

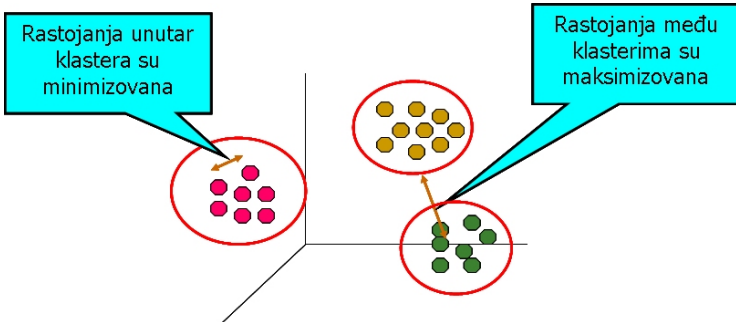
Кластер анализа

Ненад Митић

Математички факултет
`nenad@matf.bg.ac.rs`

Шта је кластер анализа?

Проналажење група објеката таквих да су објекти у групи међусобно слични (или повезани), и да су објекти у различитим групама међусобно различити (или неповезани)



Шта јесте а шта није кластер анализа?

Припадност објеката (елемената) једном кластеру не значи да су елементи међусобно слични по свим критеријумима. Тако, на пример кластери који су приказани на претходној слици су добијени према просторном груписању елемената, међутим, нема никаквих препрека да део елемената једног кластера буде по неком критеријуму сличнији елементима другог кластера него сваком од елемената кластера у коме се налазе.

Двосмисленост појма кластера



Koliko klastera?



Šest klastera



Dva klastera



Četiri klastera



Број кластера зависи од посматраног критеријума. Нпр. на слици се препознају

- два кластера, ако се посматра само просторни положај група
- четири кластера, ако се посматра распоред елемената (елементи означени крстићима и квадратима су распоређени дуж хипотетички правих линија, док елементи означени звездицама и троугловима одступају од тог правила)
- шест кластера, ако се посматра међусобна удаљеност елемената (мерена нпр. као еуклидско растојање) и постави горња граница на растојање два елемента за припадност истом кластеру

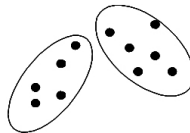
Типови кластеровања

У зависности од карактеристика кластера који се добијају као резултати, постоје различити типови кластеровања.

Код партиционог кластеровања скуп улазних података се дели у непреклапајуће подскупове (кластере) такве да сваки податак припада тачно једном кластеру



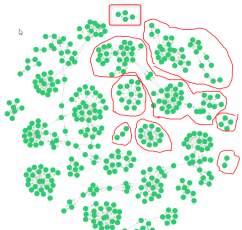
Početni podaci



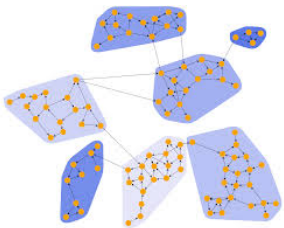
Particiono klasterovanje

Типови кластера (наставак)

Кластери засновани на графовима (енг. *graph based*)



Ако су елементи представљени као чворови повезаног графа, тада кластери могу да буду скупови објеката који су међусобно повезани, али нису повезани са објектима ван групе, односно који припадају изолованом подграфу.



Неке дефиниције допуштају да између кластера (подграфова) постоје везе, али у много мањем броју (или са много већим растојањем) него између елемената подграфа.

Алгоритми засновани на репрезентативним представницима

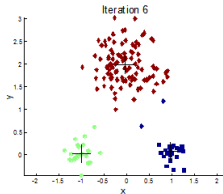
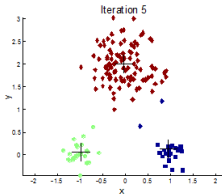
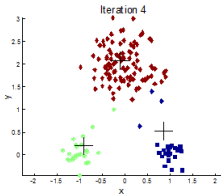
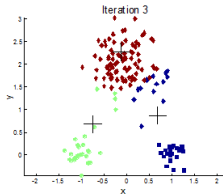
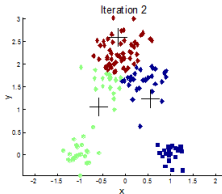
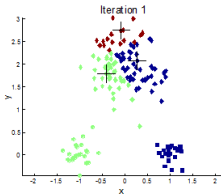
Основни принцип: За скуп од n тачака X_1, X_2, \dots, X_n у d димензионом простору циљ је пронаћи k репрезентативних тачака Y_1, Y_2, \dots, Y_k , где је k које минимизују циљну функцију

