cầu đưa ra hàm đánh giá xem thế cờ hiện tại đang lợi thế cho bên nào. Một thế cờ không thể hiện được kết quả thắng thua của trận đấu, do đó các nước đi tiếp theo được giả lập, rồi thông qua đó đánh giá lợi thế. Để làm được những điều này, những hoàn cảnh khác nhau được tự động xây dựng bởi hệ thống để phụ vụ quá trình tự học.

2.1.2. Một số kỹ thuật học máy phổ biến

2.1.2.1 Thuật toán cây quyết định

Các thuật toán cây quyết định (hay Decision tree) thuộc mô hình học có giám sát, có thể được áp dụng vào cả hai bài toán phân loại (classification) và bài toán hồi quy (regression), được đặc trưng bởi các kiến thức sau quá trình học được biểu diễn dưới dạng cây quyết định. Phương pháp dự đoán, phân loại một mẫu x được đưa ra phụ thuộc vào quá trình di chuyển của x đi từ ngọn cây về một lá.

Việc xây dựng cây quyết định trên dữ liệu huấn luyện cho trước tiến hành xác định các câu hỏi và thứ tự của chúng để có thể tạo ra những luật quyết định. Có nhiều phương pháp được sử dụng sinh ra tập luật, tương ứng với các thuật toán được sử dụng sau: CART (Breiman cùng cộng sự, 1984), ID3 (Quinlan, 1986), C4.5 (Quinlan, 1993);

+ *Thuật toán Classification And Regression Trees (CART):* Sử dụng tư tưởng tham lam nhưng thay vì sử dụng các quy tắc dừng, thuật toán CART phát triển cây đến tối đa rồi sau đó tiến hành tỉa cây thông qua xác thực chéo sao cho các bước có mức độ lỗi là thấp nhất. Dựa trên ý tưởng này, CART sử dụng việc cắt nút luật yếu nhất tại các bước tỉa để giải quyết under-fitting và over-fitting. Thuật toán CART sử dụng giá trị "Gini index" để đánh giá độ đồng nhất giữa các nút để tiến hành cắt tỉa, giá trị này có công thức tính như sau:

$$Gini: \phi(t) = 1 - \sum_{j=1}^{n} \left(p^2(j|t)\right)$$

Trong đó $p(j|t)$ là xác suất của lớp j được đánh giá bởi nút t. Tại mỗi bước tỉa, ta tính chỉ số "Gini index" trong từng đặc trưng và lấy chọn đặc trưng có tổng trọng số bé nhất và tiến hành cắt tỉa.

+ *Thuật toán ID3 và C4.5:* Áp dụng tư tưởng tham lam, thuật toán C4.5 được JR. Quinlan đề xuất năm 1993 để mở rộng thuật toán ID3 trước đó cũng do chính JR. Quinlan đề xuất [12]. Là một thuật toán nhằm phân loại, cả hai thuật toán sử dụng khái niệm về entropy trong lý thuyết thông tin để đánh giá chọn lựa giữa các luật. C4.5 sử dụng khái hiện "Gain ratio" làm thước đo. Nếu tập X chứa m phần tử khác nhau, việc đánh giá được tiến hành: Giả sử bước phân tách nút t tiếp theo được có thể được phân tách thành các nút con $t_1, t_2, ..., t_n$, với $n(t)$ là số lượng các mẫu chứa trong nút t. Định nghĩa:

Entropy của t bất kỳ:

$$\phi(t) = -\sum_{j=1}^{n} (p(j|t) \times \log log\, p(j|t))$$

Entropy kỳ vọng sau khi phân tách của X, với $f_k(t) = \frac{n(t_k)}{n(t)}$

$$\phi_X(t) = -\sum_{k=1}^{r} \left(\phi(t_k) \times f_k(t)\right)$$

Thước đo lượng thông tin sau khi phân tách nút t

$$Infomation\ Gain:\ g(X) = \phi(t) - \phi_X(t)$$

Thước đo lượng thông tin về số lượng sau phân nhánh nút t:

$$Split\ info:\ h(X) = -\sum_{k} f_k(t) \times \log log\, f_k(t)$$

Khi này mức độ "Gain ratio" của X sẽ được tính bởi công thức:

$$Gain\ ratio\ = \frac{g(X)}{h(X)}$$

Việc sử dụng "Gain ratio" thay thế việc chỉ sử dụng cho "Infomation Gain" như thuật toán ID3 trước đó giúp loại bỏ các vấn đề về hiện tượng "over-fitting", dễ bị phụ thuộc vào các đặc trưng có số lượng dữ liệu lớn và bỏ qua các đặc trưng có số lượng dữ liệu bé nhưng ảnh hưởng lớn tới kết quả. C4.5 được cắt tỉa với một công thức heuristic.

2.1.2.2. Thuật toán rừng cây quyết định Random Forest

Được L. Breiman đinh nghĩa như sau [13]:

"Một rừng cây quyết định ngẫu nhiên (Random Forest) là một bộ phân loại bao gồm một tập hợp các cấu trúc cây phân loại $\{h\,(x,\,\Theta_k),\,k\,=\,1,2,...\}$ trong đó $\{\Theta_k\}$ là vectơ ngẫu nhiên độc lập có chung phân phối và $h\,(x,\,\Theta_k)$ là cây phân loại tương ứng được sinh ra bởi Θ_k,việc xác định lớp được tiến hành thông qua quá trình đánh giá của mỗi cây, kết quả của lớp mà đầu vào x được phân loại sẽ có số phiếu là nhiều nhất bỏ phiếu."

Ý tưởng của thuật toán có thể được mô tả lại như sau:

+ Từ mẫu ban đầu tách thành nhiều bộ mẫu ngẫu nhiên có cùng phân phối, xây dựng Tập hợp các cây CART tương ứng với các mẫu được sinh ra.

+ Các cây sẽ không được thực hiện cắt tỉa.

+ Giá trị dự đoán cuối cùng đối với một mẫu xác định sẽ là giá trị trung bình của dự đoán bởi toàn bộ các cây.

Thông qua tính ngẫu nhiên và số lượng cây được tạo, Random Forest vượt qua vấn đề về "over-fitting" khi càng nhiều cây được thêm vào.

2.1.2.3. Thuật toán phân loại Naive Bayes

Thuộc mô hình học có giám sát, Naive Bayes Classification (NBC) là một họ các thuật toán phân loại dựa trên việc sử dụng tính toán xác suất sử dụng định lý Bayes:

$$P(y|x) = \frac{P(y)P(y)}{P(x)}$$

Bài toán phân loại đặt ra: Phân chia các mẫu vector $x = (x_1, x_2, ..., x_n)$ vào các lớp Y=$\{y_1, y_2, ..., y_k\}$. Mô hình của thuật toán Naive Bayes đặt ra hai giả định: Thứ nhất, các đặc trưng $x_1, x_2, ..., x_n$ là ngang hàng và đều quan trọng như nhau trong việc quyết định mẫu x thuộc lớp y_k bất kỳ nào; thứ hai, các đặc trưng $x_1, x_2, ..., x_n$ độc lập thống kê, tức việc nhận biết một hoặc nhiều giá trị x_i không cho thông tin giúp nhận biết các giá trị còn lại. Quá trình phân loại được thực hiện bằng việc tính các xác suất cho từng phân lớp y:

$$P(x_1, x_2, ..., x_n),\ i = 1, 2,..., k$$

Sử dụng giả định thứ hai, các (x_i) là độc lập với nhau, sử dụng định lý Bayes ta có thể tính được:

$$P(x|y_i) = P(x_1 \cap x_2 \cap \cdots \cap x_n|y_i) = P(x_1|y_i)P(x_2|y_i)...P(x_n|y_i)$$

Ta có tỉ lệ thuận sau:

$$P(y_i|x) \propto P(y_i) \prod_{j=1}^{n} P(x_j|y_i)$$

Thông qua công thức, với $P\left(y_i\right)$, $P(x_j|y_i)$ đã biết, các thuật toán theo tư tưởng mô hình này cố gắng xác định phân lớp của mẫu bằng cách so sánh tìm $P(y_i|x)$ với xác suất lớn nhất để đưa ra quyết định.

