TRƯỜNG ĐẠI HỌC MỞ THÀNH PHỐ HỒ CHÍ MINH

**KHOA CÔNG NGHỆ THÔNG TIN**

---🙡 🕮 🙣---



**BÀI TẬP MÔN HỌC**

**KHAI PHÁ DỮ LIỆU**

**Đề tài**

**FEATURE SELECTION**

**Sinh viên thực hiện: Phùng Tri An**

**Mã số sinh viên: 1851050002**

**Sinh viên thực hiện: Châu Chí Hạo**

**Mã số sinh viên: 1851050041**

**Lớp: DH18TH01**

**Giảng viên hướng dẫn: TS. Nguyễn Tiến Đạt**

**Tháng 05 năm 2021**

**MỤC LỤC**

[Chương 1. GIỚI THIỆU về đề tài 3](#_Toc14692274)

[**1.** TÌNH TRẠNG DỮ LIỆU TRƯỚC ĐÂY 3](#_Toc14692275)

[**2.** TÓM TẮT NHU CẦU CẦN CÓ PHƯƠNG PHÁP LỰA CHỌN DỮ LIỆU (FEATURE SELECTION) 3](#_Toc14692276)

[Chương 2. Feature selection và các phương pháp 5](#_Toc14692277)

[**1.** PHƯƠNG PHÁP LỌC (FILTER APPROACH) 5](#_Toc14692278)

[**2.** PHƯƠNG PHÁP GÓI (WRAPPER APPROACH) 14](#_Toc14692279)

[**3.** PHƯƠNG PHÁP NHÚNG (EMBEDDED APPROACH) 16](#_Toc14692280)

[Chương 3. Kết luận 17](#_Toc14692281)

[**1.** KẾT LUẬN 17](#_Toc14692282)

# Chương 1. Giới Thiệu Về Đề Tài

1. Tình trạng dữ liệu trước đây

Hàng ngàn thập kỷ qua, con người thu thập dữ liệu qua mắt nhìn, qua các hiện tượng vật lý hoặc qua được truyền từ người qua người. Họ ghi chép chồng chất một lượng dữ liệu lớn không theo một quy tắc hay công tắc gì cả. Tính khả năng của của các dữ liệu không cao dẫn đến việc tìm kiếm áp dụng vào công việc và cuộc sống trở nên khó khăn và tốn kém rất nhiều.

Vì vậy một trong những thách thức lớn hiện nay là xử lý, phân loại và phân tích dữ liệu một cách có khoa học. Nhu cầu cần có các phương pháp để giải quyết các vấn đề dữ liệu ngày một nhiều. Do đó, phương pháp lựa chọn dữ liệu là một trong những phương pháp được nghiên cứu và ứng dụng rất nhiều.

1. Tóm tắt nhu cầu cần có phương pháp lựa chọn dữ liệu (feature selection)

Các dữ liệu trước đây có tính khả dụng rất cao, rất nhiều nguồn tài nguyên cần được lựa chọn và phân tách. Vài năm trước, việc phân tích dữ liệu trở nên khó khăn hơn theo cách thủ công.

Nhưng với công nghệ 4.0 ngày nay việc phân tích dữ liệu đều được đơn giản hóa bằng khai phá dữ liệu. Khai phá dữ liệu giúp cho tìm kiếm và nạp dữ liệu ẩn trên các cơ sở mẫu, giúp cho dữ liệu được phân tách ra từng phàn rõ ràng, giảm đi sự trùng lắp không cần thiết giữa các mô hình. Đa phần dữ liệu đều là dữ liệu thô và thu thập một cách rời rạc không thống nhất và chứa các tệp tin với dung lượng lưu trữ lớn. Vì vậy, sự xuất hiện của phương pháp lựa chọn thuộc tính (feature selection)

Lựa chọn thuộc tính (feature selection) là các quy trình cho lựa chọn nhiều tập hợp con có tính năng liên quan với nhau, tối đa hóa chính xác chính xác phân loại, là quá trình bỏ đi những tính năng dư thừa ra khỏi dữ liệu gốc. Do đó feature selection đóng vai trò quan trọng trong các hệ thống phát hiện xâm nhập.

Lựa chọn thuộc tính trở thành nhiều đề tài nghiên cứu và phát triển hơn vì nó cung cấp nhiều giá trị về việc phân loại dữ liệu một cách chính xác, tối ưu và tiết kiệm nhiều kinh phí trong việc áp dụng vào thực tiễn.

