Sztuczna inteligencja

WYKŁAD 2

Klasteryzacja – podstawowe założenia

- Podstawowym zadaniem klasteryzacji (grupowania) jest dokonanie podziału zbioru przypadków C znajdujących się w bazie na grupy C_1, C_2, ..., C_k, nazywane klastrami, stanowiące podzbiory przypadków podobnych do siebie, przy czym pojęcie podobieństwa może być definiowane w różny sposób.
- Podział zbioru C powinien być dokonany w taki sposób, aby przypadki z danej grupy były bardziej podobne do siebie (*homogeniczność*) niż do jakichkolwiek przypadków z pozostałych grup (*heterogeniczność*).

Klasteryzacja – podstawowe założenia

- Istotnym zagadnieniem jest ustalenie liczby *k* grup, na które zbiór przypadków ma zostać podzielony, gdyż zazwyczaj liczba ta nie jest z góry zadana.
- Kryteria klasteryzacji dotyczą interpretacji semantycznej klastrów.
- Istotna jest odpowiedź na pytanie dlaczego dwa przypadki przypisywane są do tego samego klastra. W tej kwestii odpowiedź może być udzielona na podstawie dostępnej wiedzy.

Klasteryzacja – podstawowe założenia

- W wielu sytuacjach przypadki grupowane są razem ze względu na istniejące pomiędzy nimi zależności takie jak np. noeodróżnialność, podobieństwo, bliskość, funkcjonalność, zgodność.
- Istotnym zagadnieniem w procesie klasteryzacji zbioru przypadków jest ustalenie struktury klastrów.
- Klastry mogą być parami rozłączne lub też zachodzące na siebie.

Klasteryzacja – podstawowe założenia

- W przypadku klastrów rozłącznych mówi się o tzw. podziale ostrym. W takiej sytuacji dany przypadek należy tylko do jednego klastra.
- W przypadku klastrów zachodzących na siebie mówi się o tzw. podziale rozmytym. Przy tym podziale dany przypadek może należeć do wielu klastrów.
- Dodatkowo określany jest stopień przynależności przypadku do danego klastra.
- Stopień ten ma wartość rozmytą z przedziału [0, 1]. Wynika z tego, że przypadek może należeć do grupy tylko w pewnym stopniu.

Klasteryzacja – podstawowe założenia

- Dla podziałów rozmytych możliwe są dwie sytuacje:
  - W pierwszej z nich suma stopni przynależności danego przypadku do każdego z klastrów jest zawsze równa 1 (tzw. podział probabilistyczny).
  - W drugiej sytuacji warunek sumowania się stopni przynależności do 1 nie obowiązuje (tzw. podział posybilistyczny).

Miary odległości

- Dane są dwa punkty *x* oraz *y* w przestrzeni *m*-wymiarowej:

$$x = [x_{1,} x_{2,} \ldots, x_m]$$

$$y = [y_{1,} y_m, \ldots, y_m]$$

Miary odległości

- Odległość (metryka) Euklidesa

$$d_{Eukl}(x, y) = \sqrt{\sum_{i=1}^{n} (y_i - x_i)^2}$$

- Odległość (metryka) Manhattanu

$$d_{Manh}(x, y) = \sum_{i=1}^{n} |y_i - x_i|$$

Miary odległości

- Odległość (metryka) Minkowskiego

$$d_{Mink}(x, y) = \sqrt[q]{\sum_{i=1}^{n} |y_i - x_i|^q}$$

Miary odległości

- Dla atrybutów symbolicznych możemy zdefiniować funkcję *R* („różne od”).
- Dla *i*-tego atrybutu funkcja ma postać:

$$R(x_i, y_i) = \begin{cases} 0 & dla\ x_i = y_i \\ 1 & w\ przeciwnym\ przypadku \end{cases}$$

- Funkcja *R* może zostać zastosowana dla *i*-tego atrybutu w miarach odległości.

Normalizacja wartości atrybutów

- Przy wyznaczaniu odległości atrybuty posiadające duże wartości mogą niwelować wpływ innych atrybutów (tych posiadających mniejsze wartości).
- W celu wyeliminowania tej sytuacji należy dokonać normalizacji wartości atrybutów.

