Алгоритми кластеровања

Ненад Митић

Математички факултет
nenad@matf.bg.ac.rs

Алгоритми раздвајајућег кластеровања

- Почињу се са једним кластером који укључује све тачке
- У сваком кораку се кластер дели све док се не дође до тога да сваки кластер садржи само једну тачку или док се не јави k кластера
- За деобу може да послужи било који алгоритам кластеровања (на равном скупу података)
- Пример представника ове групе алгоритама је *алгоритам бисекције к-средина* (код њега се подела увек врши на 2 кластера)

Алгоритми раздвајајућег кластеровања

Псеудокод општег алгоритма раздвајајућег хијерахијског кластеровања

```
/* Podatak: D, Klasterovanje ravnih podataka: A */
Razdvajajuce_klasterovanje(D, A)
begin
  inicijalizovati drvo T tako da koren sadrzi D;
  repeat
    Izabrati list drveta L u T na osnovu
        predefinisane strategije;
    Koristeci algoritam A razdvojiti L na L1, ..., Lk;
    Dodati L1, ..., Lk kao decu cvora L u T;
  until kriterijum izlaska;
end
```

Алгоритми раздвајајућег хијерархијског кластеровања

Фактори који утичу на кластеровање

- Критеријум поделе (може да се користи нпр. Ward-ов K-средина - смањењ SSE)
- Метод поделе (нпр. бисекција K-средина)
- Избор наредног кластера за поделу (нпр. провера квадрата грешака и подела оног са највећом грешком)
- Обрада шума

Алгоритми раздвајајућег кластеровања

Псеудокод основног алгоритма, верзија 2

```
/* Skup slogova D */
Razdvajajuce_klasterovanje(D);
begin
  Pocetno stanje: koreni cvor sadrzi sve slogove
  repeat
    Podeliti roditelj cvor na dva dela C1 i C2
       koristeci bisekciju K sredina za
       maksimizaciju WARD-ovog rastojanja   (C1,C2)
    Konstruisati dendogram. Iz trenutnog skupa klastera
       izabrati onaj sa najvecom greskom
  until dok se ne dobiju jedinicni klasteri
end
```

Minimum Spanning Tree (MST)

- У графу са тежинама MST је ациклични подграф који садржи све чворове са минималном тежином грана
- Алгоритам Prim-a и Kruskal-a се користи за налажење MST у графу са тежинама
- У Еуклидском MST (EMST) гране представљају најкраће (Еуклидско) растојање између две тачке
- Деоба се врши уклањањем гране са највећом тежином
- MST кластеровање може да пронаже несферичне кластере

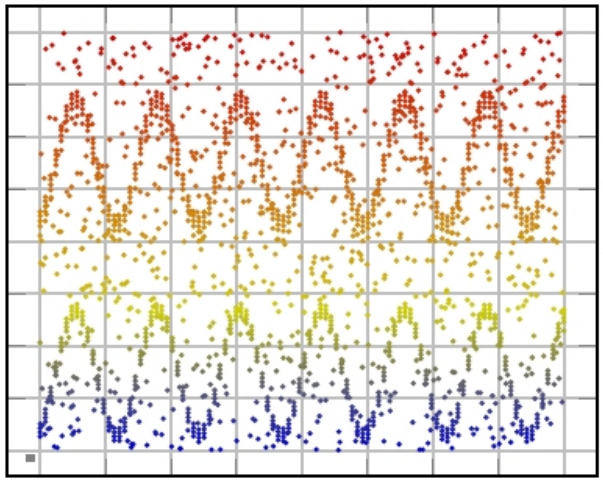
Алгоритми засновани на мрежама и густини

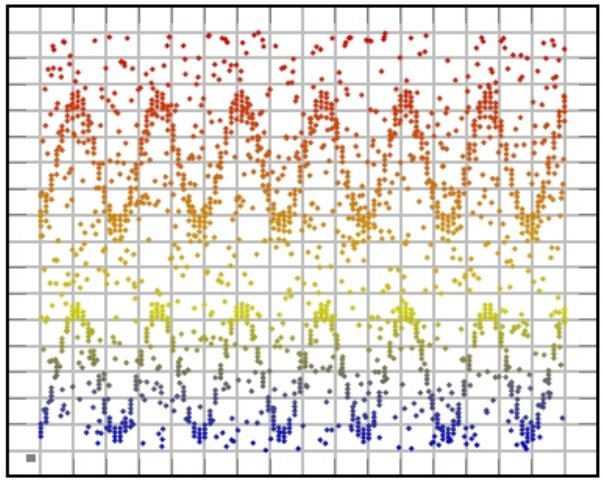
Алгоритми засновани на мрежама и густини се користе када треба одредити регионе (кластере) са високом густином тачака који су раздвојени регионима (кластерима) са мањом густином тачака. Карактеристика ове групе алгоритама је да ефикасно одређују кластер који су различитог облика.

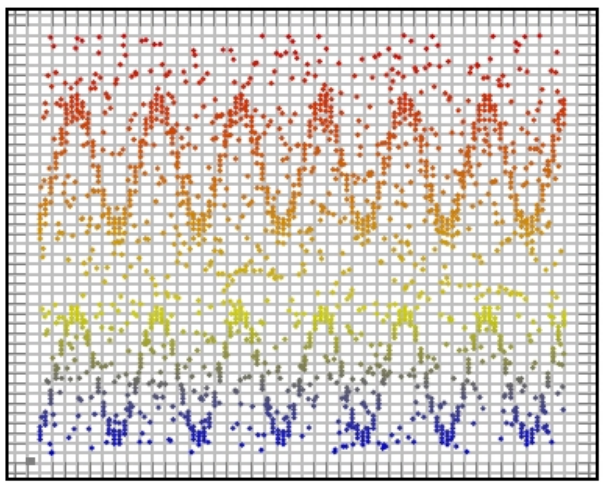
Основна идеја је следећа:

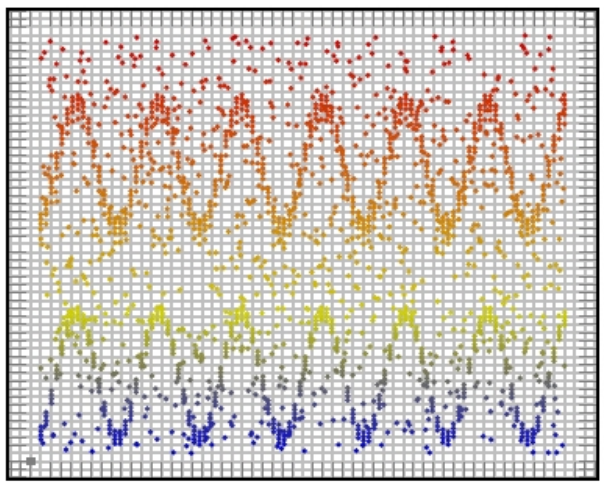
- Простор се подели на целине (ћелије, хиперкоцке) помоћу решетке или граничника другог облика. Добијене ћелије могу да буду правоугаоне, кружне, троугласте, шестогаоне или да имају неки други (могуће и неправилан) облик
- Преброје се тачке које се налазе у некој ћелији C. Уколико је број тачака већи од унапред задате величине k, за ту ћелију се каже да је *густа*
- Густе ћелије које имају заједничке ивице или (у ређим случајевима) заједничку тачку (нпр. теме хиперкоцке) формирају кластер. Ретке ћелије не припадају ни једном кластеру
- Од густине линија које одређују мрежу хиперкоцки зависи да ли ће ћелија бити густа или ретка

Алгоритми засновани на мрежама и густини









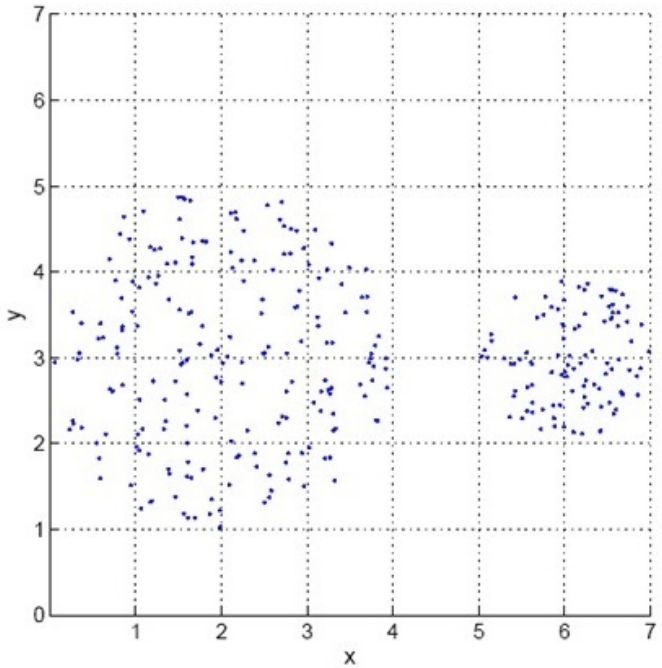
Алгоритми засновни на мрежама

Псеудокод алгоритма за кластеровање заснованог на мрежама је

```
/* Podatak: D, Gustina tacaka: t, Gustina mreze: p */
Klasterovanje_mreze(D, p, t)
begin
  Diskretizovati svaku dimenziju podatka D u p vrednosti;
  Odrediti gustinu celija mreze za gustinu tacaka t;
  Napraviti grafik u kome su guste celije
     povezane ako su susedne;
  Odrediti veze delova grafa
  return tacke u svakoj povezanoj komponenti kao klaster;
end
```

Алгоритам заснован на мрежама и густини

Пример: Еуклидска густина заснована на ћелијама - подела региона на неки број ћелија и дефинисање густине преко броја тачака у ћелијама

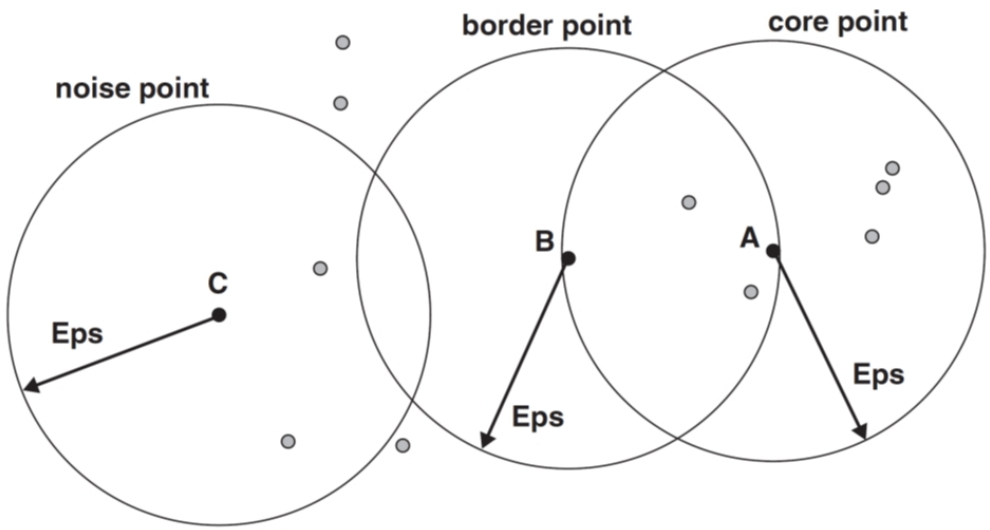


0	0	0	0	0	0	0
0	0	0	0	0	0	0
4	17	18	6	0	0	0
14	14	13	13	0	18	27
11	18	10	21	0	24	31
3	20	14	4	0	0	0
0	0	0	0	0	0	0

DBSCAN

За задату вредност ε и број t

- Тачка припада *језгру* ако се у кругу полупречника ε налази бар t других тачака
- Тачка је *на граници* ако се у кругу полупречника ε налази мање од t других тачака, али се налази бар једна тачка језгра
- Тачка је *шум* ако није нити у језгру нити на граници.