Để nghiên cứu về lựa chọn thuộc tính (feature selection) chúng em chia thành 3 cách tiếp cận sau:

* Phương pháp lọc (filter approach).
* Phương pháp bọc (wrapper approach).
* Phương pháp nhúng (embedded approach).

# Chương 2. Feature Selection Và Các Phương Pháp

1. Phương pháp lọc (filter approach)

Phương pháp lọc được thể hiện trong sơ đồ sau:

Trình bày cho người dùng

Tập hợp tất cả các tính năng

Chọn những tập con tốt nhất

Đưa vào thuật toán học

Ưu điểm của phương pháp lọc (fliter approach)

Chọn tính năng mà không phụ thuộc vào loại phân loại được sử dụng.

Độc lập với phân loại được sử dụng do đó chỉ cần thực hiện một lần chỉ với cách tiếp cận này.

Nhược điểm của phương pháp lọc (Feature approach)

Bỏ qua sự tương tác với bộ phân loại và các thuộc tính phụ.

Mỗi tính năng được xem xét riêng biệt

Xếp hạng các kỹ năng thuật toán của phương pháp lọc đucợ sử dụng là một trong các tiêu chí nguyên tắc để tiếp cận dễ dàng hơn. Các phương pháp thuật toán này khá đơn giản, bỏ đi quá nhiều sự phụ thuộc giữa các tính năng. Vì vập, tập con có thể không được tối ưu và có thể thu nhập các tập con dưa thừa. Có các thuật toán và cách kiểm tra sau để giảm bớt đi những nhược điểm trên như sau:

* 1. Chi-square test

Mục đích chi-square test là để đưa là số liệu cụ thể giữa 2 luồng dữ liệu độc lập. So sánh 2 luồng dữ lậu độc lập có liên quan với nhau hay không.

Ví dụ: Tại một thị trần có các bệnh nhận với tình trạng bệnh lý khác nhau. Xem coi giữa giới tính với bệnh lý đó có liên quan với nhau hay không? Lúc đó ta kiểm định bằng Chi bình phương theo theo công thức sau:

(1)

Trong đó:

* E là giá trị kỳ vọng (expected value).
* là giá trị quan sát (observed value).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Bệnh lý | | | |
|  |  | Tiểu đường | Đau dạ dày | Máu khó đông | Tổng |
| Giới tính | Nam | 30 | 15 | 20 | 65% |
| Nữ | 20 | 5 | 10 | 35% |
|  | Xác suất | 50% | 20% | 30% |  |

Từ bảng số liệu trên ta tính ngược lại các số kỳ vọng (expected value). Ta lấy xác suất bệnh lý nhân cho xác suất của giới tính. Lưu ý P(A&B) = P(A) \* P(B) nếu A và B độc lập với nhau. Nếu kỳ vọng như quan sát thì 2 dữ liệu hoặc biến không liên quan, ngược lại sẽ liên quan với nhau.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  |  | Bệnh lý | | | |
|  |  | Tiểu đường | Đau dạ dày | Máu khó đông | Tổng |
| Giới tính | Nam | 32.5 | 13.0 | 19.5 | 65% |
| Nữ | 17.5 | 7.0 | 9.0 | 35% |
|  | Xác suất | 50% | 20% | 30% |  |

Từ đó ta áp dụng công thức (1) ra kết quả x2 0,96.

Tính giá trị bậc tự do: (3 - 1) \* (2 – 1) = 2 (Số 3 ở đây là do có 3 bệnh lý và 2 giới tính).

Tính sig ta lấy excel hàm CHIDIST(0.96, 2)  0,62 (Xác định tính đúng đắn của giả thuyết ban đầu đưa ra).

Nếu sig lớn hơn 5% ta có thể xác nhận 2 dữ liệu hoàn toàn độc lập với nhau. 2 dữ liệu hoàn toàn không liên quan. Nếu nhỏ hơn 5% thì dữ liệu có mối quan lệ với nhau và sẽ sử dụng các thuật toán khác.

* 1. Khoảng cách Euclid.