$$O = \sum_{i=1}^n [\min_j \text{Dist}(X_i, Y_j)]$$

где је $\text{Dist}(A, B)$ функција растојања

Алгоритам к-средина: пример

Алгоритам к-средина: пример

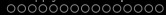


Алгоритам к-средина

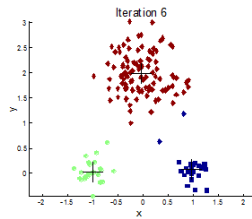
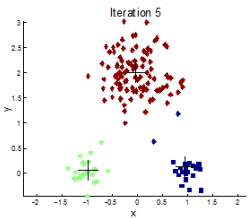
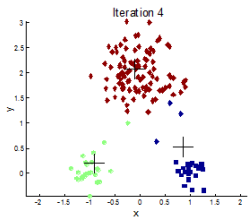
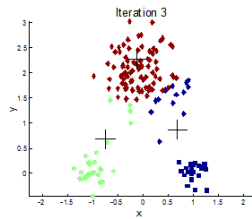
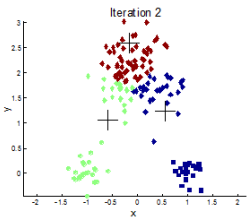
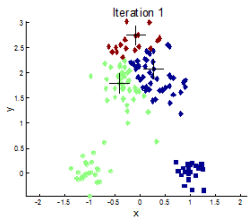
- За одређивање растојања могу да се користе различите мере (обрађене у уводном делу курса!).
- Алгоритам конвергира за поменуте мере, при чему се највећи део конвергенције дешава у првих неколико итерација
- Као услов заустављања алгоритма се задаје број тачака које промене кластер у одређеној итерацији. Ако је број тачака које промене кластер мањи од задатог прага, алгоритам се зауставља.

Алгоритам к-средина

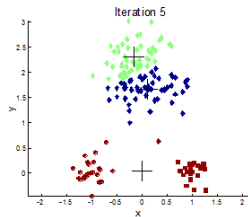
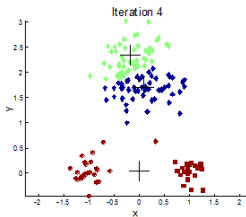
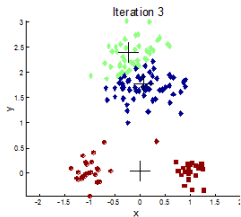
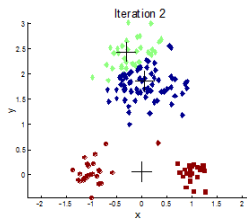
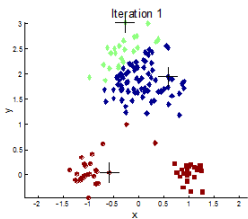
- Један од недостатака методе је често бирање почетних центроида на случајан начин. Резултат оваквог начина избора је добијање кластера који могу да се разликују од 'природних' кластера, односно добијање незадовољавајућих резултата (видети пример у даљем тексту)
- Сложеност: временска $O(n * K * I * d)$, просторна $O((n + K) * d)$ где је n број тачака, K број кластера, I број итерација, и d број атрибута. Одавде се види да је алгоритам k средина релативно неефикасан за материјал са јако великим бројем тачака јер захтева велики број израчунавања.



Важност избора почетног центроида - пример 1



Важност избора почетног центроида - пример 2



Избор почетних центроида

- Ако постоји k 'реалних' кластера тада је вероватноћа да се изабере по један центроид у сваком од њих релативно мала
 - Ако је k велико шанса за добар избор је мала
 - Ако кластери имају исту величину n , тада важи

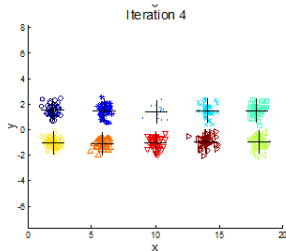
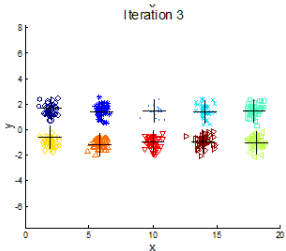
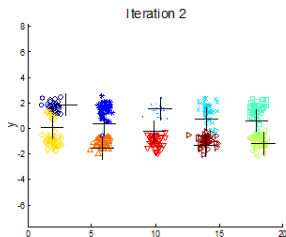
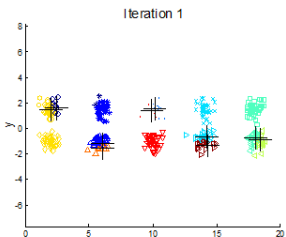
$$P = \frac{\text{број начина за избор центроида у сваком кластеру}}{\text{број начина за избор } k \text{ центроида}}$$

$$P = \frac{k!n^k}{(kn)^k} = \frac{k!}{k^k}$$

- На пример, за $k = 10$, вероватноћа је $10!/10^{10}=0.00036$
- Понекад се иницијални центроиди сами поравнају на 'прави' редослед, а понекад не

