Tạp chí Khoa học ĐHQGHN, Khoa học Tự nhiên và Công nghệ 25 (2009) 84-93

Tối ưu hóa KPCA bằng GA để chọn các thuộc tính đặc trưng nhằm tăng hiệu quả phân lớp của thuật toán Random Forest

Nguyễn Hà Nam*

Khoa Công Nghệ Thông Tin, Trường ĐH Công Nghệ, ĐHQGHN, 144 Xuân Thủy, Hà Nội, Việt Nam

Nhận ngày 2 tháng 4 năm 2007

Tóm tắt. Phân tích thành phần chính (PCA) là một phương pháp khá nổi tiếng và hiệu quả trong quá trình làm giảm số thuộc tính của tập dữ liệu đầu vào. Hiện nay phương pháp hàm nhân đã được dùng để tăng khả năng áp dụng PCA khi giải quyết các bài toán phi tuyến. Phương pháp này đã được Scholkhof và đồng nghiệp của ông đưa ra với tên gọi là KPCA. Trong bài báo này chúng tôi sẽ trình bày một cách tiếp cận mới dựa trên hàm nhân để có thể chọn ra những thuộc tính tốt nhất để tăng khả năng phân lớp của thuật toán Random Forest (RF). Chúng tôi đã sử dụng giải thuật di truyền để tìm ra hàm nhân tối ưu cho việc tìm ra cách chuyển đổi phi tuyến tốt nhất nhằm làm tăng khả năng phân lớp của RF. Cách tiếp cận của chúng tôi về cơ bản đã tăng khả năng phân lớp của giải thuật RF. Không chỉ tăng được khả năng phân lớp cho thuật toán RF, phương pháp đề nghị còn cho thấy khả năng phân lớp tốt hơn một số phương pháp trích chọn đã được công bố.

Từ khóa: PCA, Hàm nhân, KPCA, Random Forest, trích chọn thuộc tính.

1. Giới thiệu

Trong lĩnh vực nghiên cứu về khai phá dữ liệu nói chung cũng như trong nghiên cứu về các thuật toán phân lớp nói riêng, vấn đề xử lý dữ liệu lớn ngày càng trở thành vấn đề cấp thiết và đóng vai trò chủ đạo trong việc giải quyết các bài toán thực tế. Phần lớn các thuật toán phân lớp đã phát triển chỉ có thể giải quyết được với một lượng số liệu giới hạn cũng như với một độ phức tạp dữ liệu biết trước. Trong khi đó lượng dữ liệu mà chúng ta thu thập được ngày càng trở nên phong phú và đa dạng nhờ sự phát triển mạnh mẽ của khoa học kỹ thuật. Mặc dù rất nhiều kỹ thuật khai phá dữ liệu dựa trên một số nền tảng lý thuyết khác nhau đã được phát triển và ứng dụng từ rất lâu, nhưng thực tế cho thấy kết quả phụ thuộc rất nhiều vào đặc tính dữ liệu cũng như khả năng xử lý dữ liệu thô của từng nhóm nghiên cứu. Một điều hiển nhiên là với mỗi phương pháp chỉ có thể đáp ứng và xử lý tốt trên một vài dữ liệu và ứng dụng cụ thể nào đó. Trong khai phá dữ liệu thì phương pháp trích chọn đóng một vai trò quan trọng trong tiền xử lý số liệu. Hướng tiếp cận này làm tăng hiệu năng thu nhận tri thức trong các ngành như tin sinh, xử lý dữ liệu web, xử lý tiếng nói, hình ảnh với đặc tính là có rất nhiều thuộc tích (vài trăm cho đến vài trăm ngàn thuộc tính) nhưng thường chỉ có một số lượng

* Tel.: 84-4-37547813.
E-mail: namnh@vnu.edu.vn

tương đối nhỏ các mẫu dùng để huấn luyện (thường là vài trăm). Phương pháp trích chọn sẽ giúp giảm kích cỡ của không gian dữ liệu, loại bỏ những thuộc tính không liên quan và những thuộc tính nhiễu. Phương pháp này có ảnh hưởng ngay lập tức đến các ứng dụng như tăng tốc độ của thuật toán khai phá dữ liệu, cải thiện chất lượng dữ liệu và vì vậy tăng hiệu suất khai phá dữ liệu, kiểm soát được kết quả của thuật toán. Phương pháp này đã được giới thiệu từ những năm 1970 trong các tài liệu về xác suất thống kê, học máy và khai phá dữ liệu [1-7].