Ở cách tính này, mối quan hệ giữa các tính năng, thuộc tính được tính toán ở công thức Euclid. Nếu a chứa n thì số lượng tính năng n được so sánh với các tính năng n – 1 khác bằng cách tính khoảng cách giữa chúng bằng công thức sau:

* 1. Hệ số tương quan (correlation coefficient)

Hệ số tương quan (correlation coefficient) là chỉ số thống kê đo lường mức độ mạnh yếu của mối quan hệ giữa hai biến số.

Hệ số tương quan có giá trị từ -1 đến 1 có nghĩa là có lỗi trong phép do tương quan.

Hệ số tương quan âm có giá trị âm cho thấy hai biến có giá trị âm hoặc có mối quan hệ nghịch biến.

Hệ số tương quan dương có giá trị dương cho thấy hai biến có giá trị dương hoặc có mối quan hệ đồng biến.

Tương quan bằng 0 là 2 dữ liệu biến độc lập với nhau.

Công thức tính hệ số tương quan (correlation coefficient):

Trong đó :

* : Hệ số tương quan Pearson
* Cov(x,y): Hiệp phương sai của biến x và y

Độ lệch chuẩn thể hiện độ phân tán dữ liệu so với mức trung bình.

Hiệp phương sai thể hiện mối quan hệ tuyến tính giữa hai biến.

Ví dụ:

Thống kê về mối tương quan cho biết nhà đầu tư xác định tương quan giữa hai biến có thay đổi hay không? Trong ngân hàng người ta thường cổ phiếu ngân hàng từ đó xác định lãi suất cho vay dựa vào sự thay đổi dữ liệu của giá cổ phiếu. Từ đó có thể suy xét về nội bộ hoạt động trong ngân hàng:

Nếu giá cổ phiếu giảm mà lãi suất tăng thì chắc ngân hàng đó có vấn đề trong tài chính hoặc nghi vấn nội bộ khác.

Nếu giá cổ phiếu của các ngành khác cũng tăng theo ta có thể suy đoán rằng cổ phiếu ngân hàng giảm không phải do lãi suất, có thể do quy chế hoạt động kém, hoặc nội bộ gặp vấn đề,…

* 1. Information Gain

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Day | Outlook | Temp | Humidity | Wind | Play tennis |
| 1 | Sunny | Hot | High | Weak | No |
| 2 | Sunny | Hot | High | Strong | No |
| 3 | Overcast | Hot | High | Weak | Yes |
| 4 | Rain | Mild | High | Weak | Yes |
| 5 | Rain | Cool | Normal | Weak | Yes |
| 6 | Rain | Cool | Normal | Strong | No |
| 7 | Overcast | Cool | Normal | Strong | Yes |
| 8 | Sunny | Mild | High | Weak | No |
| 9 | Sunny | Cool | Normal | Weak | Yes |
| 10 | Rain | Mild | Normal | Weak | Yes |
| 11 | Sunny | Mild | Normal | Strong | Yes |
| 12 | Overcast | Mild | High | Strong | Yes |
| 13 | Overcast | Hot | Normal | Weak | Yes |
| 14 | Rain | Mild | High | Strong | No |

Bảng dữ liệu tham khảo từ Gv.Nguyễn Đình Thuân – ĐH CNTT

Định nghĩa Entropy

Entrogy đặc trưng cho độ bất định hỗn tạp của tập bất kỳ.

Entropy(S) = - p**+**log2p**+** - p**-**log2p**-**

S là tập hợp các mẫu thuộc lớp âm và dương.

P+ là tỷ lệ các mẫu thuộc lớp dương trong S.

P- là tỷ lệ các mẫu thuộc lớp âm trong S.

Ví dụ: Từ kết quả Play tennis, 9 giá trị yes và 5 giá trị âm, tương ứng với 9 giá trị dương và 5 giá trị âm. Ký hiệu ([9+, 5-])

Entropy([9+, 5-]) = -(9/14)log2(9/14) – (5/14)log2(5/14) = 0.94

**Information gain** để đo sự rút giảm trong Entropy, phản ảnh mức độ hiệu quả của một thuộc tính trong phân lớp. Sự rút giảm mong mún của entropy gây ra bởi sự phân hoạch được dựa theo công thức:

Trong đó Value(A) là tập các giá trị cho thuộc tính A.

Sv là tập con của S mà A nhận được giá trị v.