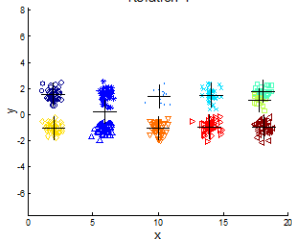
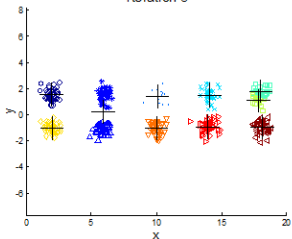
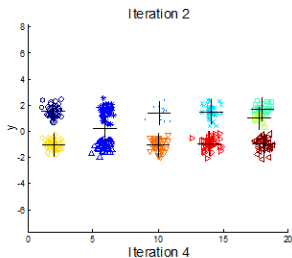
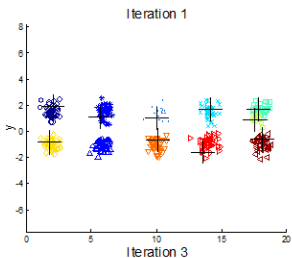
Избор почетних центроида - коректно поравнање

Важност избора почетног центроида - коректно поравнање



Избор почетних центроида - некоректно поравнање

Важност избора почетног центроида - некоректно поравнање



Избор почетних центроида

Различите технике могу да се примене ради побољшања добијених резултат или повећања шанси за добијање квалитетнијих резултата. Један део техника се односи на избор почетних центроида, док је други оријентисан на додатну обраду добијених резултата. Могуће технике су:

- Узастопна извршавања алгорита
 - Свако извршавање са нпр. случајно изабраним центроидима
 - Између њих се изабере кластер са најмањим *SSE*
- Над узорцима се примени хијерархијско кластеровање и изаберу почетни центроиди
- Изабере се m ($m > k$) почетних центроида и бирају се 'добри' центроиди између њих
 - Да би овај начин био успешан потребно је да изабрани кандидати за центроиде покривају што шири простор
- Применити приступ *K-срдине++*
- Применити методу *бисекције K-срдина*
- Извршити постпроцесирање добијених резултата

K-средине++

- Један од начина за иницијализацију скупа иницијалних центроида
- Даје боље резултате у односу на CCE
- Алгоритам може да се представи преко следећег псеудокода:

```
Izabrati slucajno tacku kako prvi centroid
for i=1 to k do
  Odrediti rastojanje  $d(x)$  svake (do sada
    neizabrane) tacke do najblizeg centroida
  Svakoј tacki dodeliti verovatnocu proporcionalnu
    njenom  $d(x)^2$ 
  Izabrati novi centroid od preostalih tacaka
    koristeći verovatnoce kao tezine
end for
```

Постпроцесирање добијених резултата

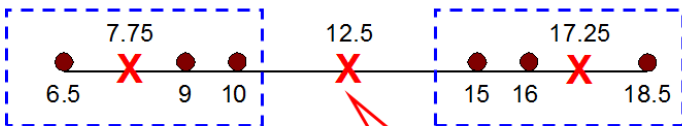
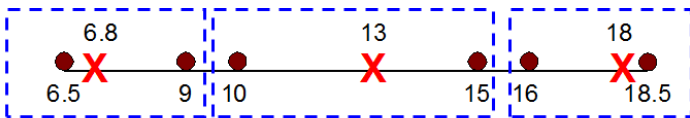
Доста често, елементи ван граница могу знатно да утичу на лоше резултате кластерованја. Квалитет кластерованја може да се додатно побољша анализом резултата и уклањањем елемената ван граница. Уклањање може да се изврши и у фази препроцесирања података. При томе треба бити опрезан, јер уклањање елемента ван граница не важи за сваку врсту апликација (нпр. не важи у случају компресије података).

Додатне технике постпроцесирања које доводе до побољшања резултата су:

- Елиминација малих кластера са елементима ван граница
- Подела кластера са високим *SSE*
- Интеграција кластера који су 'близу' и имају релативно мали *SSE*

Рад са празним кластерима

Основни алгоритам k -средина може да произведе празне кластере при извршавању. При томе 'празан' означава да се у том кластеру налази само центроид, без иједног елемента.



**Prazan
klaster**

Рад са празним кластерима

Стратегије за елиминацију празних кластера укључују замену центроида на неки од следећих начина:

- Изабрати тачку која највише учествује у SSE
- Изабрати тачку која је најдаље од текућих центроида
- Изабрати тачку из кластера са највећим SSE . Овај начин обично доводи до деобе кластера
- Ако има више празних кластера поновити поступак

Алгоритам бисекције K -средина

Алгоритам бисекције k -средина је варијанта алгоритма k -средина која може да произведе партиционо или хијерархијско кластеровање

Основна идеја: за добијање k кластера подели се скуп свих тачака у два кластера, изабере се један од њих за поделу, уз понављање поступка све док се не добије K кластера. Различити начини поделе кластера су:

- подели се највећи кластер
- подели се кластер са највећим SSE
- користи се критеријум заснован и на величини кластера и на величини SSE-а

Ова метода се често не користи за само кластеровање, већ се добијени центроиди користе за улаз у основни K -средина алгоритам кластеровања