Phân tích các thành phần cơ bản (PCA) [4] là một phương pháp khá nổi tiếng và hiệu quả trong quá trình làm giảm số thuộc tính của tập dữ liệu đầu vào. Gần đây phương pháp hàm nhân đã được áp dụng để có thể ứng dụng PCA vào giải quyết các bài toán phi tuyến tính. Phương pháp này đã được Scholkhof và đồng nghiệp của ông đưa ra với tên gọi là KPCA [9]. Trong bài báo này chúng tôi sẽ trình bày một cách tiếp cận mới dựa trên hàm nhân để có thể chọn ra những thuộc tính tốt nhất để tăng khả năng phân lớp của thuật toán Random Forest (RF). Trong phương pháp đề nghị, chúng tôi sử dụng giải thuật di truyền để tìm ra hàm nhân tối ưu cho việc tìm ra cách chuyển đổi phi tuyến tốt nhất nhằm làm tăng khả năng phân lớp của RF.

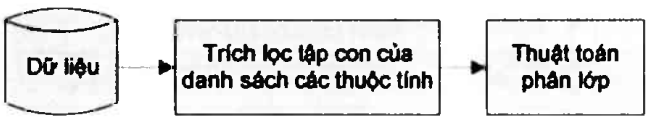
2. Cơ sở lý thuyết

2.1. Giới thiệu về trích chọn nội dung

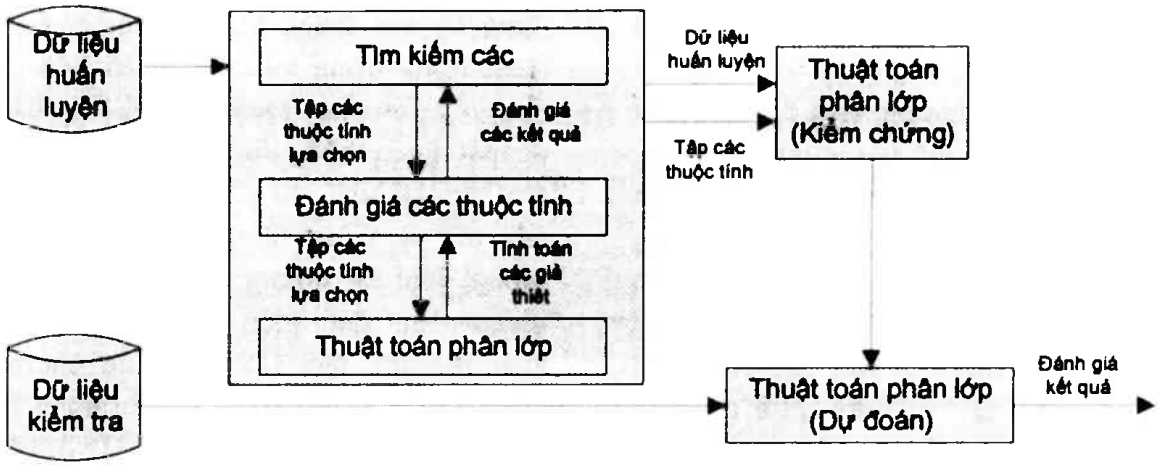
Về cơ bản việc bóc tách các thuộc tính đặc trưng bao gồm hai phần là xây dựng các thuộc tính và lựa chọn các thuộc tính đặc trưng. Xây dựng bộ các thuộc tính là một công việc rất quan trọng trong việc xử lý số liệu. Khi xây dựng dữ liệu chúng ta cần phải đảm bảo không để mất nhiều thông tin quá cũng như không quá tốn kém về mặt chi phí. Phần thứ hai có mục tiêu tìm ra những thuộc tính đại diện cho đối tượng, loại bỏ những thuộc tính thừa và gây nhiễu nhằm tăng hiệu suất của các thuật toán khai phá dữ liệu. Có rất nhiều phương pháp cũng như hướng tiếp cận khác nhau bao gồm các phương pháp kinh điển [1-3] với bộ dữ liệu tương đối nhỏ và các hướng tiếp cận hiện đại [5-7]. Tuy vậy chúng đều có một số các yêu cầu chung như sau:

- Giảm dữ liệu cần lưu trữ và tăng tốc độ của thuật toán (tính toán trên dữ liệu đó)
- Giảm bộ thuộc tính nhằm tiết kiệm không gian lưu trữ
- Tăng cường hiệu quả thuật toán: nhằm thu được tỷ lệ dự đoán đúng cao hơn
- Có tri thức về dữ liệu: thu được các tri thức về dữ liệu thông qua các phương pháp bóc tách dữ liệu để có thể tạo ra hay biểu diễn dữ liệu dễ dàng hơn.

