Cluster_Hierarchical

May 22, 2024

0.0.1 Phân cụm phân cấp

Ý tưởng chính và bài toán phân cụm phân cấp

1. **Phân cụm phân cấp là gì và tại sao nó được sử dụng?**
2. **Những bài toán nào thường được giải quyết bằng phân cụm phân cấp?**
3. **Ý tưởng chính của phân cụm phân cấp là gì?**
4. **Tại sao phân cụm phân cấp được gọi là phân cụm phân cấp?**
5. **Hai loại chính của phân cụm phân cấp là gì và chúng khác nhau như thế nào?**
6. **Làm thế nào để xây dựng một cây phân cấp từ dữ liệu?**

Input và Output của phân cụm phân cấp

7. **Dữ liệu đầu vào cho phân cụm phân cấp yêu cầu những gì?**
8. **Dữ liệu đầu vào có cần phải chuẩn hóa không? Tại sao?**
9. **Phân cụm phân cấp tạo ra những output gì từ dữ liệu đầu vào?**
10. **Làm thế nào để xác định số lượng cụm từ cây phân cấp?**
11. **Dendrogram là gì và vai trò của nó trong phân cụm phân cấp?**
12. **Làm thế nào để đọc và diễn giải một dendrogram?**

Thuật toán triển khai phân cụm phân cấp

13. **Các bước chính của thuật toán phân cụm phân cấp là gì?**
14. **Phân cụm phân cấp liên kết đơn (single linkage) là gì?**
15. **Phân cụm phân cấp liên kết đầy đủ (complete linkage) là gì?**
16. **Phân cụm phân cấp liên kết trung bình (average linkage) là gì?**
17. **Làm thế nào để chọn phương pháp liên kết trong phân cụm phân cấp?**
18. **Độ đo khoảng cách nào thường được sử dụng trong phân cụm phân cấp?**
19. **Sự khác biệt giữa phân cụm phân cấp kết hợp (agglomerative) và phân cụm phân cấp phân chia (divisive) là gì?**

Ưu điểm và nhược điểm của phân cụm phân cấp

20. **Những ưu điểm chính của phân cụm phân cấp là gì?**
21. **Những nhược điểm của phân cụm phân cấp là gì?**
22. **Phân cụm phân cấp có thể gặp vấn đề gì khi xử lý dữ liệu lớn?**
23. **Làm thế nào để phân cụm phân cấp xử lý dữ liệu ngoại lai?**
24. **Tại sao phân cụm phân cấp không yêu cầu xác định số lượng cụm trước?**
25. **Phân cụm phân cấp có thể cải thiện bằng cách sử dụng các kỹ thuật nào khác?**

Ứng dụng và ví dụ của phân cụm phân cấp

26. **Một số ứng dụng thực tế của phân cụm phân cấp trong phân tích dữ liệu là gì?**
27. **Làm thế nào phân cụm phân cấp có thể được sử dụng trong phân tích gene?**
28. **Phân cụm phân cấp có thể giúp gì trong việc phát hiện cụm trong dữ liệu thị trường tài chính?**
29. **Một ví dụ cụ thể về việc sử dụng phân cụm phân cấp trong xử lý văn bản là gì?**
30. **Làm thế nào phân cụm phân cấp có thể được áp dụng trong phân tích dữ liệu khách hàng?**

Đánh giá output thu được từ phân cụm phân cấp

31. **Làm thế nào để đánh giá chất lượng của các cụm trong phân cụm phân cấp?**
32. **Cophenetic correlation coefficient là gì và nó đánh giá gì?**
33. **Làm thế nào để xác định số lượng cụm tối ưu từ dendrogram?**
34. **Silhouette score có thể được sử dụng để đánh giá phân cụm phân cấp không?**

0.0.2 Ý tưởng chính và bài toán phân cụm phân cấp

1. **Phân cụm phân cấp là gì và tại sao nó được sử dụng?** Phân cụm phân cấp là một phương pháp phân cụm dữ liệu tạo ra một cây phân cấp (dendrogram) biểu diễn mối quan hệ giữa các điểm dữ liệu. Nó được sử dụng để hiểu và trực quan hóa cấu trúc phân cấp trong dữ liệu.

