TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

VIỆN CÔNG NGHỆ THÔNG TIN VÀ KINH TẾ SỐ

BÀI THU HOẠCH:

MÔN: TRÍ TUỆ NHÂN TẠO

ĐỀ TÀI:

CHƯƠNG TRÌNH PHẦN MỀM CÀI ĐẶT THUẬT TOÁN PHÂN CỤM PHÂN HẠCH K-MEANS

Sinh viên thực hiện: Trần Hoàng Kim Anh

Mã sinh viên: 11218387

Lớp học phần: Trí tuệ nhân tạo(223)_02

Giảng viên: TS. Lưu Minh Tuấn

Hà Nội 2023

TRƯỜNG ĐẠI HỌC KINH TẾ QUỐC DÂN

VIỆN CÔNG NGHỆ THÔNG TIN VÀ KINH TẾ SỐ

BÀI THU HOẠCH:

MÔN: TRÍ TUỆ NHÂN TẠO

ĐỀ TÀI:

CHƯƠNG TRÌNH PHẦN MỀM CÀI ĐẶT THUẬT TOÁN PHÂN CỤM PHÂN HẠCH K-MEANS

Sinh viên thực hiện: Trần Hoàng Kim Anh

Mã sinh viên: 11218387

Lớp học phần: Trí tuệ nhân tạo(223)_02

Giảng viên: TS. Lưu Minh Tuấn

Hà Nội 2023

MỤC LỤC

LỜI NÓI ĐẦU

Trong thời đại kỹ thuật số hiện nay, dữ liệu và các loại dữ liệu tràn ngập từ mọi nguồn, từ dữ liệu nhỏ của các cửa hàng buôn bán nhỏ đến dữ liệu của các doanh nghiệp lớn. Chính vì thế việc phân tích dữ liệu đã trở thành một lĩnh vực quan trọng trong nhiều ngành công nghiệp và khoa học. Một trong những phương pháp phổ biến nhất để khám phá cấu trúc trong dữ liệu là phân cụm, một nhánh của lĩnh vực học máy không giám sát. Và trong số các thuật toán phân cụm, thuật toán K-means là một trong những thuật toán được sử dụng rộng rãi nhất do tính đơn giản, hiệu quả và khả năng mở rộng vượt trội của thuật toán này.

Phân cụm, phân hạch là quá trình sử dụng các phương pháp và thuật toán để sắp xếp các dữ liệu thành các cụm, các hạch có các đặc tính chung, hay nói cách khác là sắp thành các nhóm có các đặc tính gần giống nhau. Phân cụm, phân hạch không chỉ giúp dữ liệu trở nên dễ hiểu và dễ quản lý hơn mà còn giúp người sử dụng chúng khám phá ra các mô hình và xu hướng ẩn trong dữ liệu.

Các thuật toán phân cụm nổi tiếng có thể được kể đến như K-means, DBSCAN, và phân cụm phân cấp đều được sử dụng rộng rãi trong nhiều lĩnh vực như thị giác máy tính, khai thác dữ liệu và học máy. Mỗi thuật toán đều có những ưu điểm và nhược điểm riêng, và nên lựa chọn thuật toán phù hợp vào bản chất của dữ liệu và mục tiêu của việc phân cụm.

Trong bài nghiên cứu này, em chọn thuật toán K-means, một thuật toán phân cụm phổ biến do tính đơn giản và hiệu quả của nó. Bài nghiên cứu được kỳ vọng sẽ cung cấp những tri thức cần thiết về thuật toán phân cụm, phân hạch K-means, cách thức thuật toán hoạt động và cải thiện quá trình phân nhóm dữ liệu, phân tích dữ liệu và đưa ra dự đoán từ tệp dữ liệu có sẵn. Từ cơ sở lý thuyết, bài nghiên cứu sẽ cài đặt và chạy thử thuật toán phân cụm, phân hạch K-means để xử lý tệp dữ liệu có sẵn.

CHƯƠNG 1: GIỚI THIỆU ĐỀ TÀI

1. Đề tài nghiên cứu

Thế giới ngày càng phát triển, lượng thông tin được sinh ra và trao đổi ngày càng tăng lên, tạo ra một thách thức lớn trong vấn đề xử lý thông tin. Các thông tin được lưu trữ dưới nhiều hình thức khác nhau: có thể là các bài báo, tài liệu, thống kê hay các loại dữ liệu văn bản. Điều này đòi hỏi các tổ chức và cá nhân phải có giải pháp hiệu quả cho vấn đề quản lý và xử lý loại thông tin đa dạng này.

Và trong số các thuật toán xử lý dữ liệu thành các nhóm, các cụm, K-means là thuật toán đơn giản và hiệu quả nhất cho những người mới bắt đầu. K-means tập trung vào việc tối ưu hóa khoảng cách giữa các điểm dữ liệu và trung tâm cụm mà chúng thuộc về. Điều này giúp tạo ra các cụm dữ liệu có tính chất tương đồng cao, giúp cho việc phân tích và khai thác dữ liệu trở nên dễ dàng hơn.

2. Phạm vi nghiên cứu

Bài thu hoạch *Chương trình phần mềm cài đặt thuật toán phân cụm phân hạch K-means* sẽ nghiên cứu về khái niệm, đặc điểm, ứng dụng của K-means và so sánh K-means với các phương pháp, kỹ thuật phân cụm, phân hạch tương tự và đưa ra nhận xét. Dựa trên cơ sở lý thuyết đó, bài thu hoạch đã tìm hiểu, cài đặt và chạy thử thuật toán K-means từ một bộ dữ liệu có sẵn, với đầu ra là kết quả phân cụm, phân hạch từ dữ liệu đó. Phần lập trình của thuật toán này đã được đính kèm trong file python đi kèm bản báo cáo này.