Алгоритам бисекције K-средина - пример

Недостаци алгоритма k -средина

Недостаци и ограничења алгоритма k -средина су

- не функционише за кластере произвољног облика
- не функционише за кластере различитих густина
- осетљив је на елементе ван граница који могу да доведу до јединичних или празних кластера
- проблем представља одређивање репрезентативних представника и броја кластера k

Добре стране алгоритма k -средина

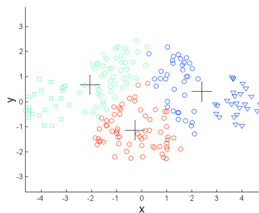
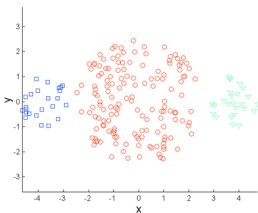
Добре стране алгоритма k -средина су

- Једноставност имплементације и примене
- Најбоље ради са глобуларним подацима
- Ако се као мера растојања користи Махаланобисово растојање, алгоритам k -средина препознаје кластере различитих густина

Неки недостаци и ограничења алгоритма k -средина су илустровани на наредним слајдовима. У сва три случаја приказана ограничења могу да се превазиђу повећањем броја кластера k и налажењем кластера који су подкластери природних кластера.

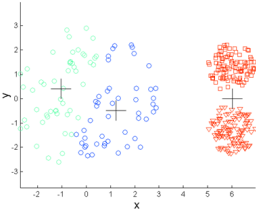
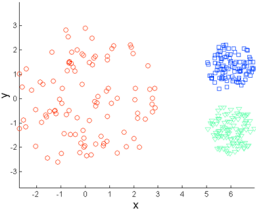
Ограничења алгорита k -средина

Примена алгорита k -средина на кластере различитих величина



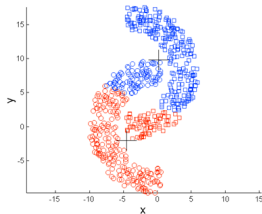
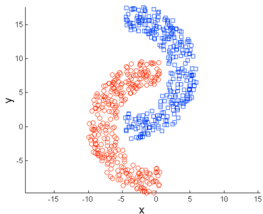
Ограничења алгорита k -средина

Примена алгорита k -средина на кластере различитих густина

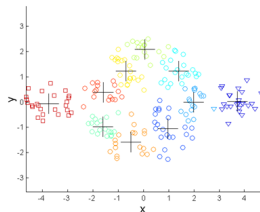
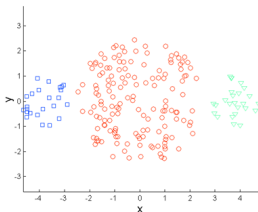


Ограничења алгорита k -средина

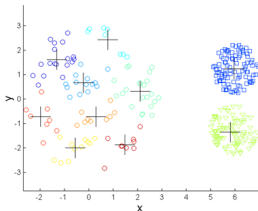
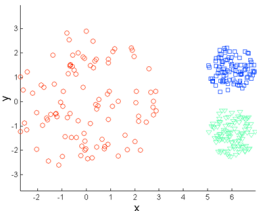
Примена алгорита k -средина на не-глобуларне кластере



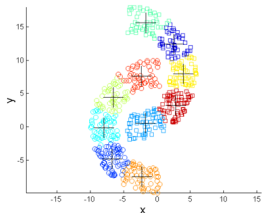
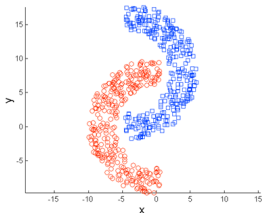
Превазилажење ограничења алг. к-средина



Превазилажење ограничења алг. к-средина



Превазилажење ограничења алг. к-средина



Алгоритам k -медијана

Алгоритам k -медијана је сличан алгоритму k -средина, при чему се као центроид користи медијана. Неке карактеристике овог алгоритма су:

- Користи се растојање *такси блок*.
- Показује се да репрезентативни представник медијана података по свакој димензији кластера C_j
- Мања је осетљивост на елементе ван граница

Алгоритам k -медоида

У алгоритму k -медоида избор центроида се увек врши из инцијалног скупа тачака. Иако није оптималан, разлог за овакав избор је утицај елемената ван граница на медијану, мада је понекад тешко израчунати центар за одређене (сложене) типове података. Као и код k -медијане и у овом алгоритму се као мера растојања користи *такси блок*.

Пример алгоритма k -медоида је приказан на наредном слајду.