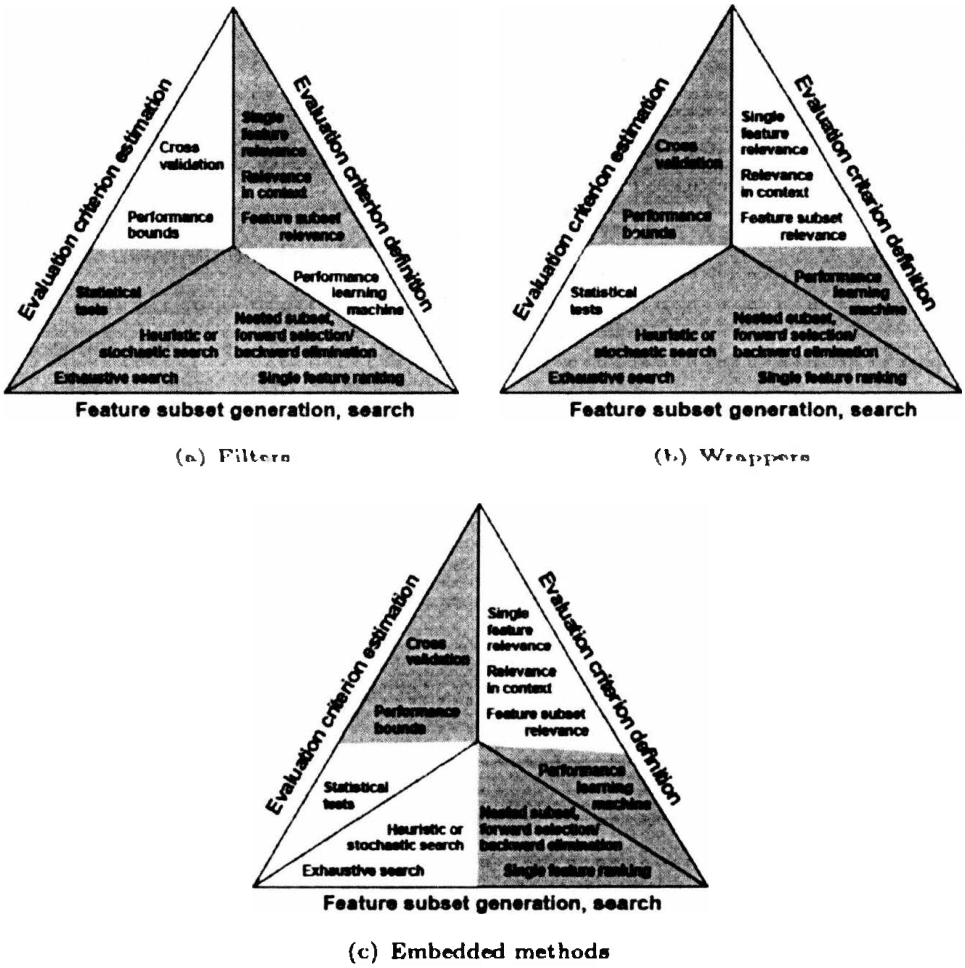
Về cơ bản chúng ta có thể phân loại các phương pháp trích chọn theo 2 cách tiếp cận khác nhau là filter/wrapper, được trình bày kỹ trong các tài liệu [1,2]. Lược đồ thực hiện của hai cách tiếp cận này được giản lược hóa trong hình vẽ 1 và 2 dưới đây.



Hình 1. Hướng tiếp cận filter (các thuộc tính được chọn độc lập với thuật toán khai phá dữ liệu) [1].



Hình 2. Hướng tiếp cận wrapper (các thuộc tính được chọn phụ thuộc theo một nghĩa nào đó với thuật toán khai phá dữ liệu) [1].



Hình 3. Ba cách tiếp cận cơ bản của trích chọn nội dung. Phần tô màu xám cho biết các thành phần mà hướng tiếp cận đó sử dụng để đưa ra kết quả cuối cùng.

Để thực hiện được các thuật toán trích chọn, chúng ta cần phải thực hiện một số công việc sau:

- Phương pháp để sinh ra tập thuộc tính đặc trưng (có thể hiểu tương ứng với các chiến lược tìm kiếm)
- Định nghĩa hàm đánh giá (đưa ra các tiêu chí để có thể xác định một thuộc tính hay nhóm thuộc tính là tốt hay không tốt)
- Ước lượng hàm đánh giá đó (kiểm chứng lại xem hàm đánh giá có thực sự phù hợp và hiệu quả với bộ dữ liệu không).

Hình vẽ 3 thể hiện sự khác nhau giữa các cách tiếp cận Filter, Wrapper và Embedded [8]. Hai phương pháp (a) và (b) đã được mô tả kỹ trong các tài liệu [1,2]. Phương pháp (c) tương đối giống cách tiếp cận (b) chỉ có điểm khác biệt là nó ghép phần sinh tập thuộc tính vào phần đánh giá trong khi huấn luyện.

2.2. Thuật toán di truyền

Có lớp các bài toán hay mà người ta chưa tìm được thuật toán tương đối nhanh để giải quyết chúng. Nhiều bài toán trong lớp này là các bài toán quy hoạch mà thường nảy sinh trong các ứng dụng cụ thể. Đối với dạng bài toán này, ta thường chỉ có thể tìm ra một thuật toán cho kết quả gần tối ưu. Ta cũng có thể dùng các thuật toán xác suất để xử lý chúng, những thuật toán này không đảm bảo cho ra kết quả tối ưu. Tuy nhiên, ta có thể giảm khá nhiều tỷ lệ sai của kết quả bằng cách chọn ngẫu nhiên đủ nhiều các "lời giải có thể". Nói một cách đơn giản, việc giải một bài toán có thể xem như việc tìm kiếm lời giải tối ưu trong một không gian các lời giải có thể. Vì cái đích của chúng ta là "lời giải tốt nhất", ta có thể coi công việc này là một quá trình tối ưu hóa. Đối với không gian nhỏ, phương pháp "vét cạn" cổ điển là đủ dùng; còn những không gian lớn hơn đòi hỏi các phương pháp tối ưu đặc biệt. Giải thuật di truyền là một trong số các phương pháp đặc biệt đó.