3. Đối tượng nghiên cứu

Bài thu hoạch sẽ tập trung vào nghiên cứu thuật toán phân cụm, phân hạch K-means, cài đặt và chạy thử thuật toán. Nhờ sự phát triển của công nghệ và các cuộc cách mạng công nghệ, K-means càng trở thành thuật toán thu hút, nhất là trong nhánh xử lý dữ liệu. K-means là thuật toán thuộc nhánh học không giám sát của học máy.

4. Mục đích nghiên cứu

Bài nghiên cứu được tạo ra với mục đích tìm hiểu lý thuyết về thuật toán phân cụm K-means, nắm được các khái niệm chính, cài đặt và chạy thử thuật toán phân cụm, phân hạch K-means, từ đó nắm bắt được cách thức một thuật toán Trí tuệ nhân tạo hoạt động và ứng dụng trong thực tế đời sống.

CHƯƠNG 2: NỘI DUNG NGHIÊN CỨU

1. Cơ sở lý thuyết

1.1. Thuật toán K-means

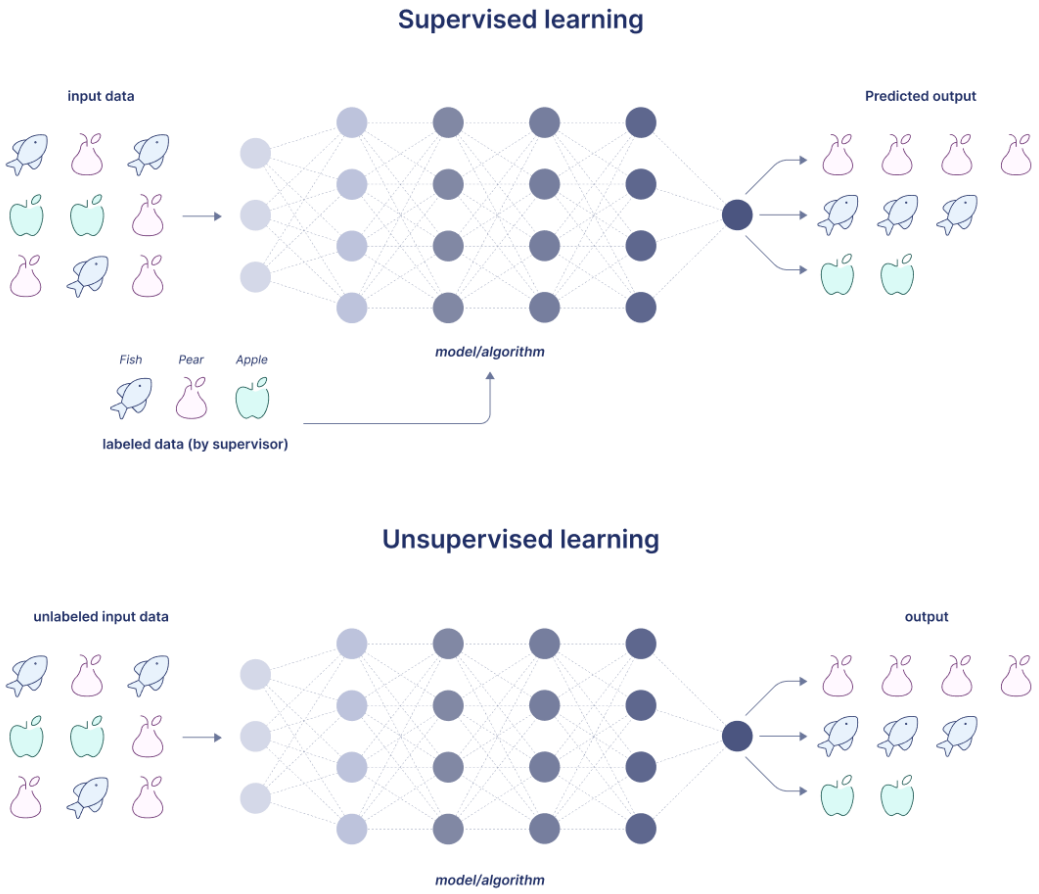
Thuật toán phân cụm, phân hạch K-means thuộc các thuật toán học máy không giám sát (Supervised learning), thuật toán này sẽ nhóm các dữ liệu chưa được gán nhãn thành các cụm khác nhau từ đó đưa ra các mô hình cụ thể để người sử dụng nắm bắt được cấu trúc dữ liệu và đưa ra các dự đoán cụ thể.

Cả Supervised learning (Học có giám sát) và Unsupervised learning (Học không giám sát) là những phương pháp kỹ thuật cơ bản của Machine Learning (Học máy). Học máy là một công nghệ phát triển từ lĩnh vực trí tuệ nhân tạo. Các thuật toán học máy là các chương trình máy tính có khả năng học hỏi về cách hoàn thành các nhiệm vụ và cách cải thiện hiệu suất theo thời gian. Học máy vẫn đòi hỏi sự đánh giá của con người trong việc tìm hiểu dữ liệu cơ sở và lựa chọn các kĩ thuật phù hợp để phân tích dữ liệu. Đồng thời, trước khi sử dụng, dữ liệu phải sạch, không có sai lệch và dữ liệu giả.