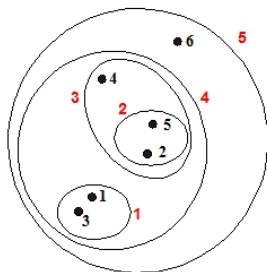
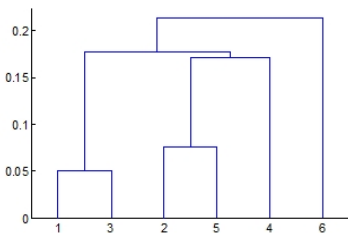
Детаљније информације о алгоритму k -медоида могу да се нађу у књизи Charu C. Aggarwal: Data Mining The Textbook, Springer, 2015

Алгоритам к-медоида

```
/* Skup podataka:  $D=\{x_1, x_2, \dots, x_n\}$ ,  
   Broj reprezent. predstavnika:  $k$  */  
klasterovanje_sa_reprezentativnim_predstavnicima(D, k)  
begin  
  inicijalni izbor skupa reprezentativnih predstavnika  
     $S=\{Y_1, Y_2, \dots, Y_k\}$  iz skupa  $D$ ;  
  repeat  
    Formiraj klustere ( $C_1, \dots, C_k$ ) dodelom svake tacke  
      iz  $D$  najblizem predstavniku iz  $S$  koristeći  
      funkciju rastojanja  $\text{Dist}(x_i, Y_j)$ ;  
    Odrediti par  $x_i$  iz  $D$  i  $Y_j$  iz  $S$  tako da zamena  
       $Y_j$  sa  $x_i$  daje najbolje moguće povećanje  
      ciljne funkcije;  
    Izvršiti zamenu  $X_i$  i  $Y_j$  samo ako je  
      povećanje pozitivno;  
  until nema poboljšanja vrednosti funkcije;  
  return ( $C_1, \dots, C_k$ );  
end
```

Алгоритми хијерархијског кластеровања

Другу велику групу алгоритама за кластеровање чине алгоритми хијерархијског кластеровања. Основна идеја ове групе алгоритама је формирање скупа угнеждених кластера који су организовани у облику хијерахије по нивоима. Резултати ове групе алгоритама се најчешће визуелизују у облику дендрограма или дијаграма са угнежденим кластерима на коме се, поред односа кластер/подкластер, види и редослед формирања кластера, односно која два подкластера се спајају у кластер. На наредној слици приказани су резултати кластеровања скупа од 6 тачака у облику дендрограма и дијаграма са угнежденим кластерима.



Алгоритми хијерархијског кластерованја

Алгоритми хијерархијског кластерованја се деле у две групе:

- алгоритме сакупљајућег кластерованја (енг. *agglomerative clustering*). Код ове групе алгоритама хијерархија кластера се формира одоздо/навише. На почетку алгоритма се свака тачка посматра као посебан кластер, и у сваком од наредних корака се врши спајање два најближа (према неком критеријуму) кластера.
- алгоритме раздвајајућег кластерованја (енг. *divisive clustering*). Код ове групе алгоритама хијерархија кластера се формира одозго/наниже. На почетку алгоритма се сматра да све тачке припадају једном кластеру (на врхи хијерархије) који се у наредним корацима дели све док се не дође до кластера који садрже само појединачне тачке.
- Традиционални хијерархијски алгоритми деле или спајају по један кластер у једном кораку и користе матрице сличности или матрице растојања (два кластера су најсличнија ако је растојање између њих најмање и обратно)
- Карактеристика ове групе алгоритама је да се иницијално не наводи број кластера који ће се формирати

Алгоритми сакупљајућег кластеровања

- Заједничка особина свих алгоритама хијерахијског кластеровања је начин формирања хијерахије - креће се од појединачних тачака које се посматрају као јединични кластери. У сваком од наредних корака алгоритама се врши сакупљање најближег пара кластера у нови кластер све док не остане један кластер
- Имплементације неких алгоритама допуштају тзв. пресецање, односно раније заустављање алгорита уколико се дође до k кластера или се дође до l -тог нивоа хијерархије, где су k и l унапред задати бројеви
- Главна разлика између алгоритама овог типа је избор функције/методе за израчунавање сличности на основу које се врши спајање два кластера. Од овог избора зависи и облик добијене хијерархије.
- За одређивање растојања могу да се користе различите мере: растојање Минковског (Еуклидско растојање, такси блок, ...), Махаланобисово растојање,

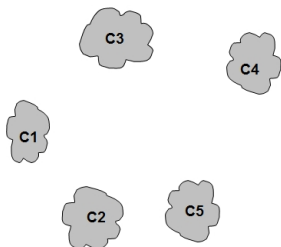
Алгоритми сакупљајућег кластеровања

Псеудокод алгоритма сакупљајућег хијерахијског кластеровања

```
/* Podatak: D, matrica slicnosti (rastojanja) M */
Sakupljajuce_klasterovanje(D)
begin
  inicijalizacija matrice slicnosti M dimenzije n x n
    na osnovu podataka D;
  repeat
    Uzeti najblizi par klastera i i j koristeći M;
    Kombinovati klastera i i j;
    Obrisati redove i kolone klastera i i j iz M i
      formirati novi red i kolonu u M za
      novodobijeni klaster;
    Uneti novi red i kolonu u M;
  until kriterijum izlaska;
  return tekuci skup klastera;
end
```