Ví dụ: Tính Gain(S, Wind)

Values(Wind) = {Weak, Strong}, S = [9+, 5-]

Sweak là nút con với giá trị “Weak” là [6+, 2-]

Sstrong là nút con với giá trị “Strong” là [3+, 3-]

Gain(S, Wind) = Entropy(S) – ((8/14)Entropy(Sweak) – (6/14)Entropy(Sstrong))

= 0.94 – ((8/14)0.811 – (6/14)1.0) = 0.048

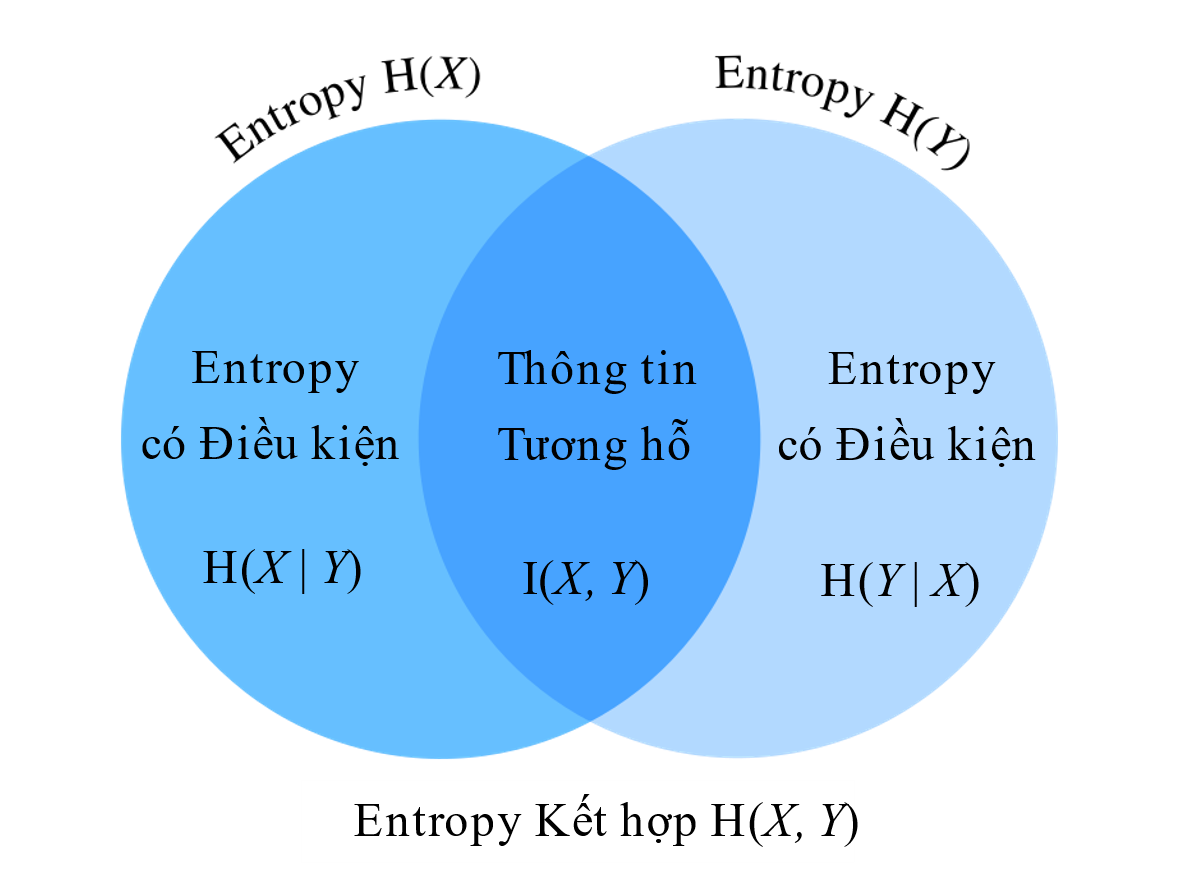
* 1. Mutual Information

Có 2 biến ngẫu nhiên X và Y. Liệu có bao nhiêu dữ liệu nằm trong Y mà không trong X, có nhiêu được bao nhiêu thông tin có trong X và Y?

Để biết đuợc điều đó thì thông tin tin tương hỗ (mutual information) là một kỹ năng tuật toán giúp phân tách điều này.