Supervised Learning (Học có giám sát) là một nhóm thuật toán sử dụng dữ liệu được gán nhãn nhằm mô hình hóa mối quan hệ giữa biến đầu vào (x) và biến đầu ra (y). Hai nhóm bài toán cơ bản trong học có giám sát là classification (phân loại) và regression (hồi quy), trong đó biến đầu ra của bài toán phân loại có các giá trị rời rạc trong khi biến đầu ra của bài toán hồi quy có các giá trị liên tục. Với Supervised Learning, bên cạnh xây dựng các mô hình mạnh, việc thu thập và gán nhãn dữ liệu tốt và hợp lý cũng đóng vai trò then chốt để giải quyết các bài toán trong thực tế.

Ngược lại, Unsupervised Learning (Học không giám sát) là một nhóm thuật toán sử dụng dữ liệu không có nhãn. Các thuật toán theo cách tiếp cận này hướng đến việc mô hình hóa được cấu trúc hay thông tin ẩn trong dữ liệu. Hay nói cách khác, sử dụng các phương pháp này thiên về việc mô tả tính chất hay đặc tính của dữ liệu. Thông thường, các thuật toán này dựa trên những thông tin sau:

- Mối quan hệ tương tự (similarity) giữa các ví dụ (được gọi là instance) trong dữ liệu như trong các thuật toán clustering (phân cụm)
- Xác suất đồng xuất hiện của các đối tượng như trong Association mining
- Các phép biến đổi ma trận để trích xuất các đặc trưng như PCA, SVD.



H2.1. So sánh giữa Học có giám sát và Học không giám sát

Thuật toán phân cụm, phân hạch K-means thuộc nhánh học máy không giám sát, sẽ gán các điểm dữ liệu vào một trong các cụm K tùy thuộc vào khoảng cách của các dữ liệu tới trung tâm của cụm. Thuật toán bắt đầu bằng cách gán ngẫu nhiên trọng tâm của các cụm trong không gian. Sau đó, mỗi điểm dữ liệu sẽ được gán vào cụm dựa theo khoảng cách của dữ liệu tới trọng tâm của các cụm. Sau khi gán mỗi điểm vào một trong số các cụm, một trọng tâm mới của cụm lại được gán. Quá trình này sẽ lập đi lập lại cho đến khi thuật toán tìm được cụm tốt nhất. Thông thường, số lượng cụm được cho trước bằng cách sử dụng một số thuật toán, và tất cả các điểm phải được gán vào trong cụm.

Trong một số trường hợp, số cụm K không được định nghĩa trước, và chúng ta phải tìm ra số lượng tối ưu nhất của K. Phân cụm K-means hoạt động tốt nhất khi dữ liệu được phân tách rõ ràng. Khi các điểm dữ liệu bị chồng chéo, thuật toán phân cụm, phân hạch sẽ không còn là thuật toán phù hợp nữa.

K-means là thuật toán phân cụm, phân hạch nổi bật và hoạt động nhanh hơn so với các thuật toán khác. Thuật toán này cung cấp sự kết hợp mạnh mẽ giữa các điểm dữ liệu. Cụm K-means không cung cấp thông tin rõ ràng về chất lượng giữa các cụm. Việc gán ban đầu các trọng tâm khác nhau có thể dẫn đến các cụm khác nhau. Ngoài ra, thuật toán K-means nhạy cảm với nhiễu, khiến thuật toán có thể mắc kẹt trong các cực tiểu cục bộ.

1.2. Mục đích của thuật toán phân cụm K-means

Mục tiêu của thuật toán phân cụm K-means là chia các dữ liệu hay tập hợp các điểm dữ liệu thành một nhóm sao cho các điểm dữ liệu trong mỗi nhóm có tính so sánh cao hơn với nhau và khác biệt so với các điểm dữ liệu trong các nhóm khác. Hay nói cách khác, đây là việc nhóm các dữ liệu thành các cụm (các nhóm) dựa trên sự tương đồng của chúng và khác biệt với các cụm khác.

1.3. Cách thức hoạt động của thuật toán phân cụm, phân hạch K-means

Giả sử được cung cấp một tệp dữ liệu của các mục, với các đặc trưng nhất định và giá trị cho các đặc trưng này (như là một vector). Nhiệm vụ là phân loại các mục này thành các nhóm. Để thực hiện điều này, ta sẽ sử dụng thuật toán K-means. K đại diện cho số lượng cụm/nhóm mà ta muốn phân loại dữ liệu của mình vào. Thuật toán sẽ phân loại các mục vào K nhóm hoặc cụm tương đồng. Để tính toán sự tương đồng đó, ta sẽ sử dụng khoảng cách Euclid làm phép đo.

Thuật toán sẽ hoạt động như sau:

1. Khởi tạo ngẫu nhiên K điểm – gọi là trung bình hay trọng tâm của các cụm.

2. Phân loại mỗi điểm dữ liệu vào cụm gần nhất, cập nhật tọa độ trung bình của dữ liệu được phân loại.

3. Lặp lại quá trình đó một số lần cho trước, sau khi kết thúc quá trình lặp lại, ta được các cụm với dữ liệu đã được phân loại.

Các điểm ở trên được gọi là trọng tâm (hay trung bình) vì chúng sẽ là giá trị trung bình của các mục sau khi đã được phân loại. Để khởi tạo các giá trị trọng tâm này, chúng ta có rất nhiều lựa chọn. Một trong các phương pháp trực quan là khởi tạo các trung bình tại các mục ngẫu nhiên trong tệp dữ liệu. Một phương pháp khác là khởi tạo các trung bình tại các giá trị ngẫu nhiên giữa các ranh giới của tập dữ liệu.

