|  |  |  |
| --- | --- | --- |
|  | **TRƯỜNG ĐẠI HỌC BÁCH KHOA**  **KHOA CÔNG NGHỆ THÔNG TIN** |  |

TIỂU LUẬN

KHOA HỌC DỮ LIỆU

**PHÂN CỤM POKEMON**

|  |  |  |
| --- | --- | --- |
| HỌ VÀ TÊN SINH VIÊN | LỚP HỌC PHẦN | ĐIỂM BẢO VỆ |
| Huỳnh Ngọc Thuật | 19N10 |  |
| Đoàn Tân Tỵ | 19N10 |  |
| Hường Lê Nguyên | 19N10 |  |

ĐÀ NẴNG, 06/2022

**TÓM TẮT**

Hiện nay nhu cầu khám phá kho dữ liệu đó phục vụ cho nhu cầu con người ngày càng cao, điều đó đòi hỏi con người phải biết khai thác dữ liệu và xử lý thông tin đó thành tri thức có ích. Một trong những kỹ thuật quan trọng trong quá trình khai phá dữ liệu và xử lý dữ liệu lớn là kỹ thuật phân cụm dữ liệu.

Pokemon là một series phim hoạt hình khá nổi tiếng có xuất xứ từ Nhật Bản các ấn phẩm cũng như các game và sản phẩm liên qua đã đem lại một nguồn thu rất lớn. Những con pokemon trong series này sẽ chứa những đặc điểm, chỉ số khác nhau và nhiệm vụ của nhóm ở bài toán này đó là sẽ đi phân nhóm những pokemon này dựa trên chỉ số đã có.

Ở đây em chọn 3 thuật toán là Kmeans Clustering, Hierarchical Clustering và DBSCAN để tiến hành phân cụm và chọn ra thuật toán tối ưu nhất dựa trên độ tách biệt giữa các cụm.

Kết quả ….

**BẢNG PHÂN CÔNG NHIỆM VỤ**

|  |  |  |
| --- | --- | --- |
| Sinh viên thực hiện | Các nhiệm vụ | Tự đánh giá theo 3 mức  (Đã hoàn thành/Chưa hoàn thành/Không triển khai) |
| Huỳnh Ngọc Thuật | - Thu thập dữ liệu   * Mô tả dữ liệu | Đã hoàn thành |
| Đoàn Tân Tỵ | - Xử lý dữ liệu  - Cụm K và thuật toán K-means | Đã hoàn thành |
| Hường Lê Nguyên | - Thuật toán Hierarchical và DBSCAN.  - Kết quả | Đã hoàn thành |

**MỤC LỤC**

# 1. Giới thiệu

Bài toán mà nhóm cần giải quyết có 2 vấn đề chính: chọn ra số cụm tốt nhất và đánh giá được mô hình phù hợp nhất với dữ liệu.

* Chọn ra số cụm tốt nhất dựa trên 3 yếu tố: Elbow Method, bằng đồ thị Hierarchical và số giá trị ngoại lệ của thuật toán DBSCAN.
* Đánh giá mô hình nào tốt nhất dựa vào: Silhouette Analysis (Mức độ tách biệc của các cụm).

# 2. Thu thập và mô tả dữ liệu

**2.1. Thu thập dữ liệu**

Dữ liệu được nhóm tìm hiểu và thu thập từ trang <https://pokemondb.net/pokedex/all>, trang này chưa đầy đủ thông tin cơ bản về chỉ số của từng con pokemon.

Dữ liệu được thu thập bằng Beautiful Soup một thư viện của Python giúp crawl dữ liệu một cách dễ dàng từ các trang web theo mục đích riêng của mình.

**2.2. Mô tả dữ liệu**

Dataset bao gồm 1076 mẫu, mỗi mẫu chứa 11 đặc trưng cần thiết của một pokemon

SV mô tả các thống kê tổng quan về tập dữ liệu (số mẫu, số đặc trưng của 1 mẫu, kiểu dữ liệu của mỗi đặc trưng, số mẫu dữ liệu trống của mỗi đặc trưng,...) và xuất ra các thống kê mô tả trực quan về các đặc trưng (ví dụ: dùng boxplot, histogram, scatter plot về độ tương quan,…).

# 3. Trích xuất đặc trưng

SV trình bày việc lựa chọn đặc trưng, làm sạch và chuẩn hóa dữ liệu, giảm chiều dữ liệu, …; và trực quan hoá kết quả/tác dụng của các quá trình trên.

# 4. Mô hình hóa dữ liệu

SV trình bày quá trình xây dựng, xác thực và kiểm thử mô hình gồm:

- Chọn các mô hình phù hợp (ít nhất 2 mô hình) và mô tả ngắn gọn cơ sở lý thuyết của các mô hình đã lựa chọn, mô tả bộ tham số của mỗi mô hình (ví dụ: số lớp của mạng nơ-ron, số node của mỗi lớp, hàm activation được sử dụng,…) thông qua API sử dụng để xây dựng mô hình.

- Chia dữ liệu thành các tập Huấn luyện/Xác thực/Kiểm thử theo tỉ lệ phù hợp.

- Trình bày các tham số của quá trình huấn luyện mô hình (ví dụ: loss function, learning rate, number of epochs,…).

- Trình bày các đồ thị thể hiện hiệu suất (đánh giá bằng Loss hoặc Accuracy) của các mô hình trên các tập Huấn luyện/Xác thực/Kiểm thử.

- Đưa ra các độ đo (metrics) dùng để đánh giá các mô hình (ví dụ: RMSE hoặc MAPE), và so sánh hiệu quả của các mô hình bằng bảng số liệu (hoặc đồ thị) dựa trên các độ đo.

# 5. Kết luận

Sinh viên tổng kết các việc đã làm và các kết quả đạt được. Nêu hướng phát triển để cải thiện kết quả hiện có.

# 6. Tài liệu tham khảo

SV liệt kê các TLTK đã trích dẫn (cite) trong báo cáo tại đây.