Có thể thấy giả định thứ hai về tính độc lập hoàn toàn giữa các biến là không thể. Tuy nhiên trên thực tế, phương pháp này loại cho kết quả rất tốt.

Sau đây là một ví dụ thể hiện các bước thực hiện, của mô hình thuật toán học máy phân loại Naive Bayes, sử dụng bộ dữ liệu chơi quần vợt.

Bảng 2.1. Bộ dữ liệu: Chơi quần vợt (Play tennis dataset)

id	quang cảnh	nhiệt độ	độ ẩm	sức gió	đi chơi?
1	nắng	nóng	cao	yếu	không
2	nắng	nóng	cao	mạnh	không
3	âm u	nóng	cao	yếu	có
4	mưa	trung bình	cao	yếu	có
5	mưa	mát	bình thường	yếu	có
6	mưa	mát	bình thường	mạnh	không
7	âm u	mát	bình thường	mạnh	có
8	nắng	trung bình	cao	yếu	không
9	nắng	mát	bình thường	yếu	có
10	mưa	trung bình	bình thường	yếu	có
11	nắng	trung bình	bình thường	mạnh	có
12	âm u	trung bình	cao	mạnh	có
13	âm u	nóng	bình thường	yếu	có
14	mưa	trung bình	cao	mạnh	không

Thực hiện đếm các chỉ số theo các quyết định của việc có đi chơi hay không, ta có các bảng đếm giá trị cùng tỉ lệ như sau:

Bảng 2.2. Các thông số đếm, tỉ lệ xác suất tương ứng

		có	không	có	không
quang cảnh	nắng	2	3	2/9	3/5
	âm u	4	0	4/9	0/5
	mưa	3	2	3/9	2/5
nhiệt độ	nóng	2	3	2/9	2/5
	trung bình	4	0	4/9	2/5
	mát	3	2	3/9	1/5
độ ẩm	cao	3	4	3/9	4/5

	bình thường	6	1	6/9	1/5
		có	không	có	không
sức gió	yếu	6	2	6/9	2/5
	mạnh	3	3	3/9	3/5
đi chơi		9	5	9/14	5/14

Thực hiện dự đoán phân lớp $Y = \{có, không\}$ đối với một ngày mới như sau: quang cảnh nắng, nhiệt độ mát, độ ẩm cao, sức gió mạnh hay tương đương với $x = (nắng, mát, cao, mạnh)$. Áp dụng công thức:

$$P(có|x) \propto P(có) \times P(nắng|có) \times P(mát|có) \times P(cao|có) \times P(mạnh|có)$$

$$\propto \frac{9}{14} \times \frac{2}{9} \times \frac{3}{9} \times \frac{3}{9} \times \frac{3}{9} = 0.0053$$

Tương tự, ta cũng có:

$$P(không|x) \propto \frac{5}{14} \times \frac{3}{5} \times \frac{1}{5} \times \frac{4}{5} \times \frac{3}{5} = 0.0206$$

Tiến hành chuẩn hóa:

$$P(có|x) = 0.0053 / (0.0053 + 0.0206) = 0.205$$

$$P(không|x) = 0.0206 / (0.0053 + 0.0206) = 0.795$$

Vậy xác suất của không đi chơi cao hơn, bộ phân loại sẽ trả về kết quả với $x = (nắng, mát, cao, mạnh)$, thì mẫu sẽ được phân loại $y = không$.

2.2. Khảo sát một số giải pháp phát hiện tấn công thay đổi giao diện website dựa trên bất thường

2.2.1. Một số giải pháp phát hiện tấn công thay đổi giao diện website dựa trên bất thường

Kim cùng cộng sự [14] đề xuất một phương pháp phát hiện dựa trên n-gram sử dụng để phát hiện các trang web động bị tấn công thay đổi giao diện với cơ chế bảo vệ từ xa. Để thực hiện điều này, cần xây dựng "hồ sơ" phát hiện từ các trang web hoạt động bình thường. Tiếp đó, nội dung HTML