Mục đích của mutual information là để tìm ra các thông tin giống nhau giữa các biến ngẫu nhiên. Ví dụ cụ thể trong 2 biến ngẫu nhiên X và Y , bỏ đi các thông tin khác nhau giữa hai dữ liệu . Thông tin chứa cả X lẫn Y là H(X, Y). Thông tin nằm trong X mà không nằm trong Y được H(X|Y), ngược lại thông tin nằm trong Y mà không nằm trong X được ký hiệu là H(Y|X). Từ đó, ta có công thức cho mutual information như sau:

I(X, Y) = H(X, Y) – H(X|Y) – H(Y|X)

Dưới đây là mô hình minh họa cho mutual information:

Hình 1: Mối quan hệ giữa mutual information với entropy có điều kiện và entropy không có điều kiện

Tính chất của mutual information:

* Mutual information có tính đối xứng I(X, Y) = I(Y, X).
* Mutual information không có giá trị âm I(X, Y) ≥ 0.
* I(X, Y) = 0 khi X và Y là hai biến độc lập.
* X là hàm nghịch đảo của Y, thì Y và X có chung toàn bộ dữ liệu khi đó
* I(X, Y) = H(Y) = H(X)

Ví dụ về mutual information

Khi có một thông tin dữ liệu về B.M.W bị bắt. Khi đó chúng ta sẽ có hai luồng phân tích là tập đoàn xe hơi BMW bị bắt vì lý do gì đó, hay đó là tên viết tắt của tên tội phạm mới bị. Nhờ mutual information ta có thể dễ dang phân biệt được luồng dữ liệu nào đúng. Nhóm một có thông tin tương hỗ tới tập đoàn xe hơi BMW như các trang xe hơi, máy móc, kỹ thuật điện tử, lắp ráp,… Nhóm hai có thông tin từ tên tội phạm có tên viết tắt là B.M.W, chẳng hạn như trên các bìa báo tội phạm, trang công an tình báo, phòng chống tội phạm, hành vi,…Từ đó, ta có thể so sánh nhóm nào xuất hiện nhiều hơn trong ngữ cảnh này.

* 1. Correlation based Feature Selection (CFS)

Correlation based Feature Selection (CFS) là kỹ năng đo lường tính hữu ích của các tính năng riêng lẻ để dư đoán dữ liệu đó sẽ nằm trong phấn lớp nào cùng với mức độ quan hệ tương quan giữa chúng. Các tính năng không liên quan để bỏ đi. Lọc đi các tính năng dư thừa dẫn đến dự đoán kém được định nghĩa trong công thức sau:

Ví dụ: Nhà bán hàng mới ra mắt dòng sản phẩm điện thoại mới. Họ thường lấy các dữ liệu về xu hướng thị trường khách hàng họ ưu chuộng những tính năng nào nhất. Từ đó nhà sản xuất sẽ đánh mạnh hơn về tính năng đó, hạn chế đi những tính năng khác. Quảng cáo ra thị trường các tính năng của điện thoại đó để kích thích người mua tăng doanh thu.

* 1. Fast Correlation-Based Filter (FCBF)

Thuật toán FCBF được chia thành 2 giai đoạn: Phân tích mức độ liên quan và phân tích dự phòng

Phân tích mức độ liên quan

Phân tích dữ liệu tương quan được sự dụng rộng rãi ở mức độ phù hợp. Sự tương quan có thể được đo lường các hệ số tuyến tính sau:

Đa phần các hệ thống trong ứng dụng đều không có tuyến tính. Sự tương quan trong hệ thống có thể đo bằng cách sử dụng Symmetrical Uncertainty (SU):

I(X, Y) = H(X) – H(X|Y)

Trong đó

* I(X|Y) là information gain của X khi quan sát Y.
* H(X), H(Y) lần lượt là entropy của biến X và Y.
* P(xi) là xác suất của biến x.

SU đo các information gain nếu kết quả từ 0 đến 1 thì FCBF sẽ loại bỏ các tính năng không cần thiết bằng cách xếp hạng độ tương quan (SU) giữa các tính năng và phân lớp của nó. Nếu SU = 1 là tính năng này hoàn toàn liên quan với phân lớp đó. Nếu SU = 0 tính năng này hoàn toàn độc lập với lớp đó.