DBSCAN

DBSCAN (Density-Based Spatial Clustering of Applications with Noise) алгоритам за задату вредност ε и број t дели тачке на које се примењује у три категорије:

- Све тачке које се налазе на растојању мањем од ε, тј. припадају језгру неке тачке се смештају у исти кластер као и тачка језгра
- Тачке на граници се придружују истом кластеру као и тачка језгра на чијој се граници налазе. У случају да су на граници два кластера тада се доноси одлука коме кластеру припадају.
- Све тачке које су шум се одбацују

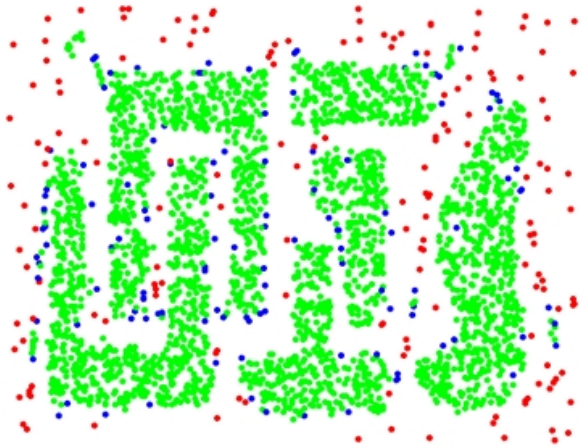
DBSCAN

Псеудокод DBSCAN алгоритма је

```
/* Podatak: D, Poluprecnik: eps, Gustina tacaka: t */
DBSCAN(D, p, t)
begin
  Odrediti jezgro, granicu i sum tacaka
     iz D za par (Eps, t);
  Formirati graf u kome su povezane tacke
     koje pripadaju jezgru ako su
     medjusobno unutar Eps;
  Odrediti povezane komponente grafa;
  Svaku tacku na granici dodeliti povezanoj
     komponenti sa kojom je najbolje povezana;
  return tacke svake povezane komponenti kao klaster;
end
```

DBSCAN

Original Points



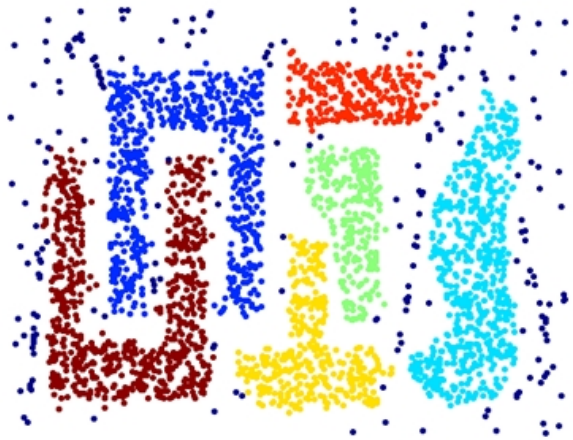
Point types: core, border and noise

Eps = 10, MinPts = 4

DBSCAN

Погодност DBSCAN алгоритма

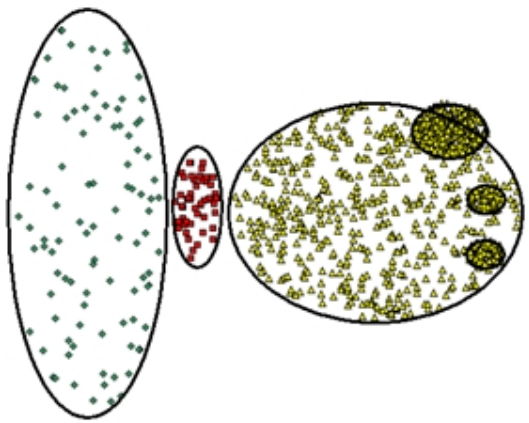
Original Points



Clusters

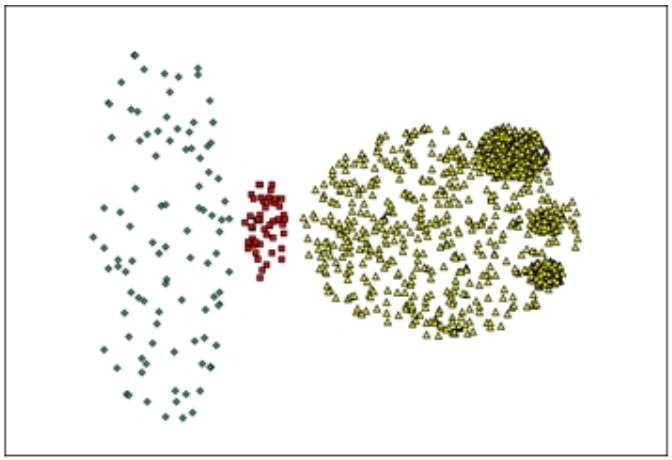
DBSCAN

Недостатак DBSCAN алгоритма

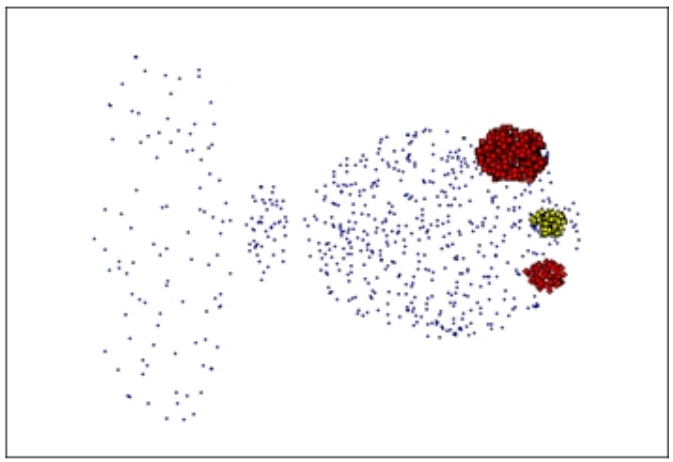


Original Points

- **Varying densities**
- **High-dimensional data**



(MinPts=4, Eps=9.75).