Ngoài ra, có thể xác định số cụm lý tưởng bằng sử dụng các thuật toán như phương pháp Elbow hay Silhouette:

- Phương pháp Elbow: Phương pháp này chọn một phạm vi các giá trị ứng viên của K, sau đó áp dụng phân cụm K-means bằng cách sử dụng từng giá trị của K. Tìm khoảng cách trung bình của mỗi điểm trong một cụm đến tâm của nó và biểu diễn nó trong một biểu đồ. Chọn giá trị của K, trong đó khoảng cách trung bình giảm đột ngột.
- Phương pháp Silhouette: Phương pháp này cũng là một phương pháp để tìm số tối ưu của các cụm và giải thích và xác nhận tính nhất quán trong các cụm dữ liệu. Phương pháp tính toán các hệ số hình bóng của mỗi điểm đo bao nhiêu điểm tương tự với cụm của chính nó so với các cụm khác.

1.4. Ví dụ về sử dụng thuật toán phân cụm, phân hạch K-means

1.4.1. Yêu cầu bài toán

Sử dụng thuật toán phân cụm, phân hạch K-means để phân loại hoa Iris. Tệp dữ liệu hoa Iris được cho sẵn trong thư viện sklearn của python. Đây là tập dữ liệu phân loại đa lớp cổ điển và phổ biến, bao gồm 150 mẫu từ ba loài Iris (Iris setosa, Iris virginica và Iris versicolor). Mỗi mẫu gồm bốn đặc trưng: chiều dài và chiều rộng của cánh hoa và lá.

1.4.2. Áp dụng thuật toán K-means

Đầu tiên, chúng ta sẽ nhập các thư viện Python cần thiết:

```
import pandas as pd
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
import matplotlib.cm as cm
from sklearn.datasets import load_iris
from sklearn.cluster import KMeans
```

Sau đó sẽ tiến hành nhập vào bộ dữ liệu hoa Iris:

```
x, y = load_iris(return_X_y=True)
```

Ta sẽ sử dụng Elbow Method để tìm ra số lượng nhóm/cụm lý tưởng để phân cụm dữ liệu:

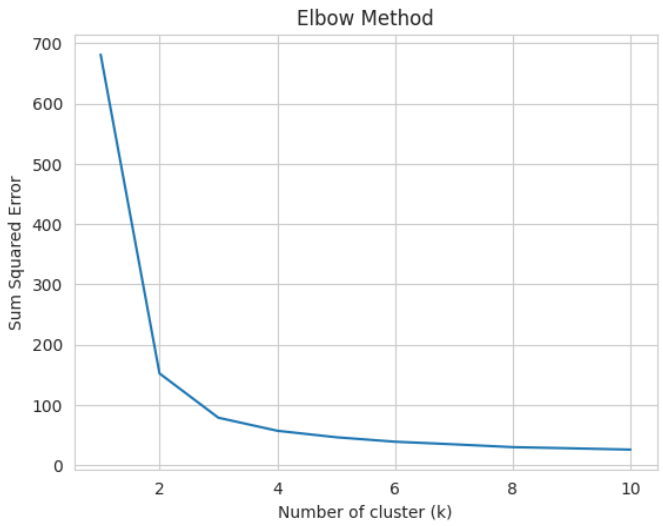
```
#Find optimum number of cluster
sse = [] #SUM OF SQUARED ERROR
for k in range(1,11):
    km = KMeans(n_clusters=k, random_state=2)
    km.fit(X)
    sse.append(km.inertia_)

sns.set_style("whitegrid")
g=sns.lineplot(x=range(1,11), y=sse)

g.set(xlabel ="Number of cluster (k)",
      ylabel = "Sum Squared Error",
      title ='Elbow Method')

plt.show()
```

Ta sẽ thu được đồ thị như sau:



Từ đồ thị trên, ta quan sát được tại k = 2 và k = 3 là hệ số lý tưởng để phân cụm. Trong trường hợp này, chúng ta sẽ xem xét hệ số k = 3.

Xây dựng mô hình phân cụm K-means:

```
kmeans = KMeans(n_clusters = 3, random_state = 2)
kmeans.fit(X)
```

Tìm trọng tâm các cụm:

```
kmeans.cluster_centers_
```

Kết quả trả về:

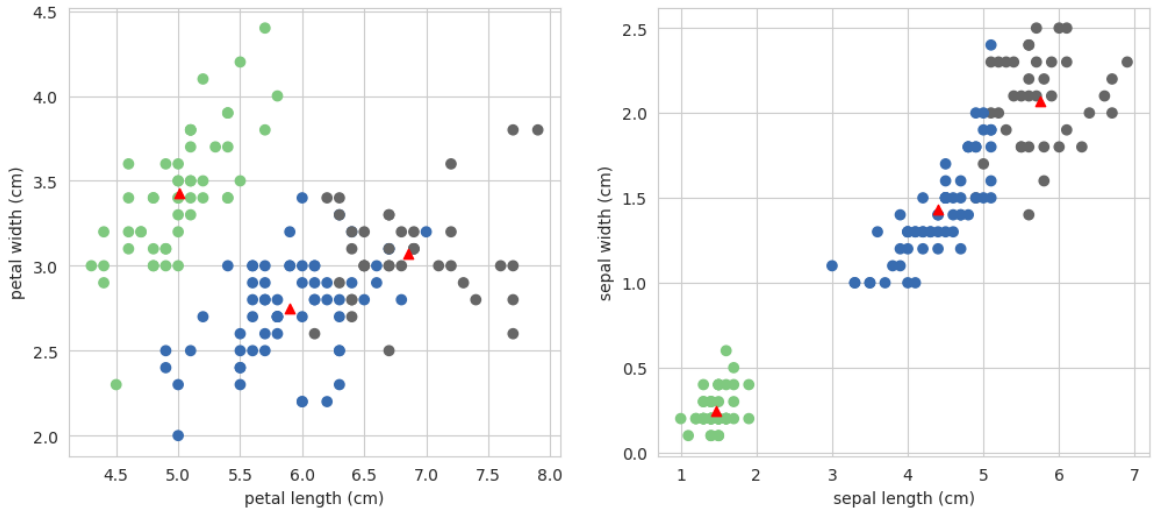
```
array([[5.006     , 3.428     , 1.462     , 0.246     ],
       [5.9016129 , 2.7483871 , 4.39354839, 1.43387097],
       [6.85      , 3.07368421, 5.74210526, 2.07105263]])
```

Gán các giá trị với trọng tâm các cụm và biểu diễn bằng đồ thị:

```
plt.figure(figsize=(12,5))
plt.subplot(1,2,1)
plt.scatter(X[:,0],X[:,1],c = pred, cmap=cm.Accent)
plt.grid(True)
for center in kmeans.cluster_centers_:
    center = center[:2]
    plt.scatter(center[0],center[1],marker = '^',c = 'red')
plt.xlabel("petal length (cm)")
plt.ylabel("petal width (cm)")

plt.subplot(1,2,2)
plt.scatter(X[:,2],X[:,3],c = pred, cmap=cm.Accent)
plt.grid(True)
for center in kmeans.cluster_centers_:
    center = center[2:4]
    plt.scatter(center[0],center[1],marker = '^',c = 'red')
plt.xlabel("sepal length (cm)")
plt.ylabel("sepal width (cm)")
plt.show()
```

Kết quả trả về:



Biểu đồ con ở bên trái hiển thị chiều dài cánh hoa so với chiều rộng cánh hoa với các điểm dữ liệu được màu sắc theo cụm, và các dấu đánh dấu màu đỏ chỉ trung tâm cụm K-means. Biểu đồ con ở bên phải hiển thị chiều dài lá đài so với chiều rộng lá đài một cách tương tự.

2. Ứng dụng của thuật toán phân cụm, phân hạch K-means

2.1. Ưu điểm và khuyết điểm của thuật toán K-means:

Giống với các thuật toán khác, thuật toán phân cụm K-means cũng có một số ưu điểm vượt trội và có những hạn chế còn tồn tại:

Thuật toán phân cụm K-means có một số ưu điểm như sau:

- Đơn giản và dễ triển khai: Thuật toán K-means dễ hiểu và triển khai, làm cho nó trở thành lựa chọn phổ biến cho các tác vụ phân cụm.
- Nhanh và hiệu quả: K-means hiệu quả về mặt tính toán và có thể xử lý các tập dữ liệu lớn với số chiều cao
- Khả năng mở rộng: K-means có thể xử lý các tập dữ liệu lớn với một số lượng lớn điểm dữ liệu và có thể được mở rộng dễ dàng để xử lý các tập dữ liệu lớn hơn.
- Linh hoạt: K-means có thể được điều chỉnh dễ dàng cho các ứng dụng khác nhau và có thể được sử dụng với các phép đo khoảng cách và phương pháp khởi tạo khác nhau.

Bên cạnh đó, thuật toán vẫn còn một số mặt hạn chế, như:

- Yêu cầu xác định số lượng cụm: Số lượng cụm k cần được xác định trước khi chạy thuật toán, điều này có thể gây khó khăn trong một số ứng dụng.
- Nhạy cảm với các điểm ban đầu: K-means nhạy cảm với việc lựa chọn ban đầu của các điểm và có thể hội tụ vào một giải pháp không tối ưu.
- Nhạy cảm với ngoại lệ: K-means nhạy cảm với ngoại lệ, có thể có tác động đáng kể đến các cụm kết quả.

2.2. Ứng dụng của thuật toán phân cụm K-means

Thuật toán phân cụm, phân hạch K-means có nhiều ứng dụng cụ thể trong thực tế, bao gồm các ứng dụng sau:

- Phân khúc thị trường: K-means được sử dụng để phân loại khách hàng thành các nhóm dựa trên các đặc điểm như hành vi mua hàng, thu nhập, tuổi, giới tính,... để các doanh nghiệp đưa ra các chiến lược và ưu đãi nhằm tăng thu hút của khách hàng đối với sản phẩm của họ qua đó làm tăng lợi nhuận mà doanh nghiệp có thể đạt được.
- Phân tích gen trong y khoa: K-means có thể được sử dụng để phân loại các gen dựa trên các đặc điểm của chúng, từ đó giúp cho các bác sĩ và các nhà nghiên cứu hiểu rõ hơn về cấu trúc gen và ảnh hưởng của chúng đến với sức khỏe con người.
- Phân đoạn hình ảnh và nén hình ảnh: K-means có thể được sử dụng để phân loại các pixel trong một hình ảnh dựa trên màu sắc, độ sáng,...
- Phát hiện tế bào ung thư: K-means có thể được sử dụng để phân loại các tế bào dựa trên các đặc điểm của chúng, giúp phát hiện ra các tế bào bất thường có thể là tế bào ung thư, qua đó tạo điều kiện cho người mắc ung thư có thể điều trị sớm trong thời gian vàng, có thể chữa khỏi hoàn toàn bệnh.
- Phát hiện bất thường (anomaly detection): K-means có thể được sử dụng để phát hiện các điểm dữ liệu bất thường bằng cách xem xét khoảng cách giữa chúng và trung tâm của cụm từ đó có thể phát hiện sớm các biểu hiện bất thường như gian lận hay không trung thực.