Phân tích dự phòng

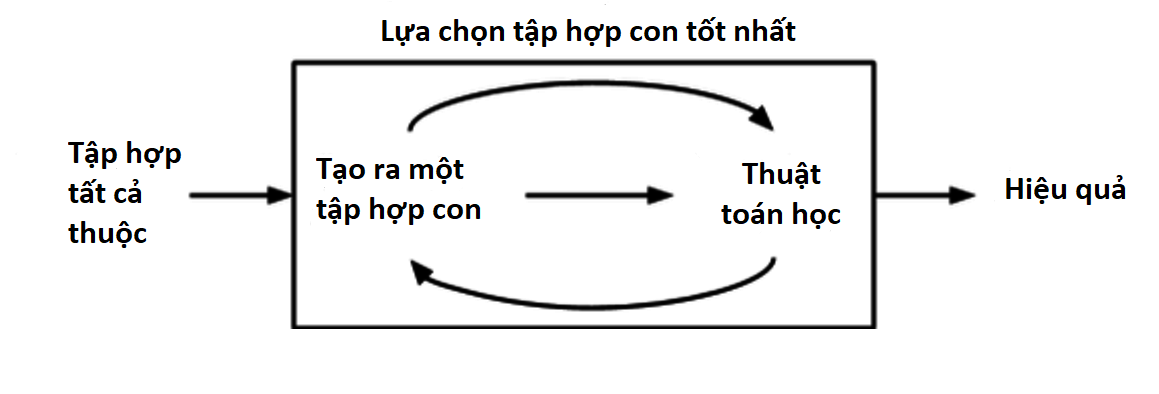
Sau khi xếp hạng các tính năng có liên quan, FCBF loại bỏ các tính năng dư thừa nhờ SU khi xếp hạng giữa tính năng và hạng và giữa tính năng và tính năng. Phân tích dự phòng có thể được định nghĩa từ ý nghĩa của tính năng chiếm ưu thế được gọi là Markov Banket (cả tính năng liên quan và không dư thừa).

Ta xét hai đặc điểm có liên quan Fi và Fj . Fj tạo thành vùng Markov Banket gần đúng cho Fi nếu:

SUj,c ≥ Sui,c và SUi,j ≥ SUi,c

Trong đó

* Sui,c là mối tương quan giữa bất kỳ đối tượng nào với lớp.
* SUi,j là mối tương quan bất kỳ của Fi và Fj.

1. Phương pháp gói (wrapper approach)
   1. Phương pháp gói:

* Xác định tốt hơn các thuộc tính tối ưu thay vì đơn thuần là các thuộc tính có liên quan.
* Thực thi bằng cách suy nghiệm thuật toán học và tập huấn luyện.
* Sử dụng phép khử ngược để loại bỏ các thuộc tính không đáng kể từ tập con.
* SVM-RFE là một trong những thuật toán lựa chọn thuộc tính sử dụng phương pháp gói.
* Cần một số thuật toán học được định nghĩa trước để xác định thuộc tính có liên quan.
* Có tương tác với thuật toán phân loại.
* Thuộc tính khớp quá được tránh sử dụng kiểm chứng chéo.
* Mặc dù tính toán tốn kém và mất nhiều thời gian hơn so với phương pháp lọc, nhưng chúng cho kết quả chính xác hơn phương pháp lọc.
* Một lợi thế khác là nó duy trì sự phụ thuộc giữa các thuộc tính và tập con thuộc tính.
  1. Thuật toán lựa chọn thuộc tính tiếp tới (forward feature selection) bắt đầu bằng một tập hợp trống và thêm thuộc tính tốt nhất vào tập hợp. Quá trình này được lặp lại cho đến khi nhận được các thuộc tính tốt nhất.

Ví dụ : {A1, A2, A3, A4, A5, …}

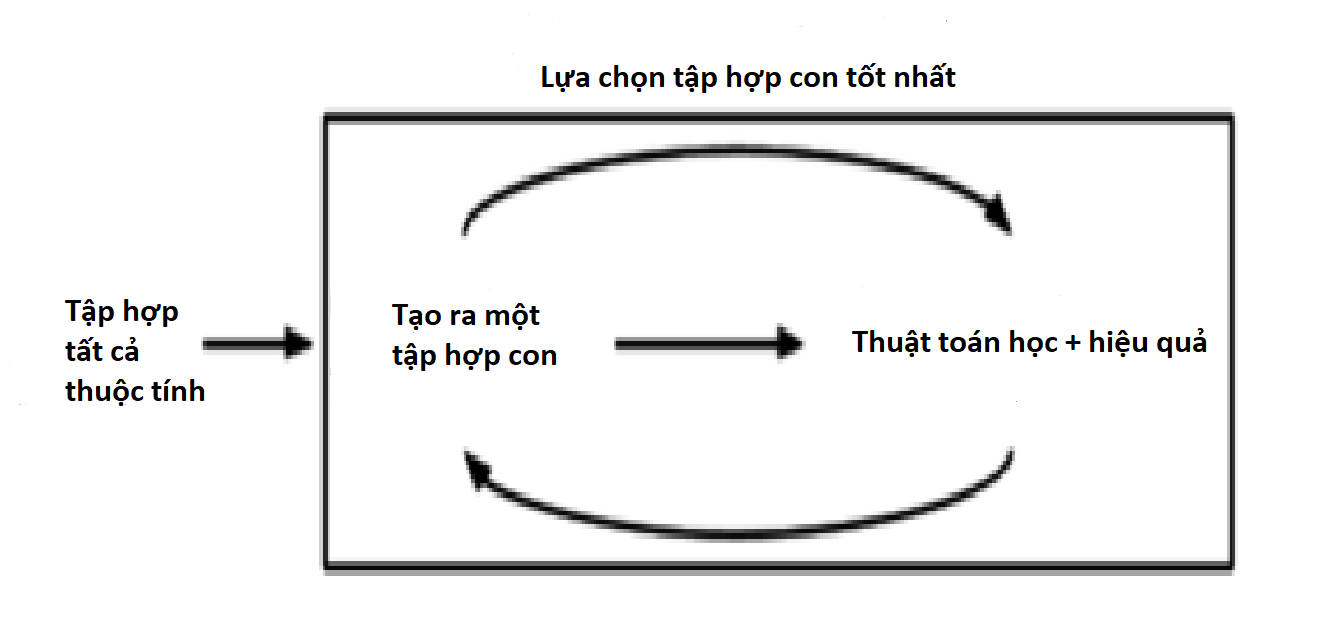
{}

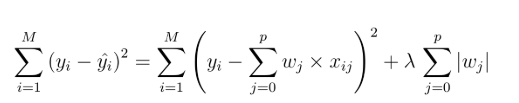
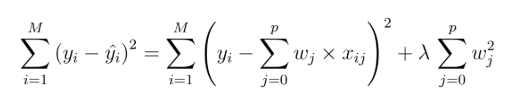
* {A1}
* {A1, A3}
* {A1, A3, A4}
* …
  1. Thuật toán loại bỏ thuộc tính ngược (backward feature elimination) bắt đầu từ toàn bộ tập hợp thuộc tính và loại bỏ thuộc tính xấu nhất tại thời điểm. Quá trình này được lặp lại cho đến khi loại bỏ các thuộc tính xấu nhất.

Ví dụ : {A1, A2, A3, A4, A5, …}

{A1, A2, A3, A4, A5, …}

* {A2, A3, A4, A5, …}
* {A2, A4, A5, …}
* …

1. Phương pháp nhúng (embedded approach)
   1. Phương pháp nhúng:

* Là phương pháp lựa chọn thuộc tính được tích hợp vào một thuật toán học và được tối ưu hóa.
* Còn được gọi là mô hình lai kết hợp giữa phương pháp lọc và phương pháp gói.
* Giảm thời gian tính toán cho việc phân loại lại các tập con khác nhau được thực hiện trong phương pháp gói nhưng cho kết quả kém chính xác vì một số thuộc tính quan trọng có thể bị mô hình lọc làm mất.
  1. Hồi quy Lasso
  2. Hồi quy Ridge

**Chương 3: Kết Luận**

1. Kết luận:

* Lựa chọn thuộc tính là một vấn đề quan trọng trong phân loại bởi vì nó có thể có ảnh hưởng đáng kể đến độ chính xác của phân loại. Nó làm giảm số chiều của tập dữ liệu, do đó việc sử dụng bộ xử lý và bộ nhớ giảm; dữ liệu trở nên dễ hiểu và dễ nghiên cứu hơn.
* Phương pháp lựa chọn thuộc tính, phương pháp lọc nên được sử dụng để cho kết quả trong thời gian ngắn hơn và các tập dữ liệu lớn. Nếu muốn kết quả chính xác và tối ưu, nên sử dụng phương pháp gói.

**Hết**