Thuật toán di truyền, cũng như các thuật toán tiến hóa nói chung, hình thành dựa trên quan niệm cho rằng: *quá trình tiến hóa tự nhiên là hoàn hảo nhất, hợp lý nhất và tự nó đã mang tính tối ưu*. Quan niệm này có thể được xem như là một tiên đề đúng và không chứng minh được, nhưng phù hợp với thực tế khách quan. Quá trình tiến hóa thể hiện tính tối ưu ở chỗ, thế hệ sau bao giờ cũng tốt hơn, phát triển hơn, hoàn thiện hơn thế hệ trước. Tiến hóa tự nhiên được duy trì nhờ hai quá trình cơ bản: sinh sản và chọn lọc tự nhiên. Xuyên suốt quá trình tiến hóa tự nhiên, các thế hệ mới luôn được sinh ra để bổ sung và thay thế cho thế hệ cũ. Cá thể nào phát triển hơn, thích ứng hơn với môi trường sẽ tồn tại, cá thể nào không thích ứng với môi trường sẽ bị đào thải. Sự thay đổi môi trường là động lực thúc đẩy quá trình tiến hóa. Ngược lại, tiến hóa cũng tác động trở lại góp phần làm thay đổi môi trường.

Trong thuật giải di truyền, các cá thể mới liên tục được sinh ra trong quá trình tiến hóa nhờ sự lai ghép ở thế hệ cha mẹ. Một cá thể mới có thể mang những tính trạng của cha mẹ (di truyền), cũng có thể mang những tính trạng hoàn toàn mới (đột biến). Di truyền và đột biến là hai cơ chế có vai trò quan trọng như nhau trong tiến hóa, dù rằng đột biến xảy ra với xác suất nhỏ hơn nhiều so với hiện tượng di truyền. Các thuật toán tiến hóa, tuy có những đặc điểm khác biệt, nhưng đều mô phỏng bốn quá trình cơ bản: Lai ghép, đột biến, sinh sản và chọn lọc tự nhiên.

Như vậy quá trình tiến hóa càng lâu thì càng có điều kiện cho các cá thể tốt được sinh ra, và chất lượng của các cá thể càng được nâng lên.

2.3. Thuật toán KPCA

Phương pháp PCA [4, 9, 10] là một phương pháp được sử dụng khá phổ biến và tương đối hiệu quả để biến đổi từ dữ liệu có số lượng thuộc tính lớn và nhiễu nhưng có độ tương quan với nhau thành một bộ dữ liệu có số chiều nhỏ hơn dựa trên các phép biến đổi tuyến tính [11]. Tuy nhiên trong nhiều ứng dụng thực tế, hiệu quả của phương pháp này rất hạn chế vì nền tảng xây dựng thuật toán dựa trên dữ liệu tuyến tính [12].

Để có thể áp dụng thuật toán này vào dữ liệu phi tuyến, đã có nhiều nghiên cứu ứng dụng các kỹ thuật khác nhau để có thể biến đổi dữ liệu đã cho thành dữ liệu được cho là tuyến tính. Nghiên cứu của Kramer [13] vào năm 1991 đã tìm cách phát triển thuật toán PCA phi tuyến dựa trên mạng nơ ron. Tuy nhiên mạng này tương đối phức tạp và rất khó tìm được giá trị tối ưu do có 5 lớp. Nghiên cứu của Dong và McAvoy [12] cũng sử dụng mạng nơ ron với giả thiết rằng sự phi tuyến của dữ liệu đầu vào có thể tương ứng với tổ hợp tuyến tính của một số đại lượng ngẫu nhiên và vì vậy có thể tách thành tổng các hàm của các đại lượng đó. Cách thức chuyển đổi đó chỉ có thể thực hiện được với một số rất hạn chế các bài toán phi tuyến.