Tuy nhiên, ta phải lưu ý rằng thuật toán K-means có một số hạn chế kể trên, và đây là thuật toán học máy không giám sát, nên phải cẩn trọng trong quá trình sử dụng.

3. So sánh với các công nghệ khác

3.1. Thuật toán DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise - phân cụm không gian dựa trên mật độ các ứng dụng với nhiễu) là một thuật toán phân cụm dựa theo mật độ. Điểm then chốt của thuật toán này là vùng lân cận của mỗi điểm trong cụm nằm trong một bán kính R cho trước phải có một số lượng tối thiểu các điểm M. Thuật toán này đã được chứng minh cực kỳ hiệu quả trong phát hiện ngoại lệ và xử lý nhiễu.

Thuật toán được hoạt động như sau:

- Loại của mỗi điểm trong tập dữ liệu được xác định. Mỗi điểm dữ liệu có thể là một trong những loại sau:

+ Điểm Lõi: Một điểm là một điểm lõi nếu có ít nhất M điểm trong vùng lân cận của nó, tức là trong vòng bán kính R đã được chỉ định.

+ Điểm Biên: Một điểm được quy định là điểm biên nếu: Vùng lân cận của nó chứa ít hơn M điểm dữ liệu, hoặc nó có thể được tiếp cận từ một số điểm lõi (nằm trong khoảng cách từ R từ một điểm lõi)

+ Điểm Ngoại lệ: Một điểm ngoại lệ là một điểm không phải điểm lõi, và cũng không đủ gần để có thể tiếp cận từ một điểm lõi.

- Điểm ngoại lệ sẽ bị loại bỏ.

- Điểm lõi là hàng xóm được kết nối và được đặt vào cùng trong một cụm.

- Các điểm biên được gán cho mỗi cụm.

Sự khác biệt giữa hai thuật toán phân cụm K-means và phân cụm DBSCAN:

	Phân cụm K-means	Phân cụm DBSCAN
Hình dạng cụm được hình thành	Ít nhiều có hình cầu lồi, có cùng kích thước đặc điểm	Hình dạng tùy ý, có thể không cùng kích thước, đặc điểm
Số lượng cụm	Cần thiết và rất nhạy cảm với số cụm được chỉ định	Không cần phải chỉ định số lượng cụm
Hiệu quả với	Các tập dữ liệu lớn	Không thể xử lý hiệu quả các bộ dữ liệu nhiều chiều
Dữ liệu ngoại lệ và xử lý nhiễu	Không hoạt động tốt	Xử lý hiệu quả
Phát hiện bất thường	Có vấn đề vì các điểm bất thường sẽ được gán vào cùng một cụm với điểm dữ liệu “bình thường”	Xác định bất thường bằng phân tách các vùng có mật độ cao với các vùng có mật độ thấp
Tham số đầu vào	Số cụm K	Bán kính R và điểm tối thiểu M
Mật độ dữ liệu	Không bị ảnh hưởng bởi mật độ khác nhau của các điểm dữ liệu	Không hoạt động tốt đối với các tập dữ liệu thưa thớt hoặc các điểm dữ liệu có mật độ khác nhau

3.2. Thuật toán phân cụm phân cấp

Thuật toán phân cụm phân cấp (Hierarchical Clustering) là mô hình phân cụm dựa trên kết nối, nhóm các điểm dữ liệu gần nhau dựa trên thước đo độ tương tự hoặc khoảng cách. Giả định rằng các điểm dữ liệu ở gần nhau thì giống nhau hoặc có liên quan hơn so với các điểm dữ liệu ở xa nhau hơn. Thuật toán này là một phương pháp phân tích cụm mà không có số lượng cụm cố định.

Sự khác biệt giữa hai thuật toán phân cụm K-means và phân cụm phân cấp:

	Phân cụm K-means	Phân cụm phân cấp
Số lượng cụm	Sử dụng số lượng cụm được chỉ định trước. Cần có hiểu biết nhất định về số lượng cụm cần chia.	Chia rẽ hoặc kết tụ. Có thể dừng ở bất kỳ số lượng cụm nào phù hợp
Số lượng cụm	Cần thiết và rất nhạy cảm với số cụm được chỉ định	Không cần phải chỉ định số lượng cụm
Trọng tâm cụm	Sử dụng trung bình hoặc trung vị	Các phương pháp tổng hợp bắt đầu bằng n cụm và kết hợp tuần tự các cụm tương tự cho đến khi chỉ thu được một cụm
Tập dữ liệu	Phù hợp tập dữ liệu lớn	Đặc biệt hữu ích khi mục tiêu là sắp xếp các cụm thành một hệ thống phân cấp tự nhiên
Kết quả	Kết quả tạo ra bằng cách chạy thuật toán nhiều lần có thể khác nhau	Kết quả luôn lặp lại
Cấu trúc cụm là hình cầu	Hoạt động tốt	Không hoạt động tốt