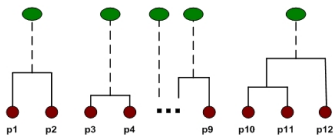
Алгоритми сакупљајућег кластеровања

- Проблем - чување матрице растојања
- Формирање новог кластера - матрица се модификује (или се прави нова)

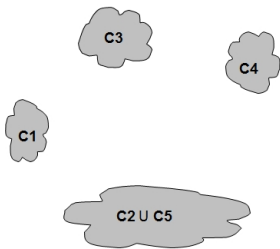


	C1	C2	C3	C4	C5
C1					
C2					
C3					
C4					
C5					

Proximity Matrix

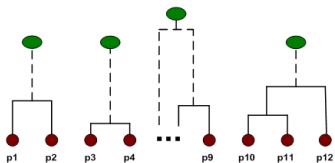


Алгоритми сакупљајућег кластеровања



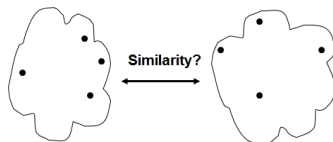
		C2 U C5	C3	C4
C1		?		
C2 U C5	?	?	?	?
C3		?		
C4		?		

Proximity Matrix



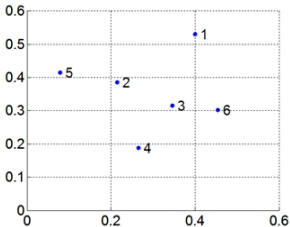
Сличност кластера

У складу са општим принципом за извршавање алгоритама (сакупљајућег) хијерархијског кластеровања који подразумева спајање два кластера која су најсличнија (на најмањем растојању), поставља се питање како дефинисати функције и методе за израчунавање сличности (растојања) два кластера P и Q са m и n елемената.



Сличност кластера

На основу матрице сличности текућег скупа кластера, свака од функција која се користи израчунава сличност између пара кластера и добијену вредност уписује у матрицу сличности.



Distance Matrix:

	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00

Методе за израчунавање сличности кластера

Неке од метода које се могу користити за израчунавање растојања два кластера P и Q са m и n елемената су:

- 1) Растојање између кластера P и Q је једнако најмањем растојању између било које две тачке које им припадају, тј.

$$D(P, Q) = \min(d(p_i, q_j)), i = 1, \dots, m, j = 1, \dots, n, p_i \in P, q_j \in Q$$

где је $d(x, y)$ мера за растојање између две тачке, а m и n број тачака у кластерима P и Q (*најбоља, најкраћа, појединачна веза*)

- 2) Растојање између кластера је једнако највећем растојању између било које две тачке које припадају кластерима P и Q , тј.

$$D(P, Q) = \max(d(p_i, q_j)), i = 1, \dots, m, j = 1, \dots, n, p_i \in P, q_j \in Q$$

где је $d(x, y)$ мера за растојање између две тачке, а m и n број тачака у кластерима P и Q (*најгора, најдужа, комплетна веза*)

Методе за израчунавање сличности кластера

- 3) Растојање између кластера P и Q је једнако просечном растојању тачака које им припадају, тј.

$$D(P, Q) = \frac{\sum_{p_i \in P, q_j \in Q, i=1, \dots, m, j=1, \dots, n} d(p_i, q_j)}{m \times n}$$

где је $d(x, y)$ мера за растојање између две тачке, а m и n број тачака у кластерима P и Q .

У рачунање просечног растојања могу да буду укључени и тежински фактори, при чему тежине зависе од броја тачака у сваком кластеру.

Модификација: Растојање између кластера P и Q је једнако просечном растојању свих парова тачака које су присутни у оба кластера, тј.

$$D(P, Q) = \frac{\sum_{i \in P \cup Q} \sum_{j \in P \cup Q, i \neq j} d(i, j)}{(m+n) \times (m+n-1)}$$

Методе за израчунавање сличности кластера

- 4) Растојање између кластера P и Q је једнако растојању између њихових центроида кластера. У наредном кораку се спајају два кластера чији су центроиди најближи. Лоше особине
- не прави разлику код спајања између кластера различитих величина ако је растојање њихових центроида једнако
 - након спајања се поново израчунавају центроиди и може да се деси да је растојање између (центроида) два кластера на нивоу k мање него растојање (центроида) неких кластера на нивоу $k - 1$

Уместо центроида за растојање кластера могу да се користе и медијане.

- 5) Нови кластер се формира од два кластера чијим спајањем се минимизује промена варијансе унутар новодобијеног кластера. Промене у варијанси после спајања кластера могу да се изразе формулом

$$\nabla SE_{ij} = SE_{ij} - SE_i - SE_j$$

где SE_x означава просечну квадратну грешку кластера x .