(MinPts=4, Eps=9.92)

Провера коректности кластеровања

Провера коректности кластеровања и исправности добијених кластера (модела) није једноставан задатак као у случају класификације где су унапред познати тачан број и ознаке класа. Приликом кластеровања увек треба имати на уму да

- Сваки алгоритам кластеровања ће пронаћи кластере у подацима без обзира да ли они заиста постоје или не. Због тога је пожељно извршити проверу да ли у материјалу уопште постоје кластери
- Не постоји најбољи алгоритам кластеровања за поједини скуп података. Савет је увек пробати више различитих алгоритама да би се дошло до коректног решења

Провера коректности кластеровања

У реалном окружењу, поготову када не постоји спољашњи критеријум за проверу, постављају се различита питања:

- Како одредити тенденцију кластеровања скупова података, тј. да ли структуре које нису случајне (и које могу да се кластерују) постоје у улазном скупу
- Да ли је добро одређен број кластера за одабрани алгоритам кластеровања који захтева уношење броја кластера унапред
- Одредити колико су резултати кластеровања добри у односу на улазне податке у случају да не постоје референце на спољашње информације
  - Унутрашњи (интерни) критеријум јако зависи од коришћеног алгоритма

Провера коректности кластеровања

Питања (наставак):

- Ако постоје спољашње информације о ознакама класе, како са њима упоредити резултате кластеровања
- У општем случају нема *спољашњег* критеријума који је на располагању за проверу. Делимично, могућа је провера преко спољашњег критеријума ако
  - постоје синтетички генерисани подаци за тестирање
  - постоје ознаке класа
- Како од два различита резултата кластеровања одредити који је бољи

Тенденција кластеровања

Постоје два начина како може да се одреди да ли су подаци погодни за кластеровање, тј. да ли у њима постоје кластери:

- Могу да се користе различити алгоритми за кластеровање и упореде добијени резултати. Ако се у свим случајевима добијају лоши кластери, то најчешће значи да у подацима не постоје кластери, односно да су подаци случајни
- Одредити да ли постоји кластер у подацима без претходног кластеровања. Технике које се најчешће примењују (поготову за податке у Еуклидском простору) укључују статистичке тестове на случајан распоред у простору. Иако су доста сложене, развијен је велики број техника које могу да се примене на податке у Еуклидском простору малих димензија.

Хопкинсова статистика

Једна од метода којом може да се одреди да ли постоје кластери у подацима без претходног кластеровања

- Нека је дат скуп од n тачака које треба кластеровати. Да би испитали да ли дати скуп садржи кластер, односно да ли је погодан за кластеровање, одабрати број $p < n$ и генерисати p тачака које су случајно распоређене у простору које покрива улазни скуп, и изабрати узорак од p тачака из оригиналног скупа од n тачака.

Хопкинсова статистика

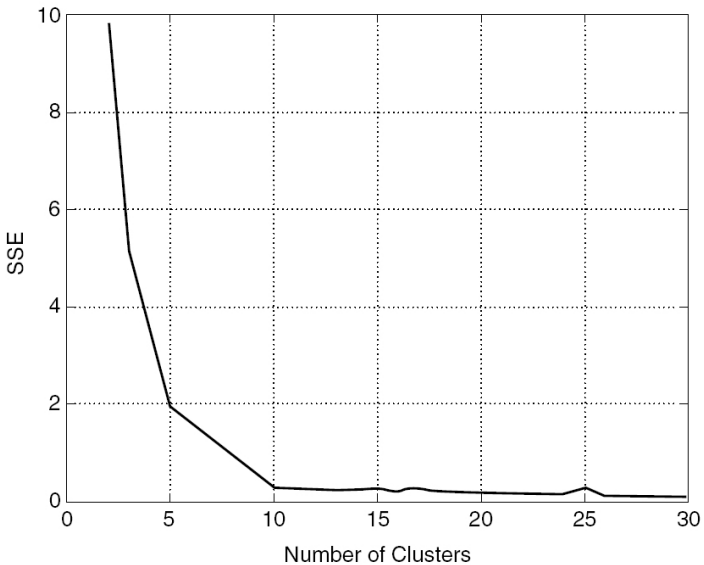
- За оба скупа од n тачака одредити растојање сваке тачке до најближег суседа у оригиналном скупу. Нека су v_i и o_i растојања до најближих суседа тачака из вештачки генерисаног скупа и скупа тачака који је узоркован из оригиналног скупа. Тада се Хопкинсова статистика дефинише као

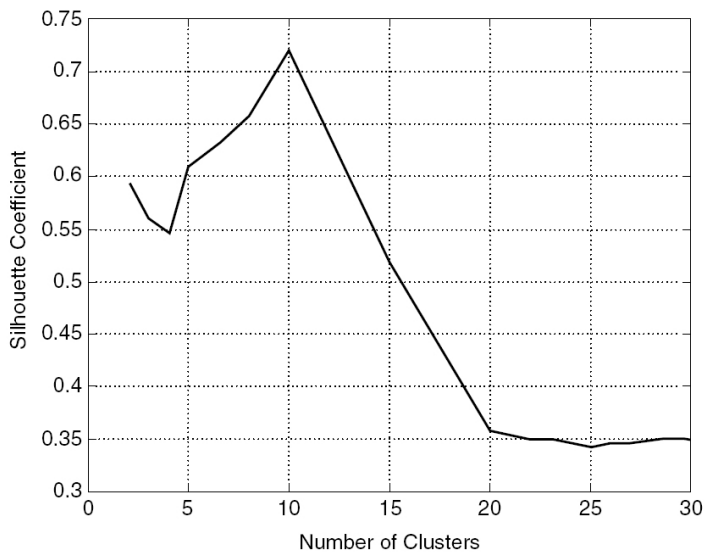
$$H = \frac{\sum_{i=1}^{p} o_i}{\sum_{i=1}^{p} v_i + \sum_{i=1}^{p} o_i}$$

- Ако случајно генерисане тачке и тачке из узорка имају слична растојања до најближих суседа, вредност ≈ 0.5. То значи да су подаци највероватније случајни и да не постоји задовољавајуће кластеровање. Вредност $\to 1$ означава да тачке са великом вероватноћом могу да се групишу у кластере.

Одређивање броја кластера

Неки алгоритми кластеровања сами одређују најбољи број кластера, док други захтевају да се број резултујућих кластера зада унапред. Постоји више метода за коректно одређивање броја кластера које се у принципу своде на узастопну примену истог алгоритма са различитим задатим вредностима кластера, графичким представљањем добијених резултата процене коректности и тражењем тачки максимума/минимума и прегиба на таквим графицима.