CHƯƠNG 3: XÂY DỰNG CHƯƠNG TRÌNH PHẦN MỀM

1. Cài đặt thử nghiệm chương trình phần mềm

1.1. Yêu cầu bài toán

Một tổ chức phi chính phủ nhân đạo quốc tế cam kết chống lại nghèo đói và lạc sẽ cung cấp cho người dân các quốc gia lạc hậu cứu trợ. Tổ chức này đã gây quỹ được một số tiền lớn, và số tiền này cần được phân bổ một cách chiến lược hiệu quả hợp lý. Vì thế, để chọn ra các quốc gia cần sự giúp đỡ khẩn cấp nhất, cần quyết định dựa trên dữ liệu cần thực hiện. Việc phân loại các quốc gia sử dụng các yếu tố kinh tế-xã hội và sức khỏe quyết định sự phát triển tổng thể của quốc gia trở nên cần thiết. Do đó, dựa trên các cụm của các quốc gia tùy thuộc vào điều kiện của họ, quỹ sẽ được phân bổ để hỗ trợ trong thời gian của thảm họa và thiên tai.

Mục tiêu cần đạt được: Phân cụm các quốc gia dựa theo các đặc trưng số học của các đặc tính.

1.2. Bộ dữ liệu

Bài toán này sử dụng bộ dữ liệu Country-data.csv thể hiện các đặc tính của các quốc gia trên thế giới. Bộ dữ liệu gồm 167 dòng và 10 cột.

Bộ dữ liệu có các thuộc tính sau đây:

- country: Tên các quốc gia

- child_mort: Tỉ lệ tử vong của trẻ dưới 5 tuổi trên 1000 trẻ được sinh ra

- exports: Xuất khẩu hàng hóa và dịch vụ bình quân đầu người. Được cho là % của GDP bình quân đầu người.

- health: Tổng chi tiêu cho sức khỏe bình quân đầu người. Được cho là % của GDP bình quân đầu người

- imports: Nhập khẩu hàng hóa và dịch vụ bình quân đầu người. Được cho là % của GDP bình quân đầu người

- income: Thu nhập ròng bình quân đầu người

- inflation: Sự đo lường tỷ lệ tăng trưởng hàng năm của tổng GDP

- life_expec: Số năm trung bình mà một đứa trẻ mới sinh sẽ sống nếu các tỉ lệ tử vong hiện tại được duy trì

- total_fer: Số lượng trẻ em sẽ được sinh ra cho mỗi phụ nữ nếu các tỷ lệ sinh sản theo tuổi hiện tại được duy trì

- gdpp: GDP bình quân đầu người. Được tính toán như tổng GDP chia cho tổng dân số.

1.3. Chạy chương trình phần mềm

Chương trình phần mềm được đính kèm cùng bản báo cáo và có giải thích đi kèm.

2. Hướng dẫn cài đặt phần mềm

Chương trình được viết bằng ngôn ngữ Python chạy trên IDE PyCharm, vì thế đầu tiên cần tải phiên bản Python thích hợp (ví dụ python 3.11) và tải PyCharm trên trang chủ chính thức của JetBrains.

Để chương trình chạy được, cần tải các thư viện sau đây:

- pandas: sử dụng để xử lý và phân tích dữ liệu, giúp làm việc với dữ liệu dễ dàng và hiệu quả.

- numpy: hỗ trợ tính toán số học, đặc biệt là trên ma trận, cung cấp nhiều hàm toán học để thực hiện các phép toán trên mảng.

- matplotlib: thư viện vẽ đồ thị 2D và 3D trong Python, cung cấp cách tạo ra biểu đồ, từ biểu đồ đơn giản đến biểu đồ phức tạp.

- seaborn: thư viện đồ thị dựa trên matplotlib cung cấp giao diện vẽ đồ thị thống kê.

- sklearn (scikit-learn): thư viện dành cho học máy, bao gồm các thuật toán phân loại, hồi quy, phân cụm và giảm chiều dữ liệu.

- mpl_toolkits: một phần của matplotlib, cung cấp một số công cụ để tạo ra biểu đồ 3D và biểu đồ trên bản đồ.

- plotly: thư viện cho phép tạo ra các biểu đồ tương tác đẹp mắt

- kaleido: thư viện được sử dụng để xuất các biểu đồ plotly thành các tệp tĩnh như PNG, JPEG, SVG hoặc PDF.

3. Hướng dẫn chi tiết sử dụng các chức năng của phần mềm

Trước khi khởi chạy phần mềm, cần kiểm tra kỹ vị trí các tệp, đặc biệt là tệp dữ liệu cần thiết, ví dụ như cần thể hiện đúng vị trí tệp dữ liệu đầu vào Country_data.csv của chương trình:

```
data = pd.read_csv('/Country-data.csv')
```

