Klasteryzacja danych o płatkach śniadaniowych

Autor: Tsimafei Kotski 158740 Grupa: L2

Celem zadania było zaimplementowanie algorytmu k-means oraz zastosowanie go do klasteryzacji danych o płatkach śniadaniowych.

Opis danych

Dane zawierają informacje o 77 płatkach śniadaniowych opisanych przez 16 atrybutów: `name`, `mfr`, `type`, `calories`, `protein`, `fat`, `sodium`, `fiber`, `carbo`, `sugars`, `potass`, `vitamins`, `shelf`, `weight`, `cups`, `rating`.

Metody przetwarzania danych

1. Usunięcie atrybutów nominalnych

Usunięto kolumny `name` i `mfr`, są to atrybuty nominalne, które nie nadają się do bezpośredniego użycia w algorytmie k-means opartym na odległościach euklidesowych.

2. Konwersja atrybutu type

Atrybut `type` został przekształcony z wartości tekstowych na numeryczne:

- C (zimny) → 0
- H (gorący) → 1

Po tych operacjach zbiór danych zawierał 77 przykładów opisanych przez 14 atrybutów numerycznych.

3. Standaryzacja danych

Zastosowano standaryzację danych przy użyciu `StandardScaler`, aby każdy atrybut miał średnią równą 0 i odchylenie standardowe równe 1. Standaryzacja jest niezbędna ze względu na różne jednostki i zakresy wartości atrybutów.

4. Selekcja atrybutów

Zastosowano metodę `VarianceThreshold` z progiem 0.1. Wszystkie 14 atrybutów zostało zachowanych (żadne nie zostały usunięte), co wskazuje, że wszystkie atrybuty są wystarczająco zmienne i wykazują znaczące wahania wartości między różnymi płatkami.

Implementacja algorytmu k-means

Inicjalizacja centroidów

Centroidy są inicjalizowane losowo poprzez wybór k losowych punktów z danych.

Przypisanie punktów do klastrów

Każdy punkt jest przypisywany do najbliższego centroidu na podstawie odległości euklidesowej.

Aktualizacja centroidów

Po przypisaniu wszystkich punktów, centroidy są aktualizowane jako średnie arytmetyczne punktów w każdym klastrze.

Warunki stopu

Algorytm zatrzymuje się, gdy spełniony jest jeden z warunków:

1. Osiągnięcie maksymalnej liczby iteracji (100)
2. Zbieżność - zmiana wszystkich współrzędnych wszystkich centroidów jest mniejsza niż epsilon = 1 * e^-4

Wyniki klasteryzacji

Parametry algorytmu

- Liczba klastrów: k = 3
- Maksymalna liczba iteracji: 100
- Próg zbieżności: epsilon = 1 * e^-4
- Ziarno losowości: 42

Liczba iteracji i zbieżność

Algorytm osiągnął zbieżność po 7 iteracjach

Rozkład płatków w klastrach

Klaster	Liczba płatków	Procent
Klaster 1	22	28.6%
Klaster 2	34	44.2%
Klaster 3	21	27.3%
Suma	77	100%

Rozkład jest stosunkowo równomierny, z lekką przewagą klastra 2.

Analiza klastrów

Centroidy klastrów

Analiza wartości centroidów (znormalizowanych) pozwala na scharakteryzowanie klastrów:

Charakterystyka klastrów

Klaster 1 - "Wysokowartościowe płatki z błonnikiem" (22 płatki): Wysoka zawartość białka, tłuszczu, błonnika i potasu. Wyższa kaloryczność, większa waga porcji.

Klaster 2 - "Standardowe płatki śniadaniowe" (34 płatki): Niska zawartość białka i błonnika, wyższa zawartość sodu i węglowodanów. Umiarkowana kaloryczność, niższa ocena konsumentów. Największy klaster.

Klaster 3 - "Zdrowa, niskokaloryczna opcja" (21 płatków): Bardzo niska zawartość cukru i kalorii, niska zawartość tłuszczu i sodu. Wyższa zawartość białka i błonnika. Więcej płatków gorących.