

Data Mining

Lab 5: Gruparea datelor

Sumar:

- Scopul grupării datelor
- Algoritmi partiționali
- Algoritmi ierarhici
- Algoritmi bazați pe densitate

1. Scopul grupării datelor

Gruparea (clusterizarea) datelor are ca scop identificarea de subseturi de date similare (într-un anumit sens), numite grupuri, clustere sau clase. Elementul specific unui proces de grupare este faptul că nu se cunosc apriori grupurile (uneori nici măcar numărul acestora) scopul algoritmilor de grupare fiind de a identifica aceste grupuri folosind doar relațiile dintre date.

2. Algoritmi partiționali

Produc o partiționare a setului de date în clustere, fiecare cluster având asociat un reprezentant (centroid). Cel mai simplu algoritm partițional este kMeans care se caracterizează prin faptul că pentru un set de date și un număr de clustere (K, specificat de către utilizator) produce un set de K centroizi care determină o alocare univocă a datelor la clustere (fiecare dată este alocată clusterului corespunzător celui mai apropiat centroid).

Structura generală a algoritmului kMeans este:

Etapa 1. Inițializare: se aleg aleator din setul de date K centroizi

Etapa 2. **Repeat**

- Asignează fiecare dată la clusterul corespunzător celui mai apropiat centroid
- Recalculează centroizii (ca medii aritmetice ale datelor din fiecare cluster)

until partiția s-a stabilizat

Observatii:

- Rezultatul grupării depinde de pozițiile inițiale ale centroizilor
- Procesul iterativ de la kMeans are ca scop minimizarea varianței într-cluster (suma pătratelor distanțelor dintre datele și centroidul corespunzător unui cluster)
- kMeans este adecvat pentru identificarea clusterelor “sferice” însă nu este adecvat în cazul clusterelor cu formă arbitrară

Implementare Weka:

- **SimpleKMeans**: varianta clasică a algoritmului; se poate opta între distanța euclidiană și distanța Manhattan
- **XMeans**: este o variantă care estimează numărul de clustere (utilizatorul specifică o valoare minimă și o valoare maximă pentru numărul de clustere iar XMeans aplică KMeans pentru fiecare dintre valori și o alege pe cea care conduce la modelul de partiționare

Exercițiul 1:

- a) Deschideți fișierul “iris.2D.arff” și eliminați atributul de clasă
- b) Identificați 3 grupuri în date aplicând algoritmul kMeans (Cluster->SimpleKMeans). Vizualizați clusterelor identificate prin click dreapta pe rezultat (din Result list) și [Visualize Cluster Assignments](#)
- c) Comparați valorile obținute pentru varianța intra-cluster (SSE = within cluster squared sum of errors) pentru diferite valori ale numărului de cluster (parametrul -N de la SimpleKMeans): 2,3,4,5
- d) Aplicați pentru același set de date algoritmul XMeans specificând 1 pentru numărul minim de cluster și respectiv 10 pentru numărul maxim de cluster
- e) Aplicați pentru același set de date algoritmul EM (Expectation-Maximization) cu valorile implicite ale parametrilor

3. Algoritmi ierarhici

Algoritmii ierarhici permit construirea unei ierarhii de partiții (dendrogramă) obținută în una dintre variantele:

- Aglomerativă (bottom-up): se pornește de la partiția în care fiecare cluster conține o singură dată și succesiv se grupează clusterelor folosind un criteriu de similaritate (single-link, complete-link, average-link) – la fiecare etapă se pot reuni două sau mai multe clusterelor. Procesul continuă până când se ajunge la nodul rădăcină al structurii care corespunde întregului set de date.
- Divizivă (top-down): se pornește de la un cluster unic ce conține toate datele și succesiv se aplică o strategie de partiționare (poate fi bazată pe un algoritm partițional de tip kMeans)

Exercițiul 2:

- a) Deschideți fișierul “data.arff”
- b) Construiți și comparați dendrogramele corespunzătoare cazurilor în care se utilizează single-link, complete-link, average-link. Indicație: se selectează Cluster->Hierarchical. Pentru vizualizarea arborelui prin click dreapta pe numele rezultatului (în Result List) se selectează VisualizeTree

4. Algoritmi de grupare bazați pe densitate

Ideea algoritmilor bazati pe densitate este de a clasifica datele ca facând parte dintr-un cluster (core points, border points) sau ca fiind zgomot pe baza densității datelor. Pentru fiecare punct se estimează densitatea numărând câte date sunt într-o vecinătate de raza dată.

Obs: sunt utilizați în special pentru gruparea datelor spațiale și permit identificarea de clusterelor de forme arbitrare.

Exercițiul 3:

- a) Deschideți fișierul “iris.2D.arff”
- b) Aplicați algoritmul DBSCAN pentru diferite valori ale parametrilor Eps (raza vecinătății) și MinPoints (numărul minim de date în vecinătate) și comparați rezultatele. Valori de test: Eps în {0.2,0.4,0.6,0.8}, MinPoints în {5,10,15,20,30}