На претходној слици приказане су вредности CCE и силуете коефицијентра за различит број кластера. Са слике се види да је у овом случају најбољи број кластера 10.

Унутрашњи критеријуми провере

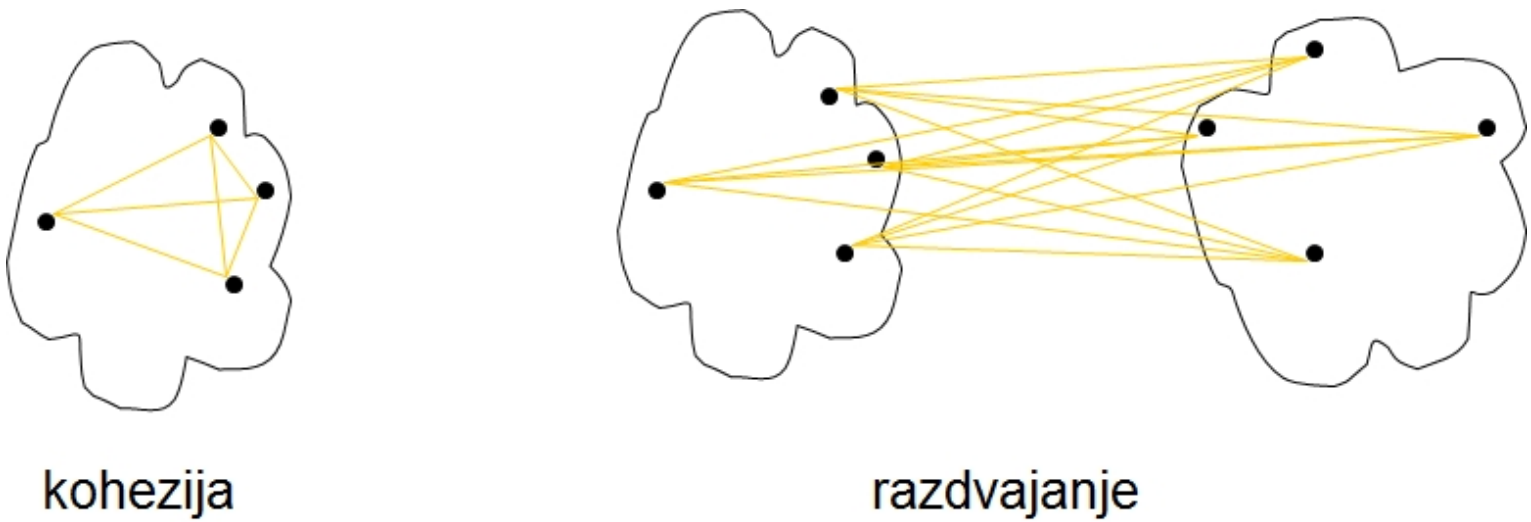
Најчешће коришћени критеријуми провере у овом случају су:

- Мере *кохезије* (компактности) кластера које показују колико су близу једно другом елементи истог кластера
- Мере *раздвајања* (изолације) кластера које показују колико су различити кластери међусобно раздвојени
- Однос растојања у кластеру/ван кластера
- Коефицијент сенке (енг. *silhouette*)
- Вероватносна мера
- Коефицијент кофенетичке корелације (енг. *Cophenetic Correlation Coefficient*, *CPCC*)

Кохезија и раздвајање

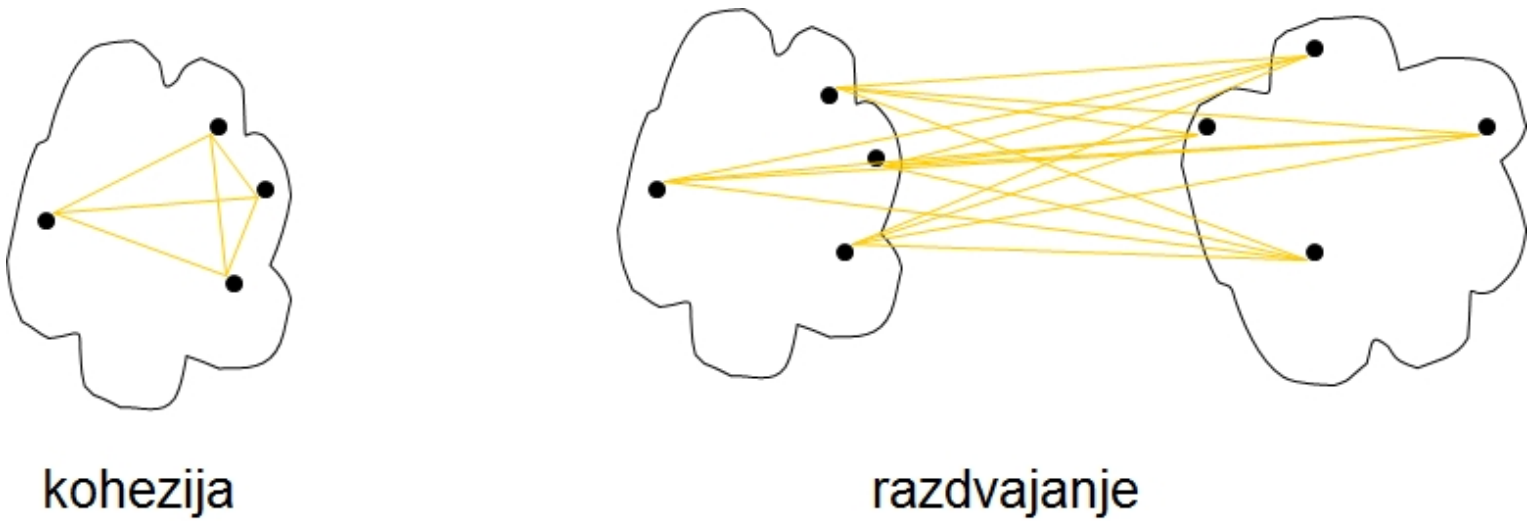
Као мера за одређивање кохезије кластера може се користити