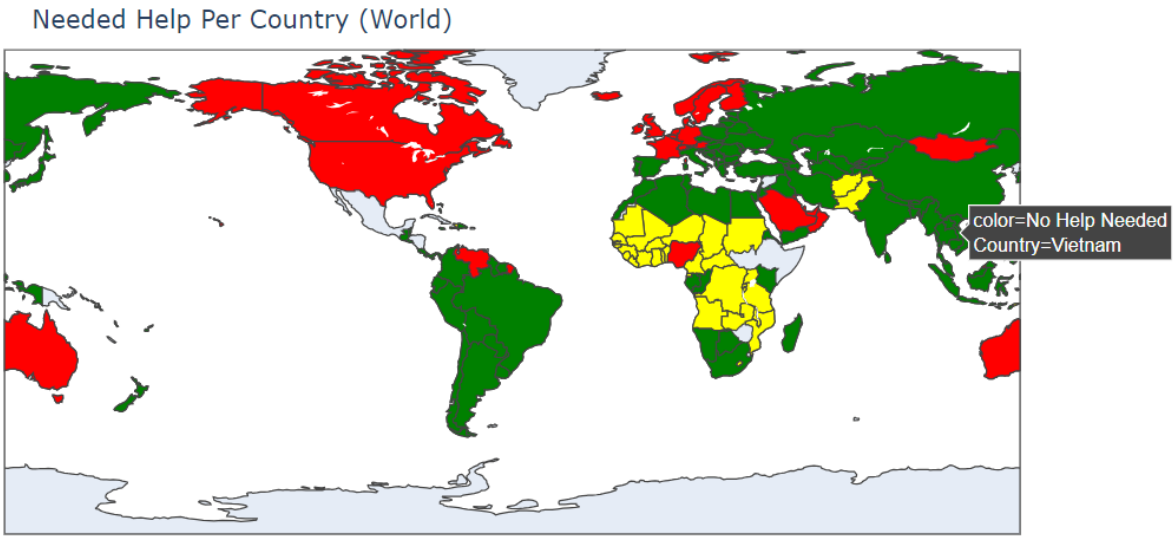
Sau đó, chúng ta sẽ kiểm tra các đặc tính của dữ liệu, thực hiện quá trình giảm chiều dữ liệu (được chú thích trong mã nguồn).

Đầu ra của chương trình sẽ là bản đồ các quốc gia cần sự giúp đỡ (cần viện trợ) được thể hiện bằng màu sắc:

- Màu vàng: Có thể cần giúp đỡ

- Màu xanh: Không cần giúp đỡ

- Màu đỏ: Cần sự giúp đỡ



Bản đồ có thể tương tác được, có thể di chuột vào quốc gia để xem trạng thái cần giúp đỡ hay không và tên quốc gia đó.

CHƯƠNG 4: ĐÁNH GIÁ KẾT QUẢ NGHIÊN CỨU VÀ KẾT LUẬN

Bài thu hoạch bắt đầu từ quá trình tìm hiểu cơ sở lý luận của thuật toán K-means, xác định khái niệm, đặc điểm cũng như cách thức hoạt động của nó. Điều này giúp nắm bắt được bản chất của lĩnh vực này và nhận thức rằng K-means không chỉ đơn giản là một công nghệ, mà còn là một công cụ mạnh mẽ để phân tích dữ liệu.

Bên cạnh đó, bài thu hoạch đã để cập đến ứng dụng đa dạng của thuật toán phân cụm K-means, bao gồm các ứng dụng của công nghệ này trong đa dạng lĩnh vực khác nhau như phân tích dữ liệu, y sinh học, xử lý ảnh, v.v. thể hiện tầm quan trọng và tính đa dạng của K-means trong thế giới hiện đại.

Bài thu hoạch cũng đã so sánh K-means với một số công nghệ nổi bật trong các thuật toán phân hạch, phân cụm khác và làm nổi bật các điểm giống nhau, khác nhau giữa các kỹ thuật. Mỗi công nghệ đều có đặc điểm ưu việt và hạn chế riêng của mình, sự linh hoạt trong việc áp dụng các công nghệ này cho từng tình huống cụ thể giúp chúng ta làm chủ các công nghệ hiện đại trong một thế giới ngày càng tiến bộ.

Cuối cùng, bài thu hoạch đã thực hiện cài đặt và chạy thử thuật toán phân cụm K-means, thể hiện được ứng dụng mạnh mẽ của thuật toán trong thực tế cuộc sống.

Kết luận, bản báo cáo này đã đưa ra cái nhìn tổng quan và cụ thể về thuật toán phân cụm K-means, và hy vọng rằng giúp mọi người hiểu rõ hơn về các thuật toán phân hạch, phân cụm cũng và ứng dụng của chúng trong thực tế.

TÀI LIỆU THAM KHẢO

[1] GeeksforGeeks, "K-means Clustering Introduction," [Online]. Available: https://www.geeksforgeeks.org/k-means-clustering-introduction/.

[2] VinBigdata, "Supervised Learning và Unsupervised Learning: Khác biệt là gì?," [Online]. Available: https://blog.vinbigdata.org/supervised-learning-va-unsupervised-learning-khac-biet-la-gi/.

[3] "K-means Clustering Algorithm: Applications, Types, & How Does It Work?," [Online]. Available: https://www.simplilearn.com/tutorials/machine-learning-tutorial/k-means-clustering-algorithm#advantages_of_kmeans.

[4] GeeksforGeeks. [Online]. Available: https://www.geeksforgeeks.org/difference-between-k-means-and-dbscan-clustering/.

[5] GeeksforGeeks, "Difference between K means and Hierarchical Clustering," [Online]. Available: https://www.geeksforgeeks.org/difference-between-k-means-and-hierarchical-clustering/.

[6] TechTarget, "AI Artificial Intelligence," [Online]. Available: https://www.techtarget.com/searchenterpriseai/definition/AI-Artificial-Intelligence.

[7] T. Deshpande, "Clustering: PCA| K-Means - DBSCAN - Hierarchical |," 21 10 2022. [Online]. Available: https://www.kaggle.com/code/tanmay111999/clustering-pca-k-means-dbscan-hierarchical/notebook.