Trong khoảng những năm cuối của thế kỷ trước, một phương pháp PCA phi tuyến mới đã được xây dựng và phát triển, có tên là KPCA (PCA dựa trên hàm nhân) bởi Scholkopf và đồng nghiệp của ông [9,10]. Phương pháp này thực hiện biến đổi phi tuyến trên hệ tọa độ bằng cách tìm các phần tử cơ bản có liên hệ phi tuyến với các giá trị đầu vào. Giả sử giá trị đầu vào là xk nằm trong không gian Rm với k=1,..., n, chúng ta có thể tính được ma trận tương quan (covariance matrix) của các giá trị đầu vào

$$Cov(x_i, x_j) = \frac{\sum_{i,j=0}^{n} (x_i - \mu_i)(x_j - \mu_j)}{n-1} \quad (1)$$

Sau đó giải hệ phương trình để tìm giá trị đặc trưng λ và véc tơ đặc trưng λv = Cv

Ý tưởng cơ bản của phương pháp hàm nhân [14] là các tính toán tương tự cũng có thể được thực hiện trong không gian tích vô hướng F có liên quan tới không gian giá trị đầu vào thông qua một biến đổi phi tuyến Φ: $R^m \rightarrow$ F và x→X. Ta có thể biểu diễn ma trận tương quan trong không gian F như sau, với giả sử là dữ liệu đã được chuyển về tâm của trục tọa độ

$$Cov(\Phi(x_i), \Phi(x_j)) = \frac{\sum_{i,j=0}^{n} (\Phi(x_j)\Phi(x_j)^T)}{n-1} \quad (2)$$

và tương tự chúng ta có thể tính được các giá trị đặc trưng tương tự như với PCA truyền thống với hàm nhân có dạng như sau

$$K_{i,j} = \left\langle \Phi(x_j)\Phi(x_j)^T \right\rangle \quad (3)$$

2.4. Thuật toán Random Forest

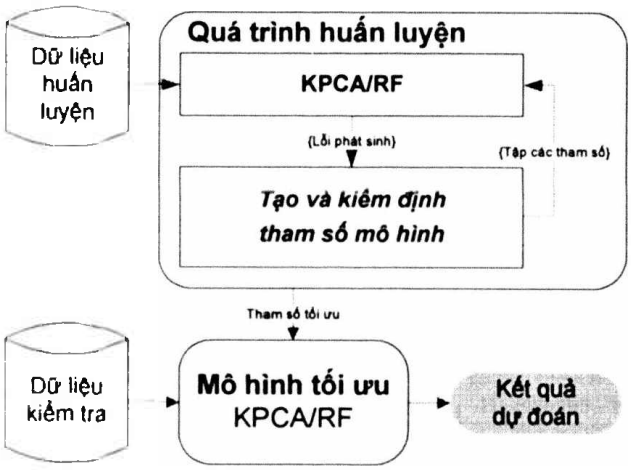
Random forest [15] là một thuật toán đặc biệt dựa trên kỹ thuật lắp ghép (ensemble techniques [4]). Về mặt bản chất thuật toán RF được xây dựng dựa trên nền tảng thuật toán phân lớp CART sử dụng kỹ thuật có tên gọi là bagging [4]. Kỹ thuật này cho phép lựa chọn một nhóm nhỏ các thuộc tính tại mỗi nút của cây để phân chia cho mức tiếp theo của cây phân lớp. Bằng cách chia nhỏ không gian tìm kiếm thành các cây nhỏ hơn như vậy cho phép thuật toán có thể phân loại một cách rất nhanh chóng cho dù không gian thuộc tính rất lớn. Các tham số đầu vào của thuật toán khá đơn giản bao gồm số các thuộc tính được chọn trong mỗi lần phân chia (mtry). Giá trị mặc định của tham số này là căn bậc hai của p với p là số lượng các thuộc tính. Tương tự như thuật toán CART, RF vẫn sử dụng công thức Gini [4] là công thức tính toán việc phân chia cây. Số lượng cây được

được tạo ra là không hạn chế và cũng không sự dụng bất kỳ kỹ thuật để hạn chế mở rộng cây. Chúng ta phải lựa chọn tham số cho biết số lượng cây (ntree) sẽ được sinh ra sao cho đảm bảo rằng sẽ mỗi một thuộc tính sẽ được kiểm tra một vài lần. Thuật toán sử dụng kỹ thuật OOB (out-of -bag) [15] để xây dựng tập huấn luyện và phương pháp kiểm tra trên nó.

3. Nội dung và kết quả nghiên cứu

3.1. Mô hình đề nghị

Kiến trúc cơ bản của hệ thống bao gồm ba phần chính: tiền xử lý số liệu, quá trình học để tìm ra tập các tham số tối ưu và cuối cùng là mô đun phân lớp số liệu chưa được sử dụng trong các quá trình trước đó.



Hình 4. Kiến trúc tổng thể của phương pháp đề nghị (KPCA-RF) với mô hình học để tìm ra hàm nhân tốt nhất..

Trong mô đun tiền xử lý, chúng tôi đã sử dụng kỹ thuật t-test [3,4] nhằm làm giảm số lượng các thuộc tính để làm giảm bớt khối lượng tính toán cũng như giảm độ nhiễu của dữ liệu. Sau đó dữ liệu được phân chia thành các tập dữ liệu huấn luyện và tập dữ liệu kiểm tra bao gồm một số mẫu là của bệnh nhân ung thư còn một số khác bình thường.