- Збир квадрата међусобног растојања тачака у кластеру или збир квадрата растојања тачака од центроида тог кластера
- Боље прилагођена алгоритмима који су засновани на одређивању растојања (као k-средина)
- Мање одговарају алгоритмима са мрежама и густином
- Апсолутне вредности растојања не пружају квалитетну информацију о квалитету самог кластера



Кохезија и раздвајање

Као мера раздвајања кластера - збир растојања између елемената у кластеру и елемената ван кластера, или растојање центроида кластера



Неки алати омогућавају коришћење више мера. На пример, CLUTO алат користи, између осталог и кохезију и раздвајање

Однос растојања у кластеру/ван кластера

Однос растојања у кластеру/ван кластера је прецизнија мера од квадарата растојања тачака. Рачуна се на следећи начин:

- Из улазног скупа се узме p парова тачака. Нека P означава скуп парова који припадају истом кластеру нађеном од стране алгоритма, а Q скуп парова тачака које припадају преосталим кластерима
- Просечно растојање унутар и изван кластера је дефинисано са

$$\text{У кластеру} = \frac{\sum_{(X_i, X_j) \in P} dist(X_i, X_j)}{\mid P \mid}$$

$$\text{Ван кластера} = \frac{\sum_{(X_i, X_j) \in Q} dist(X_i, X_j)}{\mid Q \mid}$$

$$\text{Растојање унутар/ван кластера} = \frac{\text{У кластеру}}{\text{Ван кластера}}$$

- што је вредност растојања мања кластеровање је боље и обратно

Сенка коефицијент

Нека важе следеће ознаке:

- $AvgDist_i^{in}$ нека означава просечно растојање X_i до тачака унутар кластера коме припада X_i;
- $AvgDist_i^{out}$ нека означава просечно растојање X_i до тачака кластера коме не припада X_i;
- Нека је $MinDist_i^{out} = min\{AvgDist_i^{out}\}$
- Тада се коефицијент сенке S_i у односу на i-ти објекат дефинише као

$$S_i = \frac{MinDist_i^{out} - AvgDist_i^{in}}{max\{MinDist_i^{out}, AvgDist_i^{in}\}}$$

- Ако је вредност коефицијента сенке блиска 1 кластери су добро раздвојени. Негативне вредности означавају мешавину података у кластерима и лоше кластеровање
- Добра особина коефицијента сенке је што апсолутна вредност носи информацију о квалитету кластеровања

Вероватносна мера

Вероватносна мера

- Модел са помешаним подацима за процену квалитета појединачног кластеровања
- Претпоставка: центроид мешаних података је центроид нађених кластера
- Остали параметри се рачунају користећи методу сличну EM (енг. *expectation-maximization*) алгоритму
- Корисно када се зна да кластери требају да имају специфичан облик

Коефицијент кофенетичке корелације

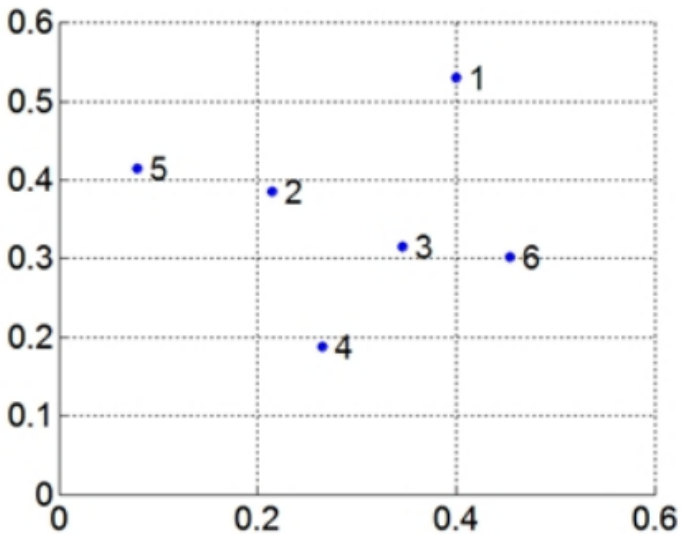
Мера квалитета за хијерархијско кластеровање

- Кофенетичко растојање два објекта је сличност на основу које технике хијерархијског сакупљајућег кластеровања први пут смештају објекте у исти кластер
- Нпр. Ако је најмање растојање између два кластера 0.2, тада све тачке у једном кластеру имају растојање 0.2 у односу на тачке другог кластера
- Коефицијент кофенетичке корелације је корелација између елемената кофенетичке и почетне матрице сличности

Коефицијент кофенетичке корелације

За процену квалитета сакупљајућег хијерархијског кластеровања за изабрани начин рачунања растојања

- Одреде се почетне матрице сличности за сваки одабрани начин рачунања растојања
- Одреде се одговарајуће кофеентичке матрице сличности
- Израчуна се коефицијент корелације за сваки пар
- Најбољи начин рачунања има највећу корелацију са кофенетичком матрицом сличности



Коефицијент кофенетичке корелације

Матрица растојања за најбољу (појединачну) везу

	p1	p2	p3	p4	p5	p6
p1	0.0	0.24	0.222	0.37	0.34	0.23
p2	0.24	0.0	0.148	0.20	0.139	0.25
p3	0.222	0.148	0.0	0.151	0.28	0.110
p4	0.37	0.20	0.151	0.0	0.29	0.22
p5	0.34	0.139	0.28	0.29	0.0	0.39
p6	0.23	0.25	0.110	0.22	0.39	0.0

Одговарајућа матрица кофенетичких растојања

p1	0.0	0.222	0.222	0.222	0.222	0.222
p2	0.222	0.0	0.148	0.151	0.139	0.148
p3	0.222	0.148	0.0	0.151	0.148	0.110
p4	0.222	0.151	0.151	0.0	0.151	0.151
p5	0.222	0.139	0.148	0.151	0.0	0.148
p6	0.222	0.148	0.110	0.151	0.148	0.0

Спољашњи критеријуми провере

Ова врста критеријума се користи када постоје тачне информације о кластерима у подацима који треба да се кластерују. У општем случају, то није могуће за највећи број реалних скупова података. Изузетак је случај када се синтетички подаци генеришу на основу познатих скупова за проверу у ком случају је могуће придружити идентификацију кластера сваком појединачном податку улазног скупа.

