8. Sailkapen gainbegiratu gabea: clustering

Espezialitatea: Konputazioa, hirugarren ikasmaila
Titulazioa: Informatika Ingeniaritzako Gradua
Konputazio Zientzia eta Adimen Artifiziala saila
Universidad del País Vasco - Euskal Herriko Unibertsitatea

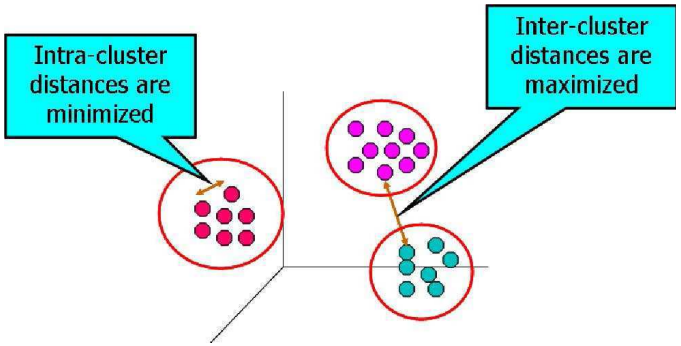
Sailkapen gainbegiratu gabea

- Datubaseko kasuak etiketatu gabe daude
- Zenbat multzotan antolatuta dauden ere ez dakigu

D	X_1	...	X_i	...	X_n
$\mathbf{x}_1$	x_1^1	...	x_i^1	...	x_n^1
...	...	...	...	...	...
$\mathbf{x}_j$	x_1^j	...	x_i^j	...	x_n^j
...	...	...	...	...	...
$\mathbf{x}_N$	x_1^N	...	x_i^N	...	x_n^N

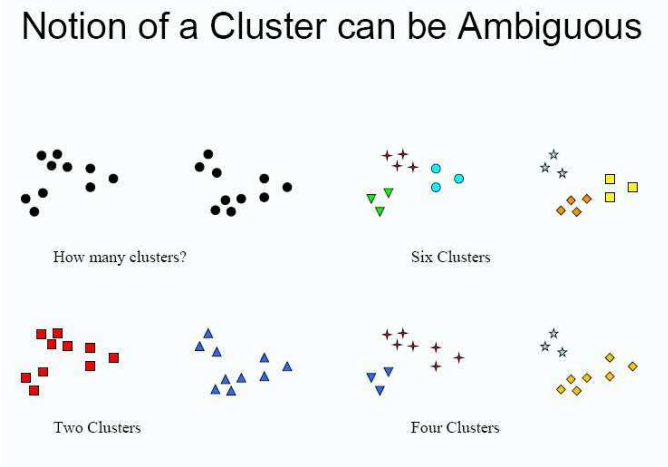
- Zer da cluster bat? Beren artean antzeko diren kasuak biltzen dituen multzoa
- Cluster bateko kasuak ez dira beste clusterretako kasuen antzekoak

Cluster analisia edo clustering



- Cluster analisia: kasuak clusterretan multzokatzea. Clusterrak aurkitzea
- Clustering ona: cluster barruko antzekotasuna altua eta cluster arteko antzekotasuna baxua duena
- Distantzia neurtzeko moduaren arabera → clusterrak

Zenbat cluster daude?



Antzekotasunaren maila finkatzea zaila; subjektiboa da...

Clustering-erako algoritmoak

Partiziozko clustering-a

- Datubaseko $D = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ kasuekin partizio bat k multzo edo cluster sortuko dira, $Cl_1, \ldots Cl_k$, non:
  - Clusterren bildura datubase osoa den: $\bigcup_{j=1}^{k} Cl_j = D$
  - Clusterrak disjuntuak diren: $Cl_j \cap Cl_i = \emptyset, \quad i \neq j$
- Cluster kopurua, k, ezezaguna da; algoritmoak abiaratzeko finkatu behar da. k balio desberdinak probatu