Tiếp theo, chúng tôi sử dụng thuật toán di truyền để tìm hệ số tốt nhất để xây dựng hàm nhân theo công thức (4) sẽ được trình bày ở phần 3.2. Hàm nhân này được sử dụng trong KPCA như một cách để biến đổi không gian ban đầu thành không gian mới với hy vọng có thể phân lớp dễ dàng và hiệu quả hơn dựa trên mô đun phân lớp RF. Ở đây thuật toán di truyền được sử dụng để tạo ra một bộ các giá trị thực β nằm trong khoảng (0, 1). Bộ giá trị này được sử dụng đề xây dựng công thức của hàm nhân nhằm biến đổi từ không gian số liệu ban đầu vào một không gian mới thông qua mô đun KPCA. Phép biến đổi này được đánh giá thông qua tỷ lệ lỗi phân lớp được tạo ra bởi mô đun RF. Quá trình tìm bộ hệ số β được thực hiện dựa trên quá trình thực hiện các thủ tục của thuật toán di truyền với hàm định giá dựa trên RF. Quá trình này được lặp lại cho tới khi đạt được kết quả tối ưu.

Sau khi kết thúc quá trình tìm tập các hệ số dựa trên thuật toán di truyền, các kết quả này sẽ được chuyển đầy đủ sang mô đun phân lớp với các dữ liệu chưa được phân loại trước đó.

3.2. Xây dựng hàm nhân và phương pháp học

Như đã trình bày ở các phần trên, việc chuyển đổi không gian phi tuyến ban đầu thành không gian tuyến tính để có thể dễ dàng thực hiện thuật toán PCA được thực hiện một cách dễ dàng và hiệu quả thông qua hàm nhân. Đã có rất nhiều hàm nhân được xây dựng và công bố cho các ứng dụng cụ thể khác nhau, tuy nhiên việc chọn ra một hàm nhân đủ tốt cho một ứng dụng hay một loại số liệu cụ thể luôn luôn là một thách thức không nhỏ đối với các nhà nghiên cứu. [10]

Ở đây chúng tôi dựa vào một số kết quả trình bày trong các tài liệu [10,14] để giới thiệu

một cách thức xây dựng hàm nhân phù hợp cho việc xử lý số liệu tin sinh học. Hàm nhân do chúng tôi xây dựng được biểu diễn như sau

$$K_c = \sum_{i=1}^{m} \beta_i \times K_i \qquad (4)$$

Thỏa mãn

$$\beta \in [0,1], \sum_{i=1}^{m} \beta_i = 1$$

Trong đó K_i là những hàm nhân đã được xây dựng trước đó, hệ số β_i thể hiện ảnh hưởng của hàm nhân thứ i vào hàm nhân chính. Để chứng minh hàm nhân vừa được xây dựng thỏa mãn các điều kiện của một hàm nhân chúng ta có thể sử dụng bổ đề 3.12 và nội dung của định lý Mercer đã được trình bày trong [14]

Hệ số β đóng một vai trò rất quan trọng trong việc tạo ra hàm nhân phù hợp với dữ liệu đầu vào. Trong quá trình học, cấu trúc của tập dữ liệu huấn luyện sẽ được học một cách tự động thông qua viêc thay đổi hệ số này. Như đã trình bày ở phần trước, chúng tôi sử dụng thuật toán di truyền để tìm ra hệ số β phù hợp nhất sao cho tối thiểu hóa được lỗi phát sinh trong quá trình học.

4. Kết quả và thảo luận

4.1. Môi trường thực nghiệm

Tất cả các thực nghiệm được thực hiện trên máy tính Pentium IV 1.8GHz. Phương pháp đề nghị được thực hiện trên ngôn ngữ R, đây là ngôn ngữ chuyên dùng trong xác suất thống kê (có thể tải về tại địa chỉ http://www.r-project.org), các mô đun KPCA và RF cũng được tải về từ địa chỉ trên.