Методе за израчунавање сличности кластера

- 6) *Ward*-ов метод и коме се уместо варијансе користи збир квадрата грешака. Нови кластер се формира од два кластера чијим спајањем се добија минимално повећање збира квадрата грешака унутар новог кластера. Тако, у општем случају грешка E је једнака

$$E = \sum_{k=1}^K \sum_{x_i \in C_k} \|x_i - c_k\|^2$$

где је K број кластера, и c_k центроид кластера C_k добијеног спајањем два кластера. Промена се рачуна само за новодобијени кластер на основу кластера од којих је настао и износи

$$\nabla E_{ij} = \frac{m_i n_j}{m_i + n_j} \|c_i - c_j\|^2$$

где m_i и n_j представљају број тачака у кластерима C_i и C_j .

Особине метода за рачунање сличности

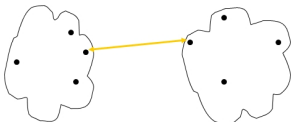
Неки недостаци сакупљајућег хијерархијског кластеровања

- После комбиновања кластери не могу да се раздвоје
 - Превазилажење: коришћење партиционог кластеровања за формирање мањих кластера на које се врши хијерархијско кластеровање (нпр. TwoStep алгоритам у SPSS Modeleru)
- Не постоји глобална функција која се директно минимизује
- У зависности од рачунања растојања јављају се
 - осетљивост на шум и елементе ван граница
 - тешкоће у обради кластера различитих величина
 - тешкоће у обради неглобуларних кластера
 - тенденција ка разбијању великих кластера

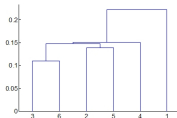
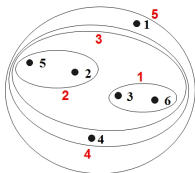
На наредним слајдовима биће илустроване предности и недостаци различитих метода за рачунање сличности кластера.

Особине метода за рачунање сличности

Најбоља (најкраћа, појединачна) веза се рачуна као минимум растојања кластера из матрице сличности. Тако би, за дата тачке и њихова растојања



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00



У првом кораку били спојени кластери (тачке) 3 и 6 јер је њихово међусобно растојање најмање (0.11), па иза њих кластери 2 и 5 (растојање 0.14), итд.

Растојање између тако добијених кластера би се рачунало по истом принципу као

$$\begin{aligned}
 \text{dist}(\{3,6\},\{2,5\}) &= \min(\text{dist}(3,2),\text{dist}(6,2),\text{dist}(3,5),\text{dist}(6,5)) \\
 &= \min(0.15,0.25,0.28,0.39) \\
 &= 0.15
 \end{aligned}$$

Особине метода за рачунање сличности

Погодност најбоље везе: може да обради не-елиптичке кластере



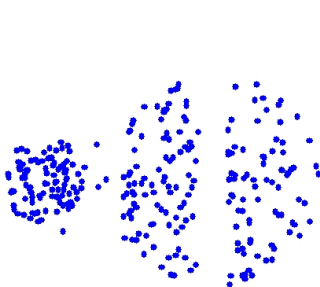
Original Points



Six Clusters

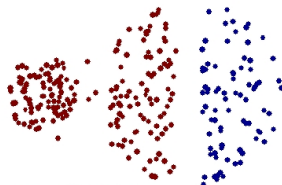
Особине метода за рачунање сличности

Недостаци најбоље везе: осетљивост на шум и елементе ван граница

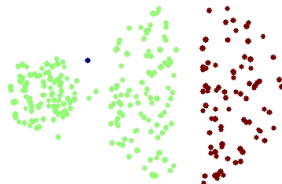


Original Points

- Sensitive to noise and outliers



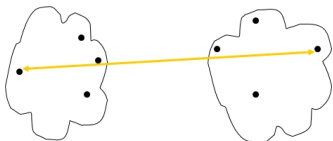
Two Clusters



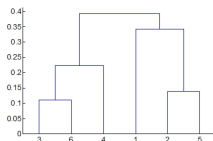
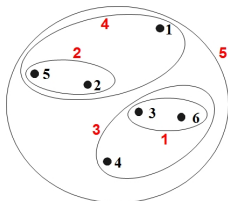
Three Clusters

Особине метода за рачунање сличности

Најгора (најдужа, комплетна) веза



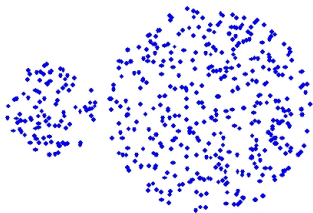
	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00



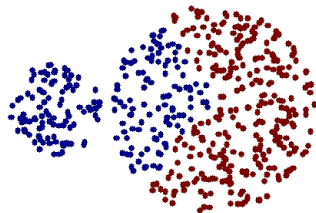
Домаћи задатак: одредити растојање кластера $\{3,6\}$ и $\{2,5\}$, и растојање кластера $\{3,6\}$ и $\{4\}$.