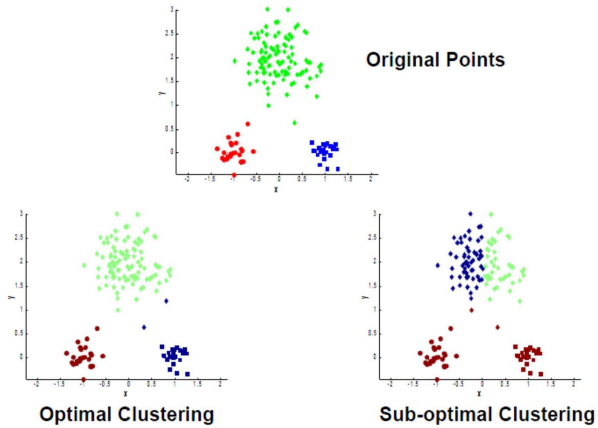
Partiziozko clustering-a. Algoritmoak

Oinarrizko ideiak

- Cluster kopuru bat finkatu: k.
- k kasu aukeratu: haziak. Auzaz? Datubaseko lehenak?
- Datubaseko kasu bakoitza gertueneko haziari dagokion clusterrari egokitu (guk distantzia Euklidestarra)
- Hazietatik hasierako partizio bat lortu: hasierako clusterrak
- Cluster bereko kasuekin clusterraren erdigunea (zentroidea) kalkulatu
- Prozesua errepikatu (datubaseko kasu bakoitza gertueneko zentroideari dagokion clusterrari egokitu), kasuei egokitutako clusterrak egonkortuko diren arte
- Algoritmo erabilienetakoa: K-means (MacQueen, 1967)

Partiziozko clustering-a. Hasierako haziak

Hasierako hazien arabera, clustering desberdinak sortzen dira
Hazi desberdinetarako clustering-ak sortu eta onena aukeratu!!



Clustering hierarkikoa

Ideia nagusiak

- Bi motako clustering hierarkikoa:
  - Behetik gorakoa: Hasieran, kasu bakoitza cluster bat da. Iterazio bakoitzean, gertuen dauden bi clusterrak bakar batean elkartu, cluster bakarra (edo k cluster) lortu arte
  - Goitik beherakoa: Hasieran, kasu guztiak cluster bakar batean. Iterazio bakoitzean, cluster bat bitan banatu clusterrak kasu bakarrekoak (edo k cluster) izan arte
- Sortutako zuhaitz hierarkikoa dendrograma baten bidez adieraz daiteke
- Dendrograma maila desberdinetatik moztean sortzen den osagai konektatu bakoitza cluster bat da.
- Cluster kopuru desberdinetako partiketak lortzen dira; aldez aurretik cluster kopurua finkatu beharrik ez!

Behetik gorako clustering hierarkikoa

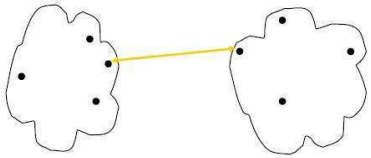
Cluster arteko distantzia

- Algoritmoaren iterazio bakoitzean, gertuen dauden bi clusterrak bakar batean elkartzen dira
- Cluster arteko distantzia neurtzeko moduak:
  - Distantzia minimoa (single linkage): gertueneko bi kasuena
  - Distantzia maximoa (complete linkage): urrunen dauden bi kasuena
  - Batazbestekoaren distantzia (average linkage): bi clusterretako kasuen arteko distantzien batazbestekoa
  - Zentroideen distantzia (centroid linkage): zentroideen artekoa
- Distantzien matrize batean gordetzen dira cluster arteko distantziak
- Bi cluster elkartzen diren bakoitzean, distantzien matrizea eguneratu egin behar da

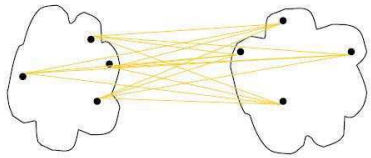
Behetik gorako clustering hierarkikoa

Cluster arteko distantzia. Adibideak

1. Distantzia minimoa (single linkage)



2. Batazbestekoaren distantzia (average linkage)



Clustering hierarkikoa.

Clusterrak sortzen...

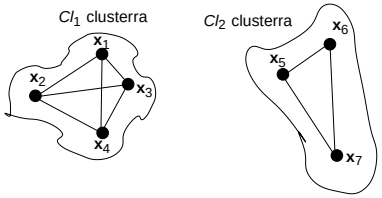
- Cluster arteko distantzia neurtzeko moduaren arabera cluster desberdinak osatuko dira; dendrograma desberdinak lortuko dira.
- Egin proba! Hierarchical clustering interactive demo.