Normalizacja wartości atrybutów

- Normalizacja min-max:
  - dla zbioru wartości atrybutu:

$$X = \{x_1, x_2, \ldots, x_k\}$$

  - znormalizowana wartość x_i jest obliczana jako:

$$x_i^* = \frac{x_i - min(X)}{zakres(X)} = \frac{x_i - min(X)}{max(X) - min(X)}$$

Algorytmy iteracyjnej optymalizacji

- W algorytmach iteracyjnej optymalizacji najlepszy podział zbioru przypadków jest wyznaczany przez iteracyjne polepszanie pewnych wskaźników jakości, startując z początkowego, najczęściej losowego, podziału.

Algorytm HCM (Hard C-Means)

- Algorytm dzieli jednoznacznie zbiór przykładów na *c* grup.
- Środek grupy jest średnią położenia wszystkich przykładów należących do tej grupy.

Algorytm HCM (Hard C-Means)

- KROK 1: Ustalenie liczby (*c*) grup, na które zbiór przypadków będzie dzielony.
- KROK 2: Losowe przypisanie *c* rekordów jako początkowych środków grup.
- KROK 3: Znalezienie dla każdego rekordu najbliższego środka grupy.
- KROK 4: Dla każdej z *c* grup znalezienie centroidu grupy i uaktualnienie położenie każdego środka grupy jako nową wartość centroidu.
- KROK 5: Powtórzenie kroków od 3 do 5 aż do osiągnięcia warunku zakończenia algorytmu.

Algorytm HCM (Hard C-Means)

- Obliczenie centroidu grupy
  - dla grupy przypadków:

$$x^1 = [x_1^1, x_2^1, \ldots, x_m^1]$$
$$x^2 = [x_1^2, x_2^2, \ldots, x_m^2]$$
$$\ldots$$
$$x^k = [x_1^k, x_2^k, \ldots, x_m^k]$$

  - centroid ma postać:

$$\left(\frac{\sum_{i=1}^{k} x_1^i}{k}, \frac{\sum_{i=1}^{k} x_2^i}{k}, \ldots, \frac{\sum_{i=1}^{k} x_m^i}{k}, \right)$$

Algorytm HCM (Hard C-Means)

- Warunkiem zakończenia algorytmu może być sytuacja gdy dla wszystkich grup wszystkie przypadki przypisane do środka tej grupy pozostają w tej grupie.
- Inaczej, algorytm kończy działanie gdy osiągnięte zostanie kryterium zbieżności – minimalizacja sumarycznego błędu kwadratowego *(p* jest przykładem z *i*-tej grupy, m_i jest centroidem *i*-tej grupy):

$$SSE = \sum_{i=1}^{c} \sum_{p \in C_i} d(p, m_i)^2$$

Algorytm FCM (Fuzzy C-Means)

- Algorytm dzieli zbiór przykładów na *c* grup.
- Przykłady mogą należeć do różnych grup z odpowiednimi stopniami przynależności.
- Suma przynależności danego przykładu do każdej z grup jest zawsze równa 1.

Algorytm PCM (Possibilistic C-Means)

- Algorytm dzieli zbiór przykładów na *c* grup.
- Przykłady mogą należeć do różnych grup z odpowiednimi stopniami przynależności.
- Suma przynależności danego przykładu do każdej z grup nie musi być równa 1.

Klasteryzacja hierarchiczna

- W klasteryzacji hierarchicznej danych tworzona jest struktura drzewiasta (dendrogram) poprzez rekurencyjne dzielenie lub łączenie istniejących grup.
- Metody aglomeracyjne:
  - na początku zakłada się, że każdy przykład stanowi oddzielną grupę,
  - w kolejnych krokach dwie grupy, które są najbliżej siebie, łączy się w nową wspólną grupę,
  - ostatecznie wszystkie przykłady należą do jednej grupy.

Klasteryzacja hierarchiczna

- Metody rozdzielające:
  - na początku zakłada się, że wszystkie przykłady należą do jednej grupy,
  - w kolejnych krokach najbardziej niepodobne przykłady rozdzielane są w osobne grupy,
  - ostatecznie każdy przykład stanowi oddzielną grupę.

Klasteryzacja hierarchiczna

- Kryteria określania odległości pomiędzy grupami:
  - *metoda pojedynczego połączenia* (metoda najbliższego sąsiedztwa) – określana jest minimalna odległość pomiędzy dwoma przykładami należącymi do różnych grup,
  - *metoda całkowitego połączenia* (metoda najdalszego sąsiedztwa) – określana jest maksymalna odległość pomiędzy dwoma przykładami należącymi do różnych grup,
  - *metoda średniego połączenia* – określana jest średnia odległość pomiędzy wszystkimi przykładami z jednej grupy i wszystkimi przykładami z drugiej grupy.