У реалним ситуацијама овакав случај ће бити само приближно могућ када се користе расположиве ознаке класа. Највећа опасност која се јавља код коришћења ознака класа је да су често те ознаке засноване на специфичним особинама скупа података који је зависан од неке апликације којој представља улаз, што може да доведе до тога да не буду коректно погођени природни кластери у улазним подацима.

Спољашњи критеријуми провере

Без обзира на недостатке, ове методе имају предност у односу на интерне критеријуме провере јер могу да избегну конзистентна одступања у извршавању при примени на више различитих скупова улазних података.

За проверу се најчешће користе

- Матрица конфузије
- Различите мере - најчешће Гинијев индекс, ентропија, прецизност, одзив, Ф-мера, ...

Критеријуми провере - домаћи задатак

Домаћи задатак: коришћењем програма СПСС Моделер добити резултате кластеровања применом алгоритама

- K-средина
- TwoStep

над улазним скуповима

- IRIS
- ADULT

Кластеровање извршити за различите вредности броја кластера (у K-средина), као и интервала за број кластера (у TwoStep).

Коментарисати добијене вредности коефицијента сенке и резултате кластеровања.

Кластеровање категоричких података

- Један начин конверзија у бинарне податке
- Одређивање центроида за категоричке податке
  - хистограм вероватноћа за сваки атрибут
  - центроид - категоричка вредност која се јавља у највећем проценту
- Рачунање сличности категоричких података
- Различити алгоритми. Нпр. ROCK (*RObust Clustering using linKs*) је заснован на сакупљајућем приступу где се кластери комбинују користећи критеријум сличности.

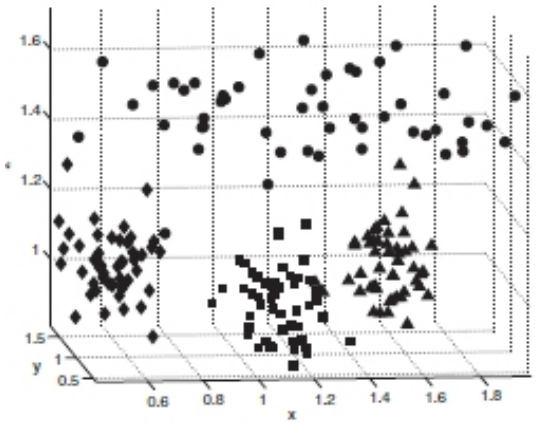
Кластеровање категоричких података

K-модално кластеровање

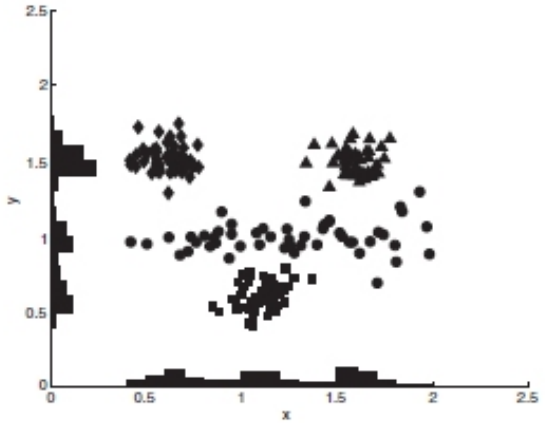
- За сваки од атрибута се одређује модална вредност (вредност са највећом фреквенцијом)
- Модална вредност сваког од атрибута се одређује независно у односу на вредности других атрибута због чега одабрана репрезентативна вредност ('центроид') не мора да припада скупу података
- к-модално кластеровање може ефективно да се користи ако су вредности категоричких атрибута равномерно распоређене
- Ако вредности категоричких атрибута нису равномерно распоређене врши се нормализација деобом фреквенције у кластеру са фреквенцијом у комплетном скупу података

K-медоид кластеровање - репрезентативна тачка је из материјала

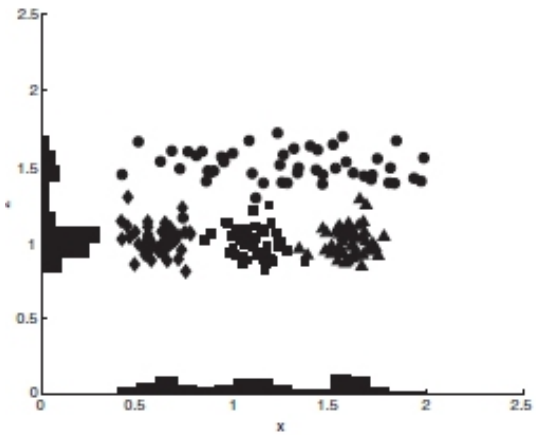
Кластеровање по потпросторима



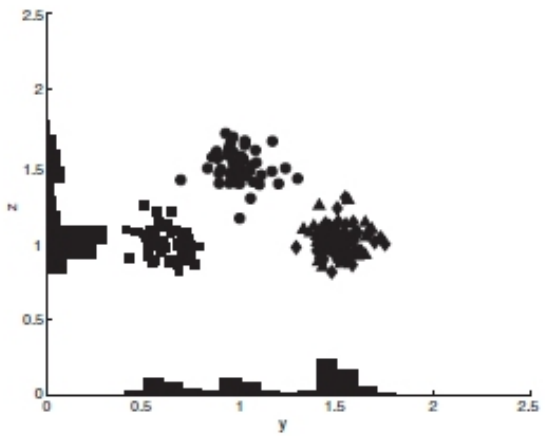
(a) Four clusters in three dimensions.



(b) View in the xy plane.



(c) View in the xz plane.



(d) View in the yz plane.