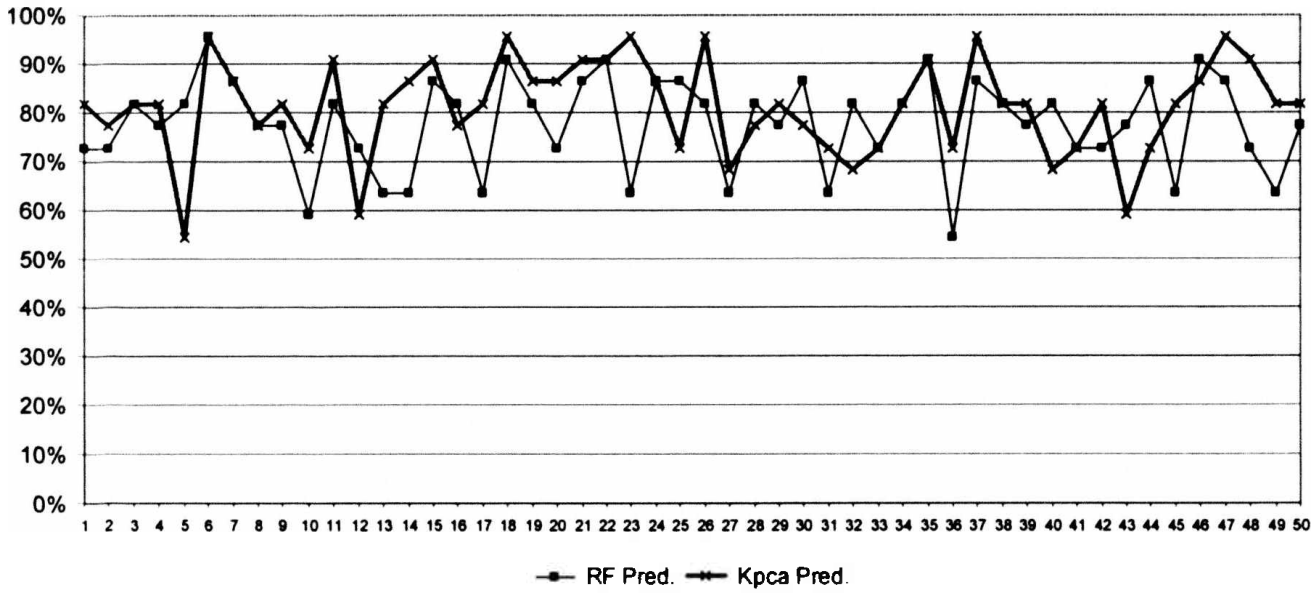
4.2. Bộ dữ liệu ung thư ruột kết

Bộ dữ liệu ung thư ruột kết (Colon Tumor cancer). Bộ dữ liệu ung thư ruột kết [16] bao gồm thông tin về gen được trích ra từ hệ thống DNA microarray. Bộ dữ liệu này bao gồm 62 mẫu với 22 mẫu của người bình thường và 40 mẫu của người có bệnh và có tổng số 2000 thuộc tính. Chúng tôi chọn ngẫu nhiên 40 mẫu làm tập huấn luyện và 22 mẫu còn lại được sử dụng làm tập kiểm tra.

4.3. Quy trình thực nghiệm và kết quả

Đầu tiên chúng tôi thực hiện việc thu gọn dữ liệu sử dụng t-test, tiếp theo giải thuật di truyền được sử dụng để tìm ra hàm nhân phù hợp cho KPCA nhằm chuyển đổi không gian tối ưu nhất cho việc áp dụng phân lớp RF. Thực nghiệm đã được thực hiện 50 lần để kiểm tra sự ổn định của phương pháp đề nghị.

Kỹ thuật t-test được áp dụng để lựa chọn khoảng 1000 thuộc tính tốt nhất và sau đó được dùng là dữ liệu đầu vào của chương trình KPCA_RF. Hình vẽ 5 so sánh kết quả giữa thuật toán RF nguyên gốc và thuật toán học của chúng tôi thông qua 50 lần thực nghiệm. Trung bình thuật toán RF cho kết quả là 77.64% với phương sai là 9.62%, còn thuật toán KPCA-RF cho kết quả đoán nhận là 81.09% với phương sai là 9.82%. Kết quả trên cho thấy thuật toán đề nghị của chúng tôi đã cho kết quả tốt hơn hẳn so với thuật toán RF cơ sở ban đầu.



Hình 5. So sánh kết quả đoán nhận giữa thuật toán RF với thuật toán đã được cải tiến KPCA-RF thông qua 50 lần thực nghiệm. Đường nét đậm thể hiện kết quả của thuật toán của chúng tôi, còn đường mảnh thể hiện kết quả của thuật toán RF..

Bảng 1 cho biết kết quả dự đoán của một số nghiên cứu có cùng hướng tiếp cận trích chọn nội dung đã công bố. So sánh với những kết quả này tỷ lệ dự đoán của hệ thống đề nghị đã đạt được kết quả tương đối khả quan.

Bảng 1. So sánh kết quả phân lớp với một số nghiên cứu trước đây với phương pháp đề nghị trên cùng bộ dữ liệu

Các phương pháp	Tỷ lệ dự đoán đúng (%)
Bootstrapped GA\SVM [17]	80.0
Combined kernel for SVM [18]	75.33±7.0
KPCA-RF	81.09+9.85.2

Kết luận

Trong bài báo này chúng tôi giới thiệu một phương pháp mới nhằm mục tiêu giảm số lượng thuộc tính của dữ liệu đầu vào trước khi áp dụng một phương pháp phân lớp đã biết. Về cơ bản thì RF là một phương pháp tương đối tốt trong việc xử lý số liệu với số chiều tương đối lớn và với số lượng mẫu huấn luyện tương đối nhỏ. Phương pháp đề nghị của chúng tôi nhằm giảm thời gian tính toán cũng như giảm độ nhiễu của dữ liệu đầu vào bằng cách áp dụng kỹ thuật hàm nhân PCA. Chúng tôi đã xây dựng hàm nhân và phương pháp tìm ra hàm nhân tối ưu thông qua việc sử dụng giải thuật di truyền. Cách tiếp cận của chúng tôi về cơ bản đã tăng khả năng phân lớp của giải thuật RF được thể hiện thông qua hình 4. Không chỉ tăng được khả năng phân lớp cho thuật toán RF, phương pháp đề nghị còn cho thấy khả năng phân lớp tốt hơn một số phương pháp trích chọn đã được công bố (Bảng 1).

Lời cảm ơn

Công trình này được tài trợ một phần từ đề tài mang mã số: QG.08.01, Đại học Quốc gia Hà Nội.

References

[1] R. Kohavi, G.H. John, Wrappers for Feature Subset Selection, *Artificial Intelligence* Vol 97 (1997) 273.

[2] A.L. Blum, P. Langley, Selection of Relevant Features and Examples in Machine Learning, *Artificial Intelligence* Vol 97 (1997) 245.

[3] Pang-Ning Tan, Michael Steinbach, and Vipin Kumar, *Introduction to Data Mining*, Addison Wesley; 1st edition, May 2, 2005.

[4] R. O. Duda, P. E. Hart, D. G. Stork, *Pattern Classification (2nd Edition)*, John Wiley & Sons Inc, 2001.

[5] Luis Carlos Molina, Luis Belanche, Àngela Nebot: Feature Selection Algorithms, A Survey and Experimental Evaluation, Technical report, Universitat Politècnica de Catalunya Departament de Llenguatges i Sistemes Informátics, France, 2002.

[6] H. Liu, L. Yu, Feature Selection for Data Mining, Technical report, Department of Computer Science and Engineering Arizona State University, America, 2002.

[7] I. Guyon, A. Elisseeff, An introduction to variable and feature selection. *Journal of Machine Learning Research* 3 (2003) 1157.

[8] I. Guyon, J. Weston, S. Barnhill, V. Vapnik, Gene Selection for Cancer Classification using Support Vector Machines, *Machine Learning*, Vol 46 (2002) 389.

[9] B. Scholkopf, A.J. Smola, K. Muller, Nonlinear component analysis as a kernel eigenvalue problem, *Neural Computation* 10 (5), 1998.

[10] B. Scholkopf, A.J. Smola, *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*, MIT press, 2002.

[11] B.M. Wise, N.B. Gallagher, The process chemometrics approach to process monitoring and fault detection, *Journal of Process Control* 6 (1996) 6.

[12] D. Dong, T.J. McAvoy, Nonlinear principal component analysis based on principal curves and neural networks, *Computers and Chemical Engineering* 20 (1996) 65.

[13] M.A. Kramer, Nonlinear principal component analysis using autoassociateive neural networks, *A.I.Ch.E. Journal* 37 (1991) 233.

[14] N. Cristianini, J. Shawe-Taylor, An introduction to Support Vector Machines and other kernel-based learning methods, Cambridge, (2000).

[15] L. Breiman, Random forest, *Technical report*, Statistics Department University of California Berkeley (2001).

[16] U. Alon, N. Barkai, D. Notterman, K. Gish, S. Ybarra, D. Mack, A. Levine.: Broad Patterns of Gene Expression Revealed by Clustering Analysis of Tumor and Normal Colon Tissues Probed by Oligonucleotide Arrays, *Proceedings of National Academy of Sciences of the United States of American* (1999).

[17] Xue-wen Chen, Gene Selection for Cancer Classification Using Bootstrapped Genetic Algorithms and Support Vector Machines, *IEEE Computer Society Bioinformatics Conference* (2003).

[18] H.N Nguyen, S.Y. Ohn, J. Park, K.S. Park, Combined Kernel Function Approach in SVM for Diagnosis of Cancer, *Proceedings of the First International Conference on Natural Computation* (2005).

Optimization of KPCA by GA for selecting relevant features to improving the effection of Random Forest classifier

Nguyen Ha Nam

Falcuty of Information Technology, College of Technology, Vietnam National University, Hanoi, 144 XuanThuy, Hanoi, Vietnam

This paper proposed a combination of kernel functions Kernel Principle Component Analysis and its learning method which is help to not only transform the input space to a lower dimension feature space but also increase the classification performance. We defined the combined kernel function as the weighted sum of a set of difference types of basis kernel function consisting of polynomial, gausian and neural kernels, which is trained by a novel learning method based on genetic algorithm. The weights of basis kernel functions in the combined kernel are determined in learning phase and used as the parameters in the decision model in the classification phase. The unified kernel and the learning method were applied to obtain the optimal decision model for the classification of a public data set for diagnosis of cancer diseases. The experiment showed fast convergence in learning phase and resulted in the optimal decision model with the better performance than other kernels. Therefore, the proposed kernel function has the greater flexibility in representing a problem space than other kernel functions.

Keywords: PCA, Kernel function, KPCA, Random Forest, Feature Selection.