**SELECT FEATURES**

**1.Tại sao cần phải Select fueatures:**

-lựa chọn tính năng như một kỹ thuật giảm kích thước,nhằm mục đích chọn một tập hợp con nhỏ của các đối tượng có liên quan từ các đối tượng địa lý gốc bằng cách loại bỏ các đối tượng địa lý không liên quan,dư thừa hoặc nhiễu .Việc lựa chọn các tính năng giúp cho hiệu suất được cải thiện,độ chính xác mô hình cao hơn,chi phí tính toán trở lên đơn giản,khả năng diễn giải mô hình được tốt hơn.

🡪mục đích :tối đa hoá mức độ liên quan và giảm sự dư thừa các tính năng.

**2.select features**

Tất cả các quy trình công việc học máy phụ thuộc vào kỹ thuật tính năng, bao gồm trích xuất tính năng và lựa chọn tính năng là những khối xây dựng cơ bản của đường ống học máy hiện đại. Mặc dù thực tế là các quy trình trích xuất đối tượng địa lý và lựa chọn đối tượng địa lý có một số điểm trùng lặp, nhưng các thuật ngữ này thường bị đánh đồng một cách sai lầm. Khai thác tính năng là quá trình sử dụng kiến ​​thức miền để trích xuất các biến mới từ dữ liệu thô giúp các thuật toán học máy hoạt động. Quá trình lựa chọn tính năng dựa trên việc lựa chọn các tính năng nhất quán, phù hợp và không dư thừa.

Các mục tiêu của kỹ thuật lựa chọn tính năng bao gồm:

* đơn giản hóa các mô hình để giúp các nhà nghiên cứu / người dùng giải thích chúng dễ dàng hơn
* thời gian đào tạo ngắn hơn
* tránh lời nguyền của chiều
* tổng quát hóa nâng cao bằng cách giảm trang bị quá mức (chính thức, giảm phương sai)

**3.Các kỹ thuật,phương pháp lựa chọn tính năng(select features)**

**3.1.Phương pháp lọc**

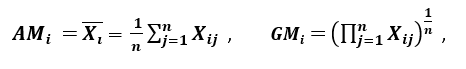
Trong Phương pháp Bộ lọc, các tính năng được chọn trên cơ sở các thước đo thống kê. Phương pháp này không phụ thuộc vào thuật toán học và chọn các tính năng như một bước tiền xử lý.

Phương pháp bộ lọc lọc ra tính năng không liên quan và các cột dư thừa khỏi mô hình bằng cách sử dụng các số liệu khác nhau thông qua xếp hạng.

Ưu điểm của việc sử dụng phương pháp bộ lọc là nó cần thời gian tính toán thấp và không trang bị quá nhiều dữ liệu.(overfitting)

Các kỹ thuật trong phương pháp này:

* Information Gain: xác định việc giảm entropy trong khi chuyển đổi tập dữ liệu. Nó có thể được sử dụng như một kỹ thuật lựa chọn đặc trưng bằng cách tính toán mức tăng thông tin của mỗi biến đối với biến mục tiêu.
* Chi-square-test: là một kỹ thuật để xác định mối quan hệ giữa các biến phân loại. Giá trị chi bình phương được tính toán giữa mỗi đối tượng địa lý và biến mục tiêu, và số lượng đối tượng địa lý mong muốn có giá trị chi bình phương tốt nhất được chọn.
* Fisher’s Score: là một trong những kỹ thuật lựa chọn đối tượng địa lý được giám sát phổ biến. Nó trả về thứ hạng của biến trên tiêu chí của ngư dân theo thứ tự giảm dần. Sau đó, chúng ta có thể chọn các biến có số điểm của ngư dan(một tỷ lệ giữa phương sai giữa lớp và phương sai trong lớp) lớn.
* Correlation Coefficient: Tương quan là thước đo mối quan hệ tuyến tính của 2 hoặc nhiều biến. Thông qua mối tương quan, chúng ta có thể dự đoán biến này so với biến khác. Logic đằng sau việc sử dụng tương quan để lựa chọn đối tượng địa lý là các biến tốt có tương quan cao với mục tiêu. Hơn nữa, các biến nên tương quan với mục tiêu nhưng không được tương quan với nhau.
* Variance Threshold: là một cách tiếp cận cơ sở đơn giản để lựa chọn đối tượng địa lý. Nó loại bỏ tất cả các tính năng mà phương sai không đáp ứng một số ngưỡng. Theo mặc định, nó loại bỏ tất cả các đối tượng không có phương sai, tức là các đối tượng có cùng giá trị trong tất cả các mẫu. Chúng tôi giả định rằng các đối tượng địa lý có phương sai cao hơn có thể chứa nhiều thông tin hữu ích hơn, nhưng lưu ý rằng chúng tôi không tính đến mối quan hệ giữa các biến đối tượng hoặc biến đối tượng và biến mục tiêu, đây là một trong những hạn chế của phương pháp lọc.
* Mean Absolute Difference (MAD): 'Chênh lệch tuyệt đối trung bình (MAD) tính chênh lệch tuyệt đối từ giá trị trung bình. Sự khác biệt chính giữa phương sai và các phép đo MAD là không có bình phương trong phương sai sau. MAD, giống như phương sai, cũng là một biến thể tỷ lệ. ' Điều này có nghĩa là MAD càng cao thì sức mạnh phân biệt càng cao.
* Dispersion ratio:một phép đo độ phân tán áp dụng giá trị trung bình số học(AM) và giá trị trung bình hình học(GM).AM và GM và RM được tính bởi:



RM

*có thể được sử dụng như một biện pháp phân tán. Độ phân tán cao hơn có nghĩa là giá trị Ri cao hơn, do đó một đặc điểm phù hợp hơn Ngược lại, khi tất cả các mẫu đối tượng có (gần đúng) cùng một giá trị, Ri gần bằng 1, cho thấy đối tượng địa lý có mức độ liên quan thấp.*

Graphical user interface, text

Description automatically generated with medium confidence**3.2 Phương pháp gói:**

-Trình bao bọc yêu cầu một số phương pháp để tìm kiếm không gian của tất cả các tập con tính năng có thể có, đánh giá chất lượng của chúng bằng cách tìm hiểu và đánh giá một bộ phân loại với tập hợp con tính năng đó. Quá trình lựa chọn tính năng dựa trên một thuật toán học máy cụ thể mà ta đang cố gắng phù hợp với một tập dữ liệu nhất định. Nó tuân theo một cách tiếp cận tìm kiếm tham lam bằng cách đánh giá tất cả các kết hợp có thể có của các tính năng so với tiêu chí đánh giá.

Các phương pháp trình bao bọc thường mang lại độ chính xác dự đoán tốt hơn các phương pháp bộ lọc.

Các kỹ thuật lựa chọn tính năng trong phương pháp này:

* Forward selection: là một quá trình lặp đi lặp lại, bắt đầu với một tập hợp các tính năng trống. Sau mỗi lần lặp lại, nó tiếp tục thêm vào một tính năng và đánh giá hiệu suất để kiểm tra xem nó có đang cải thiện hiệu suất hay không. Quá trình tiếp tục cho đến khi việc bổ sung một biến / tính năng mới không cải thiện hiệu suất của mô hình.
* Backward elimination: là một cách tiếp cận lặp đi lặp lại, nhưng nó ngược lại với chọn lọc về phía trước. Kỹ thuật này bắt đầu quá trình bằng cách xem xét tất cả các tính năng và loại bỏ tính năng ít quan trọng nhất. Quá trình loại bỏ này tiếp tục cho đến khi việc loại bỏ các tính năng không cải thiện hiệu suất của mô hình.
* Exhaustive Feature Selection: là một trong những phương pháp lựa chọn tính năng tốt nhất, đánh giá từng bộ tính năng là bạo lực. Nó có nghĩa là phương pháp này cố gắng & tạo ra từng sự kết hợp có thể có của các tính năng và trả về tập hợp tính năng hoạt động tốt nhất.
* Recursive Feature Elimination: là một cách tiếp cận tối ưu hóa tham lam đệ quy, trong đó các tính năng được chọn bằng cách lấy một cách đệ quy một tập con nhỏ hơn của các tính năng. Giờ đây, một công cụ ước tính được đào tạo với từng bộ tính năng và tầm quan trọng của từng tính năng được xác định bằng cách sử dụng coef\_attribute hoặc thôn qua  feature\_importances\_attribute.

Graphical user interface

Description automatically generated with medium confidence**3.3.Phương pháp nhúng:**

-Các phương pháp nhúng đã kết hợp những ưu điểm của cả phương pháp bộ lọc và bộ bao bọc bằng cách xem xét sự tương tác của các tính năng cùng với chi phí tính toán thấp. Đây là các phương pháp xử lý nhanh tương tự như phương pháp lọc nhưng chính xác hơn phương pháp lọc.

-Các phương pháp này cũng có tính lặp lại, đánh giá mỗi lần lặp và tìm ra một cách tối ưu các tính năng quan trọng nhất đóng góp nhiều nhất cho việc đào tạo trong một lần lặp cụ thể.

-Các kỹ thuật lựa chọn tính năng trong phương pháp này:

* Regularization: Quy định hóa thêm một điều khoản phạt cho các tham số khác nhau của mô hình học máy để tránh trang bị quá nhiều trong mô hình. Điều khoản phạt này được thêm vào các hệ số; do đó nó thu nhỏ một số hệ số về không. Những đối tượng địa lý có hệ số bằng không có thể bị xóa khỏi tập dữ liệu. Các loại kỹ thuật điều hòa là Điều hòa L1 (Điều chỉnh Lasso) hoặc Nets đàn hồi (điều hòa L1 và L2).
* Random Forest Importance: Các phương pháp lựa chọn đối tượng địa lý dựa trên cây khác nhau giúp ta đánh giá tầm quan trọng của đối tượng địa lý để đưa ra cách chọn đối tượng địa lý. Ở đây, tầm quan trọng của đối tượng địa lý chỉ định đối tượng địa lý nào có tầm quan trọng hơn trong việc xây dựng mô hình hoặc có tác động lớn đến biến mục tiêu. Random Forest là một phương pháp dựa trên cây như vậy, là một loại thuật toán đóng gói tập hợp một số lượng cây quyết định khác nhau. Nó tự động xếp hạng các nút theo hiệu suất của chúng hoặc giảm tạp chất (tạp chất Gini) trên tất cả các cây. Các nút được sắp xếp theo các giá trị tạp chất, và do đó nó cho phép cắt tỉa những cây bên dưới một nút cụ thể. Các nút còn lại tạo ra một tập hợp con của các tính năng quan trọng nhất.

**4.Phương pháp lựa chọn kỹ thuật phù hợp:**

**Graphical user interface

Description automatically generated with medium confidence**