Кластеровање по потпросторима

Кластеровање по потпросторима

- Подаци могу да се кластерују према подскупу атрибута
- Различити кластери се јављају у зависности од различитих димензија подскупова
- Алгоритми засновани на различитим методама
  - CLIQUE (*CLustering In QUEst*) - концептуално заснован на сличном принципу као и *Apriori*
  - DENCLUE (*DENsity CLUstEring*) - моделира укупну густину скупа тачака као збир утицаја функције придружених свакој тачки

CLIQUE

- Комбинује кластеровање засновано на густини и мрежама за проналажење потпростора погодних за кластеровање.
- Проналазе се ћелије са густином већом од прага. Густе ћелије се спајају ако се додирују ивицом.
- Тако спојене ћелије формирају кластере
- Одређује било који број кластера у произвољном броју димензија.
- Одређује кластере различитих облика и величина

MAFIA

- MAFIA (Merging of Adaptive Finite IntervAls)
- Варијанта CLIQUE - конструише ћелије адаптивне величине у свакој димензији
- За кластеровање бира само потпросторе обухваћене атрибутима чија је ентропија мања од прага
- Може да одреди кластере са јако малом густином
- Ограничење - висока цена израчунавања

ENCLUS

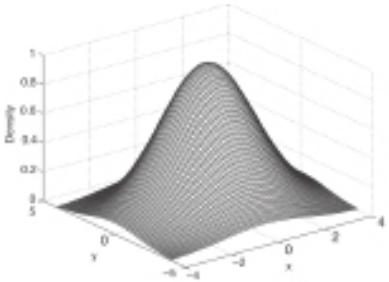
- ENCLUS (ENtropy-based CLUStering)
- Варијанта CLIQUE која користи критеријум за избор потпростора заснован на ентропији - конструише ћелије адаптивне величине у свакој димензији
- За одређивање броја ћелија у свакој димензији користи хистогртаме
- По одређивању ћелија наставља као и CLIQUE користећи Априори принцип
- Омогућава парлелизацију и може да обради велику количину високо скалабилних података

DENCLUE

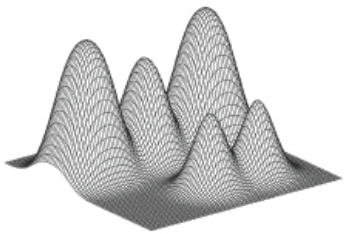
- Приступ заснован на концепту функције утицаја која математички моделира утицај тачака података на своје суседе
- Густина у некој тачки се процењује као збир утицаја свих тачака података
- Функција густине у тачки x се одређује као збир функција утицаја свих тачака на тачку x
- Тачка привлачења је тачка која одговара локалном максимуму функције густине
- Кластер се дефинише као скуп свих тачака повезаних (преко тачака високе густине) са неком тачком привлачења x, при чему је вредност функције густине у x већа од задатог прага.

DENCLUE - илустрација

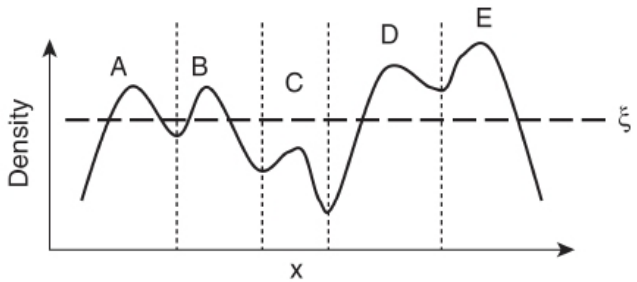
$$K(y) = e^{-distance(\mathbf{x},\mathbf{y})^2/2\sigma^2}$$



Set of 12 points.



Overall density—surface plot.



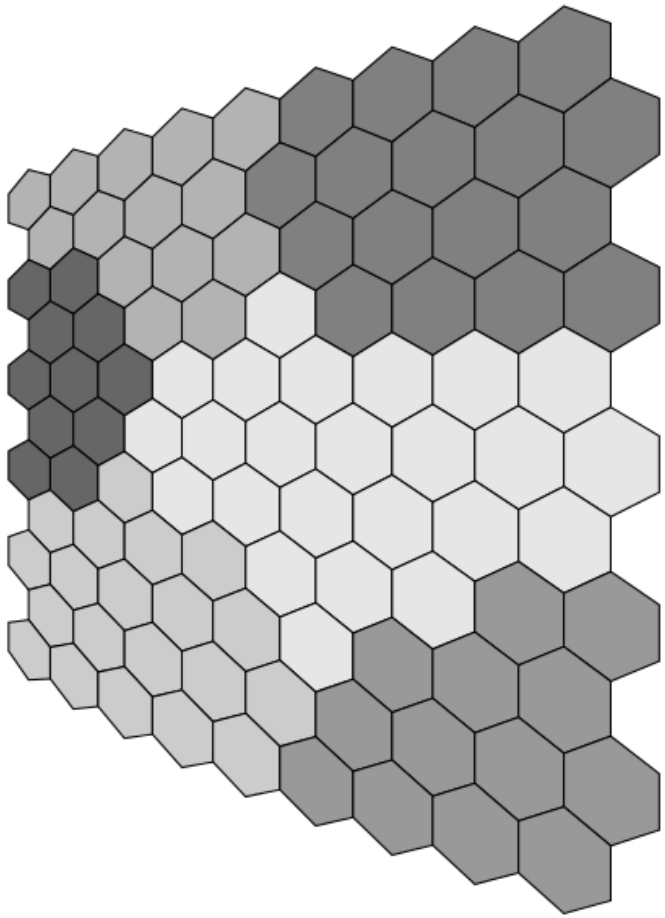
Кластеровање скалабилних података

- Комплетни подаци не могу да се сместе у меморију
- Алгоритми засновани на различитим методама
  - CLARA (*Clustering Large Applications*) и CLARANS (*Clustering Large Applications on RANdomized Search*) су засновани на уопштењу приступа кластеровању помоћу к-медоида
  - CURE (*Clustering Using REpresentatives*) је сакупљајући алгоритам за хијерархијско кластеровање
  - BIRCH (*Balanced Iterative Reducing and Clustering using Hiearchies*) представља уопштење алгоритма к-средина на хијерархијску методу одозго-наниже

SOM

- Погодне за кластеровање (и за класификацију)
- Сличне KHH
- Укључују топографску организацију центроида (неурона)
- Сваки центроид је одређен паром координата
- При раду ажурирају се текући центроид и центроиди који су му у близини по топографској оријентацији

SOM



Основни SOM алгоритам