Особине метода за рачунање сличности

Недостатак најгоре везе: тенденција разбијања великих кластера и нагињање глобуларним кластерима



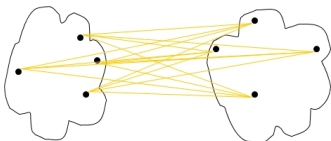
Original Points



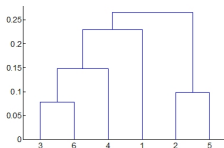
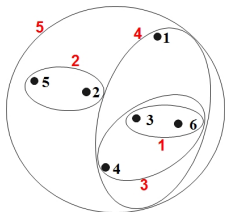
Two Clusters

Особине метода за рачунање сличности

Просек растојања парова елемената из два кластера



	p1	p2	p3	p4	p5	p6
p1	0.00	0.24	0.22	0.37	0.34	0.23
p2	0.24	0.00	0.15	0.20	0.14	0.25
p3	0.22	0.15	0.00	0.15	0.28	0.11
p4	0.37	0.20	0.15	0.00	0.29	0.22
p5	0.34	0.14	0.28	0.29	0.00	0.39
p6	0.23	0.25	0.11	0.22	0.39	0.00



Домаћи задатак: одредити растојање кластера $\{3,6\}$ од осталих кластера.

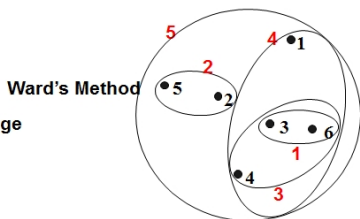
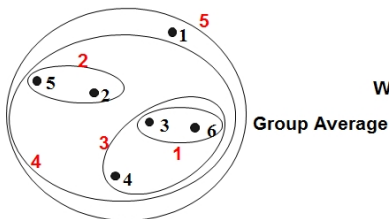
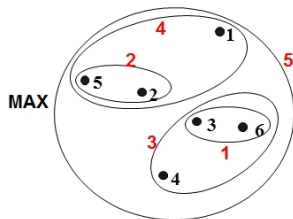
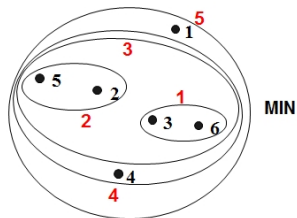
Особине метода за рачунање сличности

Просек растојања парова елемената из два кластера

- Компромис између појединачне и комплетне везе
- Погодност: мање је осетљива на шум и елементе ван граница
- Недостаци: наклоност ка глобуларним кластерима

Особине метода за рачунање сличности

Резултати кластеровања различитим методама



Временска и просторна сложеност

Временска и просторна сложеност

- Просторна $O(n^2)$ где је n број тачака (због матрице сличности)
- Временска $O(n^3)$: n корака у којима се рачунају елементи матрице сличности ($O(n^2)$)
- Ако се растојања за сваки кластер чувају у облику сортиране листе, цена тражења сличних кластера у итом кораку може да се смањи на $(n - i + 1)$, а због додатне цене чувања елемената у облику листе, укупна цена може да се смањи на $O(n^2 \log(n))$

Lance-Williams-ова формула за сличност кластера

Све претходно наведене функције за рачунање сличности могу да се посматрају као вредности *Lance-Williams*-ова формуле за изабране параметре.

Ова особина даје могућност да све технике сакупљајућег кластеровања које могу да се представе *Lance-Williams*-овом формулом не морају да чувају оригиналне тачке, већ је могуће да се матрица сличности ажурира код сваког спајања.

Lance-Williams-ова формула за сличност кластера

Нека је кластер R добијен спајањем кластера A и B , и нека је $p(\dots)$ функција сличности. Сличност кластера R и Q је једнака

$$p(R, Q) = \alpha_A p(A, Q) + \alpha_B p(B, Q) + \beta p(A, B) + \gamma |p(A, Q) - p(B, Q)|$$

Метода	α_A	α_B	β	γ
Појединачна веза	1/2	1/2	0	-1/2
Комплетна веза	1/2	1/2	0	1/2
Просек групе (UPGMA)	$\frac{n_a}{n_a+n_b}$	$\frac{n_b}{n_a+n_b}$	0	0
Тежински просек гр. (WPGMA)	1/2	1/2	0	0
Центроид (UPGMC)	$\frac{n_a}{n_a+n_b}$	$\frac{n_b}{n_a+n_b}$	$\frac{-n_a \times n_b}{(n_a+n_b)^2}$	0
Медијана (WPGMC)	1/2	1/2	-1/4	0
Ward-ова метода	$\frac{n_a+n_q}{n_a+n_b+n_q}$	$\frac{n_b+n_q}{n_a+n_b+n_q}$	$\frac{-n_q}{n_a+n_b+n_q}$	0

где су n_a, n_b и n_q бројеви елемената у кластерима A, B и Q

Ознаке - U: unweighted; W: weighted; PGM: pair group method; A: average; C: centroid