2. **Những bài toán nào thường được giải quyết bằng phân cụm phân cấp?** Phân cụm phân cấp thường được sử dụng trong các bài toán như phân tích gene, phân tích dữ liệu khách hàng, xử lý văn bản, và trực quan hóa dữ liệu.

3. **Ý tưởng chính của phân cụm phân cấp là gì?** Ý tưởng chính của phân cụm phân cấp là xây dựng một cây phân cấp, trong đó mỗi nút lá biểu diễn một điểm dữ liệu và mỗi nút không lá biểu diễn một cụm các điểm dữ liệu.

4. **Tại sao phân cụm phân cấp được gọi là phân cụm phân cấp?** Nó được gọi là phân cụm phân cấp vì nó xây dựng một cây phân cấp (dendrogram) biểu diễn mối quan hệ phân cấp giữa các điểm dữ liệu.

5. **Hai loại chính của phân cụm phân cấp là gì và chúng khác nhau như thế nào?** Hai loại chính của phân cụm phân cấp là phân cụm phân cấp kết hợp (agglomerative) và phân cụm phân cấp phân chia (divisive). Phân cụm kết hợp bắt đầu bằng việc xem mỗi điểm dữ liệu là một cụm riêng lẻ và sau đó hợp nhất các cụm lại với nhau, trong khi phân cụm phân chia bắt đầu bằng một cụm toàn bộ và sau đó chia nhỏ nó ra.

6. **Làm thế nào để xây dựng một cây phân cấp từ dữ liệu?** Một cây phân cấp được xây dựng bằng cách tính toán khoảng cách giữa các điểm dữ liệu và sau đó hợp nhất các cụm gần nhất lại với nhau cho đến khi tất cả các điểm dữ

liệu được hợp nhất vào một cụm duy nhất hoặc khi đạt đến số lượng cụm mong muốn.

Input và Output của phân cụm phân cấp

7. **Dữ liệu đầu vào cho phân cụm phân cấp yêu cầu những gì?** Dữ liệu đầu vào cho phân cụm phân cấp yêu cầu một ma trận khoảng cách hoặc một tập hợp các điểm dữ liệu để tính toán khoảng cách giữa các điểm.

8. **Dữ liệu đầu vào có cần phải chuẩn hóa không? Tại sao?** Có, dữ liệu đầu vào thường cần phải chuẩn hóa để đảm bảo rằng tất cả các biến có trọng số tương đương trong việc tính toán khoảng cách, tránh việc các biến có đơn vị lớn hơn chi phối kết quả phân cụm.

9. **Phân cụm phân cấp tạo ra những output gì từ dữ liệu đầu vào?** Phân cụm phân cấp tạo ra một cây phân cấp (dendrogram) và một phân cụm các điểm dữ liệu dựa trên cây phân cấp này.

10. **Làm thế nào để xác định số lượng cụm từ cây phân cấp?** Số lượng cụm có thể được xác định bằng cách cắt cây dendrogram tại một mức độ nào đó, sao cho các cụm được tạo ra có ý nghĩa và phù hợp với mục đích phân tích.

11. **Dendrogram là gì và vai trò của nó trong phân cụm phân cấp?** Dendrogram là một biểu đồ cây biểu diễn mối quan hệ phân cấp giữa các điểm dữ liệu. Nó giúp trực quan hóa quá trình hợp nhất các cụm và xác định số lượng cụm tối ưu.