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html

- Behetik gorako clustering hierarkikoan, algoritmoaren urrats batean bi cluster elkartu badira, hala mantenduko dira ondoreneko urratsetan ere; elkartuak izateko erabakia ez da ondoreneko urrats batean atzera botako.
- Dendrogramaren egitura eta indizea aztertuz, datubaseko kasuak zenbat clusterretan banatu nahi diren erabaki daiteke. Cluster kopuru optimoa?

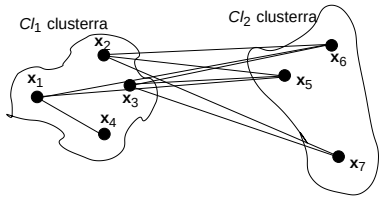
Ebaluazioa. Sortutako clusterrak egokiak al dira?

1. Kanpoko neurketa: clustering algoritmoak sortutako clusterrak zein neurritan bat datozen kanpoko ebaluatzaile batek (aditu batek?) sortutakoekin. Entropia erabil daiteke
2. Barneko neurketa: Kanpo-informaziorik erabili gabe, sortutako clusterren egituraren ontasun maila neurtzea.
   - Cluster kohesioa: cluster bereko kasuen antzekotasuna (SSE: Sum of Squared Error)



Ebaluazioa. Barneko neurketa

- Cluster bereizketa: Sum of Squares Between

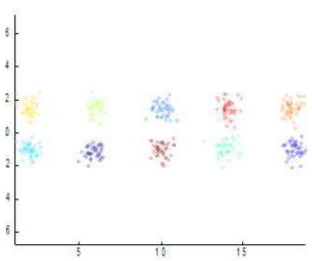


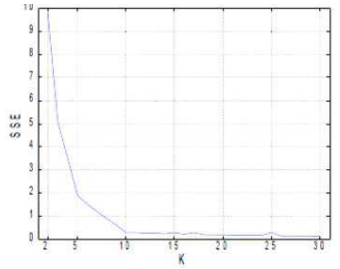
- Silhouette koefizientea: Kohesioa eta bereizketa konbinatzen dituen neurria

Ebaluazioa. Barneko neurketa

SSE neurriaren erabilera

- Clusterren kohesioa neurtzeko erabil daitekeen SSE neurriari esker bi clustering konparatu ahal izango ditugu
- SSE ere erabil daiteke cluster kopuruaren estimazio bat kalkulatzeko.





Ebaluazioa. Kanpoko neurketa. Adibidea

External Measures of Cluster Validity: Entropy and Purity

Table 5.9. K-means Clustering Results for LA Document Data Set

Cluster	Entertainment	Financial	Foreign	Metro	National	Sports	Entropy	Purity
1	3	5	40	506	96	27	1.2270	0.7474
2	4	7	280	29	39	2	1.1472	0.7756
3	1	1	1	7	4	671	0.1813	0.9796
4	10	162	3	119	73	2	1.7487	0.4390
5	331	22	5	70	13	23	1.3976	0.7134
6	5	358	12	212	48	13	1.5523	0.5525
Total	354	555	341	943	273	738	1.1450	0.7203

entropy For each cluster, the class distribution of the data is calculated first, i.e., for cluster j we compute p_{ij}, the 'probability' that a member of cluster j belongs to class i as follows: $p_{ij} = m_{ij}/m_j$, where m_j is the number of values in cluster j and m_{ij} is the number of values of class i in cluster j. Then using this class distribution, the entropy of each cluster j is calculated using the standard formula $e_j = \sum_{i=1}^{L} p_{ij} \log_2 p_{ij}$, where the L is the number of classes. The total entropy for a set of clusters is calculated as the sum of the entropies of each cluster weighted by the size of each cluster, i.e., $e = \sum_{i=1}^{K} \frac{m_i}{m} e_j$, where m_j is the size of cluster j, K is the number of clusters, and m is the total number of data points.

purity Using the terminology derived for entropy, the purity of cluster j, is given by $purity_j = \max p_{ij}$ and the overall purity of a clustering by $purity = \sum_{i=1}^{K} \frac{m_i}{m} purity_j$.

85

Bibliografia

- http://en.wikipedia.org/wiki/Cluster_analysis
- http://en.wikipedia.org/wiki/Silhouette_(clustering)
- A Tutorial on Clustering Algorithms

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/index.html

Hierarchical clustering interactive demo

http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/AppletH.html