 GHz, RAM 4GB. Phương pháp học máy được thực hiện trên ngôn ngữ R là ngôn ngữ chuyên dùng trong xác suất thống kê, có địa chỉ www.r-project.org
Dữ liệu thực nghiệm: bộ dữ liệu mô tả về bệnh ung thư dạ dày gồm 137 bản ghi, 119 thuộc tính. Các bản ghi trong bộ dữ liệu được phân thành hai lớp ký hiệu là normal (bệnh nhân bình thường) và cancer (bệnh nhân bị ung thư).
Kết quả thực nhiệm:
Thực thi thuật toán RF trên bộ dữ liệu Stomach gốc 20 lần, mỗi lần chạy lại thực hiện kiểm chứng chéo 5 lần với số cây lần lượt là 100,300,500,800,1000 ta được kết quả như sau:

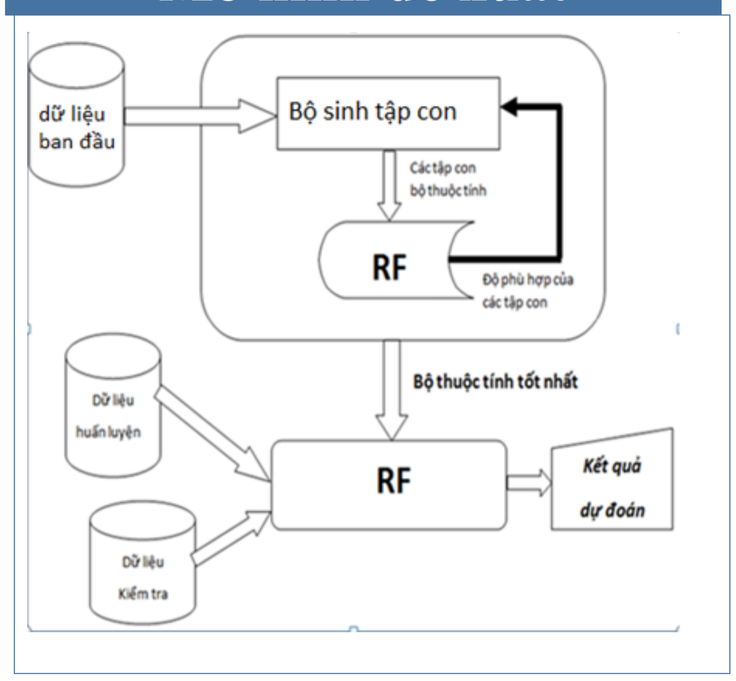


Hình 1. Biểu đồ so sánh kết quả chạy RF 20 lần trên bộ dữ liệu mới và bộ dữ liệu ban đầu với số cây bằng 100,300,500,800,1000



Hình 2. thời gian chạy trung bình của 20 lần chạy RF trên bộ dữ liệu mới và bộ dữ liệu ban đầu với số cây bằng 100,300,500,800,1000

Mô hình đề xuất



Hình 1. Mô hình đề xuất.

Thảo luận

Tỉ lệ đoán nhận của RF với bộ thuộc tính mới tăng lên rõ ràng, ước tính tăng khoảng 5%, thuật toán RF cho kết quả đoán nhận trung bình là 78% trên bộ dữ liệu ban đầu, còn RF chạy trên bộ dữ liệu sau khi lựa chọn các thuộc tính bằng thuật toán đề xuất cho kết quả đoán nhận trung bình là 83%
Tỉ lệ đoán nhận trên bộ thuộc tính mới tăng lên cho thấy bộ thuộc tính mới đã loại bỏ được một số thuộc tính nhiễu, thuộc tính dư thừa. Còn thời gian giảm đi là vì số lượng thuộc tính đã giảm xuống tương đối nhiều, cụ thể từ 119 thuộc tính ban đầu, sau khi lựa chọn bộ thuộc tính mới còn là 36 thuộc tính, như vậy số thuộc tính đã giảm khoảng 69% số thuộc tính ban đầu. Điều đó chứng tỏ phương pháp thực nghiệm mà báo cáo đưa ra cho hiệu quả tương đối tốt. Bộ thuộc tính mới đáp ứng được các mong muốn là nâng cao hiệu suất phân lớp và giảm thời gian học và thời gian kiểm thử. Đặc biệt, độ lệch chuẩn khi chạy RF trên bộ thuộc tính mới chỉ bằng 1/4 đến 1/3 độ lệch chuẩn khi chạy RF trên bộ

Kết luận

Báo cáo này đã tập trung nghiên cứu, tìm hiểu về thuật toán di truyền và Random Forest cùng với một số phương pháp tiền xử lý dữ liệu khác. Từ những tìm hiểu này, báo cáo đề xuất hướng cải tiến hiệu quả phân lớp của thuật toán RF theo phương pháp tìm ra bộ thuộc tính tối ưu nhỏ nhất từ một bộ thuộc tính rất lớn của dữ liệu ban đầu.
báo cáo đã tiến hành thực nghiệm để chứng minh tính đúng đắn của thuật toán. Thực nghiệm đã sử dụng bộ dữ liệu được lấy từ các công trình nghiên cứu trước đó là dữ liệu gen của các bệnh nhân bị ung thư dạ dày (Stomach).
Phương pháp đề xuất làm cho thuật toán phân lớp RF chạy nhanh hơn, ổn định hơn và có khả năng đoán nhận chính xác hơn. Tuy nhiên, phương pháp đề xuất này có nhược điểm là phải tiêu tốn một khoảng thời gian chạy để tìm ra bộ thuộc tính tối ưu tương đối lớn. Nhưng lại giảm được thời gian huấn luyện và kiểm thử cho tất cả các lần sử dụng bộ dữ liệu về sau này.

Liên hệ

HÀ VĂN SANG
Khoa HTTT Kinh Tế - Học Viện Tài chính
Email: sanghv@hvtc.edu.vn
Website: www.sanghv.com
Phone: 0982165568

Tài liệu tham khảo

1. [1]Nguễn Hà Nam, tối ưu hóa KPCA bằng GA để chọn các thuộc tính đặc trưng nhằm tăng hiệu quả phân lớp của thuật toán Random Forest, Tạp chí Khoa học ĐHQGHN, Khoa học Tự nhiên và Công nghệ 25 (2009) 84-93
2. [2]Nguyễn Đình Thúc, Lập trình tiến hóa, Nhà xuất bản giáo dục, 2001
3. [3]Huan Liu and Hiroshi Motoda, Computational Methods of Feature Selection, Chapman & Hall/CRC, 2008
4. [4]YongSeog Kim and Filipppo Meczenc, Feature Selection in Data Mining, 2005
5. [5]Jacek Jarmulak and Susan Craw, Genetic Algorithms for Feature Selection and Weighting, IJCAI 99 workshop, 1999
6. [6]Jihoon Yang and Vasant Honavar, Feature Subset Selection Using a Genetic Algorithm, Artifical Intelligence Research Group
7. [7]Krzysztof J.Cios, Witold Deddrycz, Roman W.Swiniarski, Lukasz A.Kurgan, Data Mining A Knowledge Discovery Approach, Springer, 2007
8. [8]Luis Carlos Molina et at, Feature Selection for Algorithms: A Survey and Experimental Evaluation, 2000
9. [9]Ron Kohavi and George H. John, Wrapper for Feature Subset Selection, AIJ special issuse on relevance, 1996