Referat Algoritmul K-means

Student: Sacu Danut Sacu

Grupa: 3143 A

Facultatea: FIESC / Calculatoare

**Algoritmi de grupare**

(clustering)

Dandu-se puncte intr-un spatiu oarecare – deseori un spatiu cu foarte multe dimensiuni – grupeaza punctele intr-un numar mic de clustere, fiecare cluster constand din puncte care sunt “apropiate” intr-un anume sens.

Cateva explicatii:

1. Cu multi ani in urma, in timpul unei izbucniri a holerei in Londra, un medic a marcat localizarea cazurilor pe o harta, obtinand un desen care arata ca in FIG. 14. Vizualizate corespunzator, datele au indicat ca aparitiile cazurilor se grupeaza in jurul unor intersectii , unde existau puturi infestate, aratand nu numai cauza holerei ci indicand si ce e de facut pentru rezolvarea problemei. Din pacate nu toate problemele de data mining sunt atat de simple, deseori deoarece clusterele sunt in atat de multe dimeansiuni incat vizualiarea este foarte dificila.

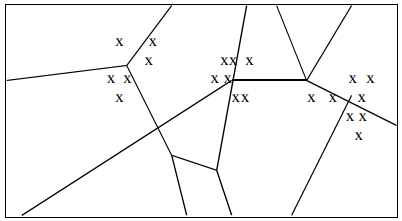


Figura 14: Clustere ale cazurilor de holera au indicat unde erau puturile infestate.

1. Skycat a grupat in clustere 2x10^9 obiecte ceresti in stele, galaxii, etc. Fiecare obiect era un punct intr-un spatiu cu 7 dimensiuni, unde fiecare dimensiune reprezenta nivelul radiatiei intr-o banda a spectruului. Proiectul Sloan Sky Survey este o incercare mult mai ambitioasa de a cataloga si grupa intregul univers vizibil.
2. Documentele pot fi percepute ca puncte intr-un spatiu multi- dimeansional in care fiecare dimensiune corespunde unui cuvant posibil. Pozitia documentului intr-o dimensiune este data de numarul de ori in care cuvantul apare in document (sau doar 1 daca apare, 0 daca nu). Clusterele de documente in acest spatiu corespund deseori cu grupuri de documente din acelasi domeniu.

**Abordari pentru grupare (clustering)**

La nivel inalt, putem imparti algoritmii de grupare in doua mari clase.

1. **Abordarea pe tip centroid**. Ghicim centroizii sau punctele centrale pentru fiecare cluster si asignam punctele la clusterul avand cel mai apropiat centroid.
2. Abordarea ierarhica. Incepem prin a considera ca fiecare punct formeaza un cluster. Comasam repetat clusterele apropiate prin folosirea unei masuri pentru apropierea a doua clustere (e.g. distanta dintre controizii lor), sau pentru cat de bun va fi clusterul rezultat (e.g. distanta medie de la punctele din cluster la noul centroid rezultat).

**Algoritmii prezentati:**

Vom considera urmatorii algoritmi de grupare; ei se diferentiaza dupa cum utilizeaza sau nu o distanta euclidiana si dupa abordarea folosita de tip centroid sau ierarhica.

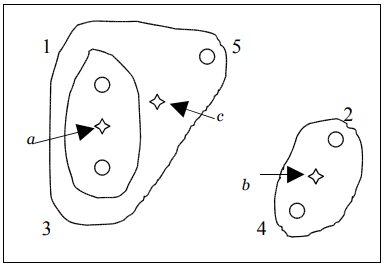
1. BFR [Bradley – Fayyad - Reina]: De tip centroid; utilizeaza o masura euclidiana, cu clustere formate in jurul centroidului printr-un process Gaussian in fiecare dimensiune.
2. Fastmap: Nu e realmente un algoritm de grupare ci un mod de a construe un spatiu Euclidian cu putine dimeansiuni pornind de la o masura a distantei.
3. GRFGPF: De tip centroid dar utilizeaza doar o masura a distantei si nu un spatiu Euclidian.
4. CURE: Ierarhic si Euclidian, acest algoritm se ocupa de clustere avand forme neobisnuite.

**Algoritmul k-means**

Acest algoritm este un algoritm popular care tine datele in memoria centrala si pe care se bazeaza algoritmul BFR. K-means alege k centroizi de cluster si asigneaza punctele la acestea alegand centroidul cel mai apropiat de punctul respectiv. Pe masura ce punctele sunt asignate la un cluster, centroidul acestuia poate migra.

Pentru un exemplu foarte simplu cu cinci puncte in doua dimensiuni sa privim fig. 15.

Presupuneam ca asignam punctele 1,2,3,4,5 in aceasta ordine, cu k = 2. Atunci punctele 1 si 2 sunt asignate celor doua clustere si devin centroidul lor pentru moment.



Cand consideram punctul 3, sa presupunem ca este mai apropiat de 1, deci 3 se adauga clusterului continand 1 iar centroidul acestuia se muta in punctul marcat ca a. Presupunem ca atunci cand asignam 4 gasim ca 4 este mai aproape de 2 decat de a, deci 4 se alatura lui 2 in clusterul acestuia iar centrul se muta in b. In final, 5 este mai aproape de a decat de b, deci el se adauga la clusterul {1,3} al carui centroid se muta in c.

**Exemplul 7.5:**

Sa consideram datele din Fig.16. In mod clar k=3 este numarul correct de clustere dar sa presupunem ca intai incercam k=1. In acest caz toate punctele sunt intr-un singur cluster si distanta medie la centroid va fi mare.

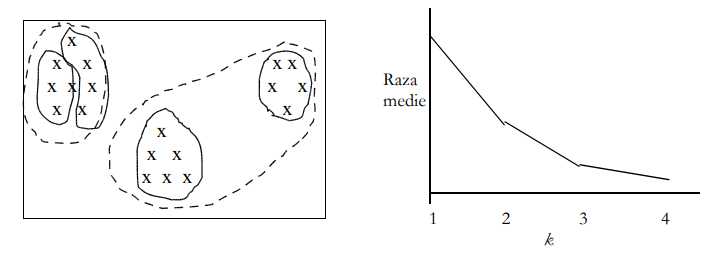


Figura 16. Descoperim ca numarul correct de clustere este k = 3;

Presupunem ca apoi incercam k=2. Unul dintre cele trei clustere va fi un cluster iar celelalte doua vor fi fortate sa creeze un singur cluster, asa cum arata linia punctate. Distanta medie a punctelor la centroid de va micsora astfel considerabil.

Daca luam k=3 atunci fiecare dintre clusterele vizibile va forma un cluster iar distanta medie de la puncta la centroizi se va micsora din nou, asa cum arata graficul din Fig 16. Totusi, daca marim k la 4 unul dintre adevaratele clustere va fi partitionat artificial in doua clustere apropiate, asa cum arata liniile continui. Distanta medie la centroid va scadea putin dar nu mult. Acest esec de a merge mai departe ne arata ca valoarea k=3 este corecta chiar daca datele sunt in atat de multe dimensiuni incat nu putem vvizualiza clusterele.

Bibliografie:

<http://cifr.cs.pub.ro/ullman/cluster1-ro.pdf>