12. **Làm thế nào để đọc và diễn giải một dendrogram?** Để đọc và diễn giải một dendrogram, bạn bắt đầu từ gốc cây và xem xét các nhánh cây. Các điểm dữ liệu nằm trên cùng một nhánh được xem là cùng một cụm. Khoảng cách giữa các nhánh biểu thị mức độ khác biệt giữa các cụm.

Thuật toán triển khai phân cụm phân cấp

13. **Các bước chính của thuật toán phân cụm phân cấp là gì?**
    - Tính toán khoảng cách giữa tất cả các điểm dữ liệu.
    - Khởi tạo mỗi điểm dữ liệu là một cụm riêng lẻ.
    - Lặp lại quá trình hợp nhất các cụm gần nhất cho đến khi đạt số lượng cụm mong muốn hoặc tất cả các điểm dữ liệu hợp nhất vào một cụm duy nhất.

14. **Phân cụm phân cấp liên kết đơn (single linkage) là gì?** Liên kết đơn (single linkage) xác định khoảng cách giữa hai cụm là khoảng cách ngắn nhất giữa bất kỳ cặp điểm nào, mỗi điểm thuộc một cụm khác nhau.

15. **Phân cụm phân cấp liên kết đầy đủ (complete linkage) là gì?** Liên kết đầy đủ (complete linkage) xác định khoảng cách giữa hai cụm là khoảng cách dài nhất giữa bất kỳ cặp điểm nào, mỗi điểm thuộc một cụm khác nhau.

16. **Phân cụm phân cấp liên kết trung bình (average linkage) là gì?** Liên kết trung bình (average linkage) xác định khoảng cách giữa hai cụm là trung bình khoảng cách giữa tất cả các cặp điểm, mỗi điểm thuộc một cụm khác nhau.

17. **Làm thế nào để chọn phương pháp liên kết trong phân cụm phân cấp?** Phương pháp liên kết được chọn dựa trên đặc điểm của dữ liệu và mục tiêu phân tích. Liên kết đơn thường tốt cho việc phát hiện các cụm dài và mỏng, trong khi liên kết đầy đủ và trung bình tốt hơn cho việc phát hiện các cụm có hình dạng đồng đều.

18. **Độ đo khoảng cách nào thường được sử dụng trong phân cụm phân cấp?** Các độ đo khoảng cách thường được sử dụng bao gồm khoảng cách Euclid, khoảng cách Manhattan,

và khoảng cách Cosine.

19. **Sự khác biệt giữa phân cụm phân cấp kết hợp (agglomerative) và phân cụm phân cấp phân chia (divisive) là gì?** Phân cụm phân cấp kết hợp bắt đầu với mỗi điểm dữ liệu là một cụm và hợp nhất chúng lại, trong khi phân cụm phân cấp phân chia bắt đầu với một cụm lớn và chia nó thành các cụm nhỏ hơn.

Ưu điểm và nhược điểm của phân cụm phân cấp

20. **Những ưu điểm chính của phân cụm phân cấp là gì?**
    - Không yêu cầu xác định số lượng cụm trước.
    - Có thể trực quan hóa cấu trúc phân cấp của dữ liệu.
    - Linh hoạt với nhiều phương pháp liên kết và độ đo khoảng cách.
21. **Những nhược điểm của phân cụm phân cấp là gì?**
    - Tốn nhiều thời gian và bộ nhớ với dữ liệu lớn.
    - Khó xác định số lượng cụm tối ưu từ dendrogram.
    - Nhạy cảm với nhiễu và ngoại lai trong dữ liệu.
22. **Phân cụm phân cấp có thể gặp vấn đề gì khi xử lý dữ liệu lớn?** Phân cụm phân cấp có độ phức tạp tính toán cao và yêu cầu nhiều bộ nhớ, làm cho việc xử lý dữ liệu lớn trở nên chậm chạp và không hiệu quả.
23. **Làm thế nào để phân cụm phân cấp xử lý dữ liệu ngoại lai?** Phân cụm phân cấp có thể bị ảnh hưởng bởi các điểm dữ liệu ngoại lai, dẫn đến việc hình thành các cụm không hợp lý. Có thể cần loại bỏ hoặc xử lý ngoại lai trước khi thực hiện phân cụm.
24. **Tại sao phân cụm phân cấp không yêu cầu xác định số lượng cụm trước?** Vì phân cụm phân cấp xây dựng một cây phân cấp, số lượng cụm có thể được xác định sau khi xem xét dendrogram, không cần xác định trước.
25. **Phân cụm phân cấp có thể cải thiện bằng cách sử dụng các kỹ thuật nào khác?**
    - Sử dụng các phương pháp gần đúng để giảm thời gian tính toán.
    - Kết hợp với các phương pháp giảm chiều dữ liệu để giảm kích thước dữ liệu đầu vào.
    - Sử dụng các kỹ thuật phát hiện và xử lý ngoại lai trước khi thực hiện phân cụm.

Ứng dụng và ví dụ của phân cụm phân cấp

26. **Một số ứng dụng thực tế của phân cụm phân cấp trong phân tích dữ liệu là gì?**
    - Phân tích gene và phát hiện cụm trong dữ liệu sinh học.
    - Phân tích dữ liệu khách hàng và xác định phân khúc thị trường.
    - Xử lý văn bản và phân loại tài liệu.
27. **Làm thế nào phân cụm phân cấp có thể được sử dụng trong phân tích gene?** Phân cụm phân cấp có thể được sử dụng để phân tích biểu hiện gene và phát hiện các cụm gene có biểu hiện tương tự.
28. **Phân cụm phân cấp có thể giúp gì trong việc phát hiện cụm trong dữ liệu thị trường tài chính?** Phân cụm phân cấp có thể phát hiện các cụm tài sản có hành vi giá tương tự, hỗ trợ trong việc xây dựng danh mục đầu tư và phân tích rủi ro.

29. **Một ví dụ cụ thể về việc sử dụng phân cụm phân cấp trong xử lý văn bản là gì?** Trong xử lý văn bản, phân cụm phân cấp có thể được sử dụng để phân loại các tài liệu thành các cụm chủ đề tương tự, hỗ trợ trong việc tổ chức và tìm kiếm thông tin.

30. **Làm thế nào phân cụm phân cấp có thể được áp dụng trong phân tích dữ liệu khách hàng?** Phân cụm phân cấp có thể phân tích dữ liệu khách hàng để xác định các phân khúc khách hàng với hành vi mua hàng tương tự, hỗ trợ trong việc phát triển chiến lược marketing và dịch vụ khách hàng.

Đánh giá output thu được từ phân cụm phân cấp

31. **Làm thế nào để đánh giá chất lượng của các cụm trong phân cụm phân cấp?**

    - Sử dụng các chỉ số đánh giá như Silhouette score, Davies-Bouldin Index.
    - So sánh với các phương pháp phân cụm khác.
    - Đánh giá trực quan bằng dendrogram và kiểm tra tính hợp lý của các cụm.

32. **Cophenetic correlation coefficient là gì và nó đánh giá gì?** Cophenetic correlation coefficient đo lường mức độ bảo toàn của khoảng cách giữa các điểm dữ liệu trong dendrogram so với khoảng cách ban đầu, giúp đánh giá chất lượng của cây phân cấp.

33. **Làm thế nào để xác định số lượng cụm tối ưu từ dendrogram?** Số lượng cụm tối ưu có thể được xác định bằng cách cắt cây dendrogram tại một mức độ sao cho các cụm được tạo ra có ý nghĩa và phù hợp với mục tiêu phân tích.

34. **Silhouette score có thể được sử dụng để đánh giá phân cụm phân cấp không?** Có, Silhouette score có thể được sử dụng để đánh giá chất lượng của các cụm trong phân cụm phân cấp bằng cách đo lường mức độ tách biệt và tính nhất quán của các cụm.