|  |  |  |
| --- | --- | --- |
|  | BAN CƠ YẾU CHÍNH PHỦ  “HỌC VIỆN KỸ THUẬT MẬT MÃ” | Mẫu 2 |

BÁO CÁO CHUYÊN ĐỀ SỐ 4.3.1

“Các tiêu chí và phương pháp đánh giá đặc trưng”

NHIỆM VỤ: “Nghiên cứu và ứng dụng nền tảng học sâu để xây dựng hệ thống phát hiện mã độc trực tuyến”.

Mã số: 06/2022/CB.

Cơ quan chủ trì: Học viện Kỹ thuật Mật mã

Chủ nhiệm: ThS. Lê Đức Thuận

Hà Nội - 2023

|  |  |  |
| --- | --- | --- |
|  | BAN CƠ YẾU CHÍNH PHỦ  “HỌC VIỆN KỸ THUẬT MẬT MÃ” |  |

BÁO CÁO CHUYÊN ĐỀ SỐ 4.3.1

“Các tiêu chí và phương pháp đánh giá đặc trưng”

NHIỆM VỤ: “Nghiên cứu và ứng dụng nền tảng học sâu để xây dựng hệ thống phát hiện mã độc trực tuyến”.

Mã số: 06/2022/CB.

Cơ quan chủ trì: Học viện Kỹ thuật Mật mã

Chủ nhiệm: ThS. Lê Đức Thuận

|  |  |
| --- | --- |
| **Người thực hiện chuyên đề** | **Cơ quan chủ trì** |
| *(Họ tên và chữ ký)* | *(Họ tên và chữ ký)* |

Hà Nội - 2023

MỤC LỤC

[MỤC LỤC 1](#_Toc129697163)

[DANH MỤC HÌNH ẢNH 2](#_Toc129697164)

[CÁC TIÊU CHÍ VÀ PHƯƠNG PHÁP ĐÁNH GIÁ ĐẶC TRƯNG MÃ ĐỘC 3](#_Toc129697165)

[1.1. Trích lọc đặc trưng (Feature Extraction) 3](#_Toc129697166)

[1.2. Biến đổi đặc trưng 4](#_Toc129697167)

[1.3. Lựa chọn đặc trưng 5](#_Toc129697168)

[1.4. Các phương pháp đánh giá để lựa chọn đặc trưng 6](#_Toc129697169)

[1.5. Tiêu chí đánh giá 8](#_Toc129697170)

# DANH MỤC HÌNH ẢNH

[Hình 1: Trích lọc đặc trưng Autoencoder 3](#_Toc129694167)

[Hình 2: Loại bỏ dữ liệu nhiễu 6](#_Toc129694168)

[Hình 3: Phương phpá lựa chọn đặc trưng 6](#_Toc129694169)

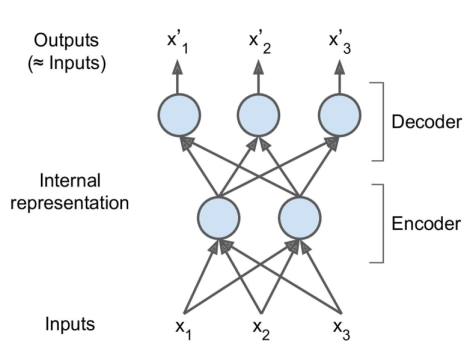
[Hình 4: Phương pháp Filter 7](file:///D:\DT_Ban%202022\bản%20word\Báo%20cáo%20tiến%20độ\Nội%20dung%204\BC_CĐ%204.3.1.%20Các%20tiêu%20chí%20và%20phương%20pháp%20đánh%20giá%20đặc%20trưng.docx#_Toc129694170)

[Hình 5: Phương pháp Wrapper 8](file:///D:\DT_Ban%202022\bản%20word\Báo%20cáo%20tiến%20độ\Nội%20dung%204\BC_CĐ%204.3.1.%20Các%20tiêu%20chí%20và%20phương%20pháp%20đánh%20giá%20đặc%20trưng.docx#_Toc129694171)

# CÁC TIÊU CHÍ VÀ PHƯƠNG PHÁP ĐÁNH GIÁ ĐẶC TRƯNG MÃ ĐỘC

## Trích lọc đặc trưng (Feature Extraction)

Ở những bộ dữ liệu cao chiều thì huấn luyện mô hình và dự báo cần tiêu tốn rất nhiều chi phí tính toán. Chính vì thế *trích lọc đặc trưng* là một kĩ thuật giúp giảm chiều giữ liệu mà ở đó cho phép chúng ta lựa chọn hoặc kết hợp các biến đầu vào thành những *đặc trưng* dự báo nhưng vẫn thể hiện một cách chính xác và nguyên vẹn của dữ liệu gốc. *Trích lọc đặc trưng* được áp dụng trong nhiều bài toán khác nhau của machine learning.

* Autoendcoder: Là kĩ thuật khá hiệu quả trong *self - supervised learning*. Kĩ thuật này sẽ tự mã hoá dữ liệu đầu từ không gian cao chiều sang một không gian thấp chiều (quá trình *encoder*). Sau đó giải mã ngược lại từ không gian thấp chiều sang không gian cao chiều (quá trình *decoder*) sao cho thông tin đầu ra của quá trình giải mã và đầu vào phải gần bằng nhau.

Hình 1: Trích lọc đặc trưng Autoencoder

* Bag-of-Words: Hay còn gọi là *thuật toán túi từ* thường được sử dụng trong *xử lý ngôn ngữ tự nhiên (Natural Language Processing - NLP)* và *trích lọc thông tin (information retrieval)*. Thuật toán cho phép chúng ta trích lọc thông tin từ các đoạn văn bản, mẩu tin, trang web bằng cách xây dựng một *túi từ* và tìm cách mã hoá nội dung văn bản thành một véc tơ tần suất của từ mà không quan tâm đến thứ tự của từ và cấu trúc ngữ pháp.
* Image Processing: Đây là những thuật toán được sử dụng để phát hiện đặc trưng trên ảnh như hình dạng (*shaped*) và cạnh (*edges*). Đó có thể là những phương pháp trích lọc đặc trưng trên ảnh thủ công như [HOG](https://phamdinhkhanh.github.io/2019/11/22/HOG.html) và [SHIFT](http://luthuli.cs.uiuc.edu/~daf/courses/ComputerVisionTutorial2012/EdgesOrientationHOGSIFT-2012.pdf) hoặc sử dụng bộ trích lọc đặc trưng thông qua tích chập [CNN](https://phamdinhkhanh.github.io/2019/08/22/convolutional-neural-network.html).

## Biến đổi đặc trưng

Biến đổi đặc trưng là những kĩ thuật giúp biến đổi dữ liệu đầu vào thành những dữ liệu phù hợp với mô hình nghiên cứu. Những dữ liệu này thường có tương quan cao đối với biến mục tiêu và do đó giúp cải thiện độ chính xác của mô hình. Bên dưới là một số phương pháp chính được áp dụng trong biến đối đặc trưng:

* **Chuẩn hóa biến**: Chuẩn hoá biến nhằm mục đích tạo ra sự đồng nhất đơn vị giữa các biến đầu vào và giảm thiểu những tác động xấu lên mô hình do sự khác biệt về độ lớn giữa các biến. Các kĩ thuật liên quan đến chuẩn hoá đơn vị cho biến đầu vào còn được gọi là Feature Scaling bao gồm: Chuẩn hoá MinMax (Minmax scaling), chuẩn hoá độ dài đơn vị (Unit length scaling), chuẩn hoá phân phối chuẩn (Standardization).
* **Biến đổi biến theo hàm**: Trong trường hợp dữ liệu có phương sai thay đổi (heteroscedasticity) thì chúng ta có thể sử dụng một số hàm biến đổi biến đầu vào để tạo ra những biến có phương sai ổn định và dạng phân phối gần với phân phối chuẩn hơn như logrith, căn bậc 2, căn bậc 3.
* **Tạo biến tương tác**: Các biến tương tác là những biến kết hợp từ nhiều biến đầu vào chẳng hạn như  Biến tương tác có thể là tích của hai hoặc nhiều biến. Trong một mô hình có ít biến đầu vào thì sử dụng biến tương tác có thể giúp tạo ra nhiều biến giải thích mới giúp ích cho mô hình.
* **Tạo biến bậc cao**: Biến bậc cao là những biến được tạo thành từ biến đầu vào bằng cách luỹ thừa với giá trị bậc cao, có thể là bậc 2, 3,… Chẳng hạn với biến đầu vào là x1 thì biến bậc cao của nó là 
* **Dữ liệu về vị trí địa lý**: Từ vị trí địa lý có thể suy ra vùng miền, thành thị, nông thôn, mức thu nhập trung bình, các yếu tố về nhân khẩu,…
* **Dữ liệu thời gian**: Các dữ liệu chuỗi thời gian thường tồn tại tính chu kì và mùa vụ. Chính vì vậy, các kĩ thuật biến đổi biến thời gian thành đặc trưng ghi nhận tính chất chu kì và mùa vụ sẽ giúp tăng cường khả năng giải thích của mô hình đối với biến mục tiêu. Chúng ta có thể lựa chọn chu kì của thời gian là buổi sáng/chiều/tối trong ngày; ngày trong tháng; tuần trong tháng; tháng trong năm hoặc quí trong năm tuỳ theo qui luật mùa vụ được thể hiện ở biến mục tiêu.

## Lựa chọn đặc trưng

Feature Selection là một phương pháp giảm số lượng các biến thông tin đầu vào trong mô hình Machine Learning của bạn, bằng cách chỉ sử dụng những dữ liệu liên quan, có ý nghĩa và loại bỏ các dữ liệu nhiễu.

Feature Selection là quá trình tự động chọn các tính năng liên quan đến mô hình học máy của bạn, dựa trên vấn đề mà bạn cần giải quyết.

Chúng ta thực hiện công việc này bằng cách gom lại hoặc loại trừ các tính năng không ảnh hưởng đến kết quả đầu ra. Điều này giúp giảm dữ liệu nhiễu và giảm lượng dữ liệu đưa vào ban đầu.

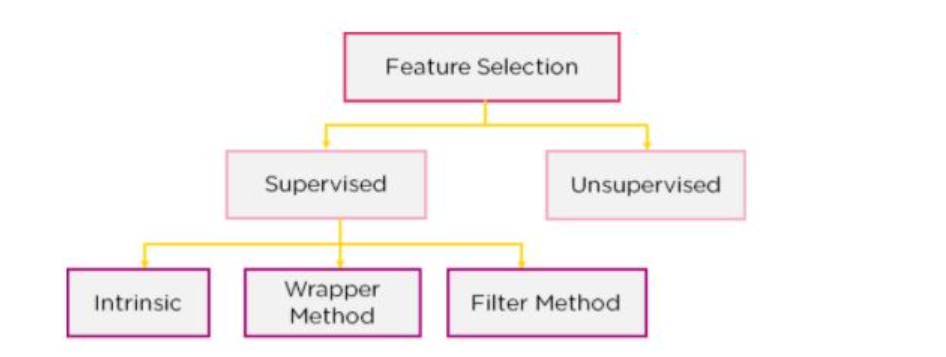


Hình 2: Loại bỏ dữ liệu nhiễu

## Các phương pháp đánh giá để lựa chọn đặc trưng

Có 2 loại mô hình chính:

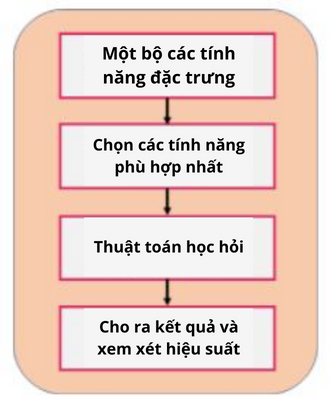
1. **Mô hình được giám sát:** Sử dụng lớp nhãn đầu ra để lựa chọn đối tượng. Hệ thống này sử dụng các biến mục tiêu để chọn ra các biến có thể tăng tính hiệu quả cho mô hình
2. **Mô hình không giám sát:**Không cần lớp nhãn đầu ra để lựa chọn đối tượng.



Hình 3: Phương phpá lựa chọn đặc trưng

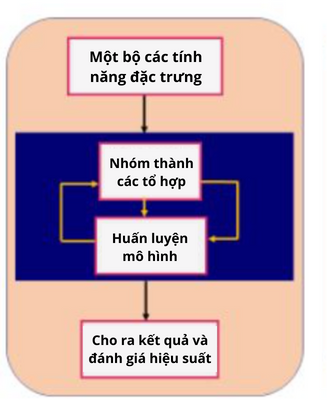
Với mô hình được giám sát, chúng ta có thể chia thành 3 mô hình nhỏ nữa:

* Bộ lọc (Filter Method): Trong phương pháp này, các tính năng bị loại bỏ dựa trên mối quan hệ của chúng với đầu ra, hay nói cách khác là cách chúng tương quan với đầu ra. Chúng ta sẽ xem xét các tính năng đặc trưng đó có tương quan tích cực hay tiêu cực với các nhãn đầu ra không, sau đó loại bỏ các tính năng không phù hợp



Hình 4: Phương pháp Filter

* Phương pháp Wrapper: Chúng ta sẽ chia dữ liệu thành các tập hợp con và huấn luyện hệ thống Machine Learning. Dựa trên kết quả đầu ra của mô hình, chúng ta sẽ thêm và bớt các tính năng, sau đó huấn luyện lại mô hình.
* Phương pháp Intrinsic: Là phương pháp kết hợp cả 2 phương pháp nêu trên để tạo thành một tập hợp các tính năng đặc trưng tốt nhất



Hình 5: Phương pháp Wrapper

## Tiêu chí đánh giá

Feature Selection là quá trình giảm số lượng features khi phát triển các ML model. Việc làm này xuất phát từ 2 tiêu chí:

* Giảm thời gian tính toán, huấn luyện mô hình.
* Tăng hiệu năng của các mô hình.

Feature Selection, đôi khi hay bị hiểu nhầm thành Dimensionality Reduction. Thực ra, chúng đều là các công việc phải thực hiện trong quá trình chuẩn bị dữ liệu cho các ML model. Chúng vừa có sự giống nhau, vừa có sự khác nhau:

* Giống: Về bản chất, 2 cái đều giảm số lượng chiều của dữ liệu, từ cao xuống thấp.
* Khác: Feature Selection giảm chiều dữ liệu bằng cách loại bớt một số features không liên quan đi, chỉ giữ lại các features liên quan nhất đến mục đích cần dự đoán. Trong khi đó, Dimensionality Reduction giảm chiều dữ liệu bằng cách ánh xạ tất cả các features hiện có sang một miền không gian có số chiều nhỏ hơn miên không gian ban đầu. Đối với Feature Selection, các features còn lại đều là các features cũ, nằm trong số các features ban đầu. Còn đối với Dimentionality Reduction, các features còn lại đều là các features mới hoàn toàn.