Journal of educational equipment: Applied research, Volume 2, Issue 309 (March 2024)
ISSN 1859 - 0810

Phát hiện xâm nhập mạng sử dụng thuật toán di truyền lai với cây quyết định

Nguyễn Mạnh Hùng*

**Khoa Công nghệ Thông tin, Trường Đại học Hải Phòng.*
Received: 26/02/2024; Accepted: 06/03/2024; Published: 15/03/2024

Abstract: *Machine Learning techniques such as Genetic Algorithms and Decision Trees have been applied to the field of intrusion detection for more than a decade. Machine Learning techniques can learn normal and anomalous patterns from training data and generate classifiers that then are used to detect attacks on computer systems. In general, the input data to classifiers is in a high dimension feature space, but not all of features are relevant to the classes to be classified. In this paper, we use a genetic algorithm to select a subset of input features for decision tree classifiers, with a goal of increasing the detection rate and decreasing the false alarm rate in network intrusion detection. We used the KDDCUP 99 data set to train and test the decision tree classifiers. The experiments show that the resulting decision trees can have better performance than those built with all available features.*

Keywords: *Genetic Algorithm, Decision Trees, Intrusion Detection*

1. Đặt vấn đề

Hệ thống phát hiện xâm nhập (IDSs) đã trở thành một tiêu điểm chính của các nhà khoa học máy tính cũng như các cuộc tấn công máy tính đã trở thành một mối đe dọa ngày càng tăng đối với thương mại điện tử. Cộng đồng bảo mật máy tính đã phát triển một loạt các hệ thống phát hiện xâm nhập để ngăn chặn các cuộc tấn công vào hệ thống máy tính. Có hai loại chính của hệ thống phát hiện xâm nhập: phát hiện bất thường và phát hiện lạm dụng. Các hệ thống phát hiện bất thường tìm cách xác định độ lệch so với các mô hình hành vi bình thường được xây dựng từ các bộ dữ liệu đào tạo lớn. Hệ thống phát hiện lạm dụng so sánh hành vi sử dụng hệ thống với chữ ký được trích xuất từ các cuộc tấn công đã biết. Hai loại hệ thống này có điểm mạnh và điểm yếu riêng. Trước đây có thể phát hiện các cuộc tấn công mới, nhưng nói chung đối với hầu hết các hệ thống hiện có như vậy, có tỷ lệ báo động sai cao vì rất khó để tạo ra các hồ sơ hành vi bình thường thực tế cho các hệ thống được bảo vệ. Loại thứ hai có thể phát hiện các cuộc tấn công đã biết với độ chính xác rất cao thông qua việc khớp mẫu trên các chữ ký đã biết, nhưng không thể phát hiện các cuộc tấn công mới vì chữ ký của chúng chưa có sẵn để khớp mẫu. Trong bài viết này, chúng tôi chỉ xem xét các hệ thống phát hiện lạm dụng.

Kỹ thuật học máy gần đây đã được áp dụng rộng rãi để phát hiện xâm nhập. Các phương pháp tiếp cận ví dụ bao gồm cây quyết định [1], Thuật toán di truyền và Lập trình di truyền [5][6], naive Bayes [1], kNN và mạng thần kinh [4][8]. Một vấn đề quan trọng là làm thế nào để chọn các thuộc tính của dữ liệu đào tạo đầu vào mà việc học sẽ diễn ra. Vì không phải mọi thuộc tính của dữ liệu đào tạo có thể liên quan đến nhiệm vụ phát hiện cho nên việc chọn một bộ dữ liệu tốt sẽ rất quan trọng để cải thiện độ tin cậy các phân loại. Gartner et al. [7] sử dụng các máy vector hỗ trợ để tìm trọng lượng tính năng tối ưu cho bộ phân loại Bayes. Theo hiểu của chúng tôi, việc kết hợp một thuật toán di truyền với các phân loại cây quyết định chưa được thử để phát hiện xâm nhập. Trong bài báo này, chúng tôi sử dụng một thuật toán di truyền để tìm một tập hợp con tối ưu các tính năng cho các phân loại cây quyết định dựa trên bộ dữ liệu KDDCUP 99 liên quan đến các đặc điểm của bốn loại tấn công: Probe, DOS, U2R và R2L.

2. Nội dung nghiên cứu

2.1. Cây quyết định

Bộ phân loại cây quyết định của Quinlan là một trong những kỹ thuật học máy nổi tiếng nhất. Một cây quyết định được tạo thành từ các nút quyết định và các nút lá. Mỗi nút quyết định tương ứng với một thử nghiệm X trên một thuộc tính duy nhất của dữ liệu đầu vào và có một số nhánh, mỗi nhánh xử lý kết quả của thử nghiệm X. Mỗi nút lá đại diện cho một lớp là kết quả của quyết định cho một trường hợp.

Quá trình xây dựng một cây quyết định về cơ bản là một quá trình phân chia và chinh phục. Một tập hợp T dữ liệu đào tạo bao gồm các lớp k (C 1, C 2, ..., C k). Nếu T chỉ bao gồm các trường hợp của một lớp duy nhất, T sẽ là một chiếc lá. Nếu T không chứa trường hợp, T là một lá và lớp liên kết với lá này sẽ được gán với lớp chính của nút mẹ của nó (đây là sự lựa chọn của C4.5). Nếu T chứa các trường hợp của các lớp hỗn

hợp (tức là nhiều hơn một lớp), một bài kiểm tra dựa trên một số thuộc tính ai của dữ liệu đào tạo sẽ được thực hiện và T sẽ được chia thành n tập con (T 1, T 2, ..., T n), trong đó n là số lượng kết quả của bài kiểm tra trên thuộc tính ai . Quá trình xây dựng cây quyết định tương tự được thực hiện đệ quy trên mỗi T j , điều kiện $1 \leq j \leq n$, cho đến khi mỗi tập con thuộc về một lớp duy nhất.

2.2. Thuật toán di truyền

Thuật toán di truyền (GAs) [9] đã được áp dụng thành công để giải quyết các vấn đề tìm kiếm và tối ưu hóa. Ý tưởng cơ bản của GA là tìm kiếm một không gian giả thuyết để tìm ra giả thuyết tốt nhất. Một nhóm các giả thuyết ban đầu called một dân số được tạo ngẫu nhiên và mỗi giả thuyết được đánh giá với một chức năng thể chất. Các giả thuyết có thể lực cao hơn có xác suất được chọn cao hơn để tạo ra thế hệ tiếp theo. Một phần nhỏ các giả thuyết tốt nhất có thể được đào tạo lại cho thế hệ tiếp theo, phần còn lại trải qua các hoạt động di truyền như chéo và đột biến để tạo ra các giả thuyết mới. Quy mô của một dân số là như nhau cho tất cả các thế hệ trong việc thực hiện của chúng tôi. Quá trình này được lặp lại cho đến khi đáp ứng tiêu chí fitness được xác định trước hoặc đạt được số lượng thế hệ tối đa được đặt trước.

Một GA thường có bốn thành phần. Một dân số của các cá nhân mà mỗi cá nhân trong dân số đại diện cho một giải pháp khả thi. Một chức năng thể chất là một chức năng đánh giá mà chúng ta có thể biết liệu một cá nhân có phải là một giải pháp tốt hay không. Một chức năng lựa chọn quyết định làm thế nào để chọn những cá nhân tốt từ dân số hiện tại để tạo ra các chi tiếp theo. Các nhà khai thác di truyền như chéo và đột biến khám phá các khu vực mới của không gian tìm kiếm trong khi vẫn giữ một số thông tin hiện tại cùng một lúc.

2.3. Lựa chọn tính năng dựa trên GA cho cây quyết định

Thuật toán lựa chọn thuộc tính dựa trên GA được đề xuất của chúng tôi dựa trên mô hình bọc như đã thảo luận trong phần 2. Trong thuật toán cải tiến của chúng tôi, thành phần tìm kiếm là GA và thành phần đánh giá là cây quyết định. Một mô tả chi tiết về thuật toán này được hiển thị trong Hình 1. Dân số ban đầu được tạo ngẫu nhiên. Mỗi cá nhân của dân số có 41 gen, mỗi gen đại diện cho một tính năng của dữ liệu đầu vào và có thể được gán cho 1 hoặc 0. 1 có nghĩa là tính năng được đại diện được sử dụng trong quá trình xây dựng cây quyết định; 0 có nghĩa là nó không được sử dụng. Kết quả là, mỗi cá nhân trong dân số đại diện cho một sự lựa chọn tính năng có sẵn. Đối với mỗi cá nhân trong dân số hiện tại, một cây quyết định được xây dựng bằng cách sử dụng chương trình C4.5. Cây quyết định kết quả này sau đó được kiểm tra trên chín bộ dữ liệu xác nhận, tạo ra chín tỷ lệ lỗi phân loại. Sự phù hợp riêng lẻ này là tổng hợp của các tỷ lệ lỗi phân loại này. Tỷ lệ lỗi phân loại càng thấp, thể chất của cá nhân càng tốt.

Một khi các giá trị thể chất của tất cả các cá nhân của dân số hiện tại đã được tính toán, GA bắt đầu di truyền thế hệ tiếp theo như sau:

(1) Chọn cá nhân theo phương pháp Xếp hạng [2].

(2) Sử dụng hai điểm chéo để trao đổi gen giữa cha mẹ để tạo ra con cái.

(3) Thực hiện một đột biến cấp độ nhỏ cho mỗi con cái.

(4) Giữ hai cha mẹ ưu tú và thay thế tất cả các cá thể khác của dân số hiện tại bằng con cái

Hình 2.1: Thuật toán di truyền và cây quyết định

Thủ tục trên được thực hiện lặp đi lặp lại cho đến khi đạt được số lượng tối đa của các thế hệ (100). Cuối cùng, cá nhân tốt nhất của thế hệ trước được chọn để xây dựng bộ phân loại cây quyết định cuối cùng, được thử nghiệm trên bộ dữ liệu thử nghiệm.

2.4. Tiến hành khai phá dữ liệu

Mục đích chính của công việc này là để xem liệu lai giải thuật di truyền và cây quyết định có thể tạo ra một phân loại tốt hơn các cuộc tấn công so với biểu diễn tốt nhất hiện tại của cây quyết định riêng lẻ hay không?. Chúng tôi sử dụng 10% dữ liệu đào tạo KDDCUP99 (489843 trường hợp) và dữ liệu kiểm thử đầy đủ (311029 trường hợp) cho các thí nghiệm của chúng tôi. Chúng tôi chia dữ liệu đào tạo và dữ liệu thử nghiệm thành bốn bộ dữ liệu đào tạo nhỏ hơn và bộ dữ liệu thử nghiệm theo bốn loại tấn công (Probe, DOS, R2L và U2R). Ví dụ: bộ dữ liệu đào tạo và bộ dữ liệu thử nghiệm cho DOS bao gồm tất cả các cuộc tấn công DOS và tất cả các trường hợp bình thường trong dữ liệu đào tạo và kiểm tra ban đầu. Đối với mỗi thể loại tấn công, chúng tôi chia dữ liệu đào tạo của nó thành mười tệp riêng biệt có kích thước bằng nhau. Một được chọn làm bộ dữ liệu đào tạo và phần còn lại là bộ xác thực. Chúng tôi chạy các thí nghiệm cho cả bốn loại tấn công và xây dựng một cây quyết định cho mỗi loại.

Có thể thấy rằng lựa chọn Hỗn hợp luôn tạo ra kết quả tốt hơn so với cây quyết định một mình trên

tập dữ liệu xác thực (bên trái). Khoảng tin cậy 95% được xây dựng xung quanh giá trị trung bình của 20 lần chạy (2 độ lệch chuẩn) và cũng thấp hơn nhiều so với giá trị được tạo ra bởi cây quyết định. Một trend tương tự được nhìn thấy trong biểu đồ cho tập dữ liệu thử nghiệm (bên phải). Giá trị trung bình được tạo ra bởi thuật toán Hỗn hợp có thể cải thiện vượt ra ngoài giá trị bằng cây quyết định vào cuối cuộc chạy, khi phương sai giảm và thuật toán GA hội tụ. Lý do cho sự gia tăng lớn về hiệu suất cho bộ xác nhận so với bộ thử nghiệm có thể được quy cho thực tế là các giá trị xác nhận nằm trong vòng lặp phản hồi của mô hình bọc. Con lai có thể tạo ra những cây quyết định tốt hơn với bộ huấn luyện vì tỷ lệ lỗi của các bộ xác nhận là chức năng thể chất. Điều này tạo ra sự thiên vị đối với dữ liệu xác thực. Ngoài ra, lỗi bộ thử nghiệm cao hơn nhiều so với lỗi thiết lập xác nhận vì bộ dữ liệu thử nghiệm KDDCUP99 giới thiệu các cuộc tấn công không bao giờ thấy before khó phát hiện trong bối cảnh các hệ thống phát hiện lạm dụng.

Để phân tích sâu hơn về cách GA tạo ra kết quả tốt hơn, tần suất sử dụng gen của các cá nhân hàng đầu được kiểm tra cho danh mục DOS. Ý nghĩa and tên đầy đủ của mỗi tính năng được đưa ra trong. Gen của cá nhân tốt nhất trong mỗi thế hệ được theo dõi cho 20 lần chạy. Mục đích là để khám phá những gen có tầm quan trọng có tần số trên hoặc dưới một số ngưỡng cách xa giá trị 0,50. Một giá trị tần số 0,50 có nghĩa là, về mặt xác suất, nó không matter cho dù tính năng đó đang bật hay tắt. Như đã thấy trong Bảng 2, thế hệ thứ 0 chứa sự phân bố tần số gen đồng đều nhất. Tuy nhiên, vẫn có thể thấy rằng cá nhân được chọn hàng đầu đã chứa một số đặc điểm cơ bản của các gen quan trọng (protocol_type, src_bytes). Khi các thế hệ tiến triển, các gen quan trọng và không quan trọng bắt đầu di chuyển đến các thái cực tương ứng của chúng. Tần số thấp chỉ ra rằng một gen cụ thể không quan trọng trong khi tần số cao chỉ ra rằng một gen là quan trọng. Ví dụ, một số gen (dịch vụ, nóng, d_h_count, same_srv_rate) được loại bỏ theo thời gian và tiến triển về 0 trong khi những gen khác (dst_bytes, wrong_fragment diff_srv_rate) được tăng cường và tăng lên về phía một. Sử dụng một số tính năng không quan trọng có thể khiến cây quyết định thực hiện "cách dễ dàng" để phân vùng dữ liệu tối đa hóa lợi ích thông tin; tuy nhiên, nó không tạo ra một quyết định phân vùng thông minh. Phần GA của thuật toán đã có thể loại bỏ các tính năng không quan trọng và xác định những tính năng cần thiết để phân loại hiệu quả.

3. Kết luận

Thuật toán di truyền và lai cây quyết định có thể vượt trội hơn thuật toán cây quyết định mà không cần lựa chọn tính năng. Chúng tôi tin rằng sự cải tiến này là do thực tế là cách tiếp cận lai có thể tập trung vào các tính năng có liên quan và loại bỏ các tính năng không cần thiết hoặc gây mất tập trung. Bộ lọc ban đầu này có thể cải thiện khả năng phân loại của cây quyết định. Thuật toán mất nhiều thời gian hơn để thực hiện so với cây quyết định tiêu chuẩn; tuy nhiên, quá trình không xác định của nó có thể tạo ra cây quyết định tốt hơn. Quá trình đào tạo chỉ cần được thực hiện một lần. Quá trình phân loại mất cùng một khoảng thời gian cho các hệ thống lai và không lai

Tài liệu tham khảo

1. Amor, N.B., Benferhat, S., và Elouedi, Z. Naive Bayes vs cây quyết định trong các hệ thống phát hiện xâm nhập. Trong *Proc. 2004 ACM Symp. on Applied Computing*. pp. 420-424, 2004.

2. Balthrop, J., Esponda, F., Forrest, S., và Glickman, M. Bảo hiểm và khái quát hóa trong một hệ thống miễn dịch nhân tạo. Trong *Proc. Hội nghị tính toán di truyền và tiến hóa. (GECCO)*, trang 3-10, 2002.

3. Botha, M., Solms, R. V., Perry, K., Loubser, E., và Yamoyany, G. Việc sử dụng trí tuệ nhân tạo trong một hệ thống phát hiện xâm nhập lai. Trong *Kỷ yếu saicsit 2002*, trang 149–155, 2002.

4. Crosbie, M., và Spafford, G. Áp dụng lập trình di truyền để phát hiện xâm nhập. In *Proc.1995 AAAI Symposium on Genetic Programming*, pp. 1-8.

5. Dasgupta, D., và Gonzalez, F. A. Một hệ thống hỗ trợ quyết định thông minh để phát hiện và phản ứng xâm nhập. Trong *Proc. Hội thảo Int'l về phương pháp toán học, mô hình và vòm. Đối với bảo mật mạng máy tính*, trang 1-14, 2001.

6. Gartner, T., và Flach, P. A. WBCsvm: Weighted Bayesian Classification dựa trên máy vector hỗ trợ. Trong *Proc. Hội nghị quốc tế lần thứ 18* về học máy (ICML- 2001), pp. 156-161.

7. Ghosh, A., và Schwartzbard, A. Một nghiên cứu trong việc sử dụng mạng lưới thần kinh cho sự bất thường và lạm dụng Phát hiện. *Hội nghị chuyên đề bảo mật USENIX lần thứ* 8, trang 141-151, 1999.

8. Hà Lan, J. H. (1975). *Thích nghi trong các hệ thống tự nhiên và nhân tạo*. Nhà xuất bản Đại học Michigan (tái bản năm 1992 bởi MIT Press, Cambridge, MA).

9. Huang, Z., Pei, M., Goodman, E., Huang, Y., và Li, G. Thuật toán di truyền tối ưu hóa chuyển đổi tính năng: so sánh với các phân loại khác nhau. Trong *Proc. GECCO 2003*, trang 2121-2133.

10. KDDCUP 1999 http://kdd.ics.uci.edu/databases/kddcup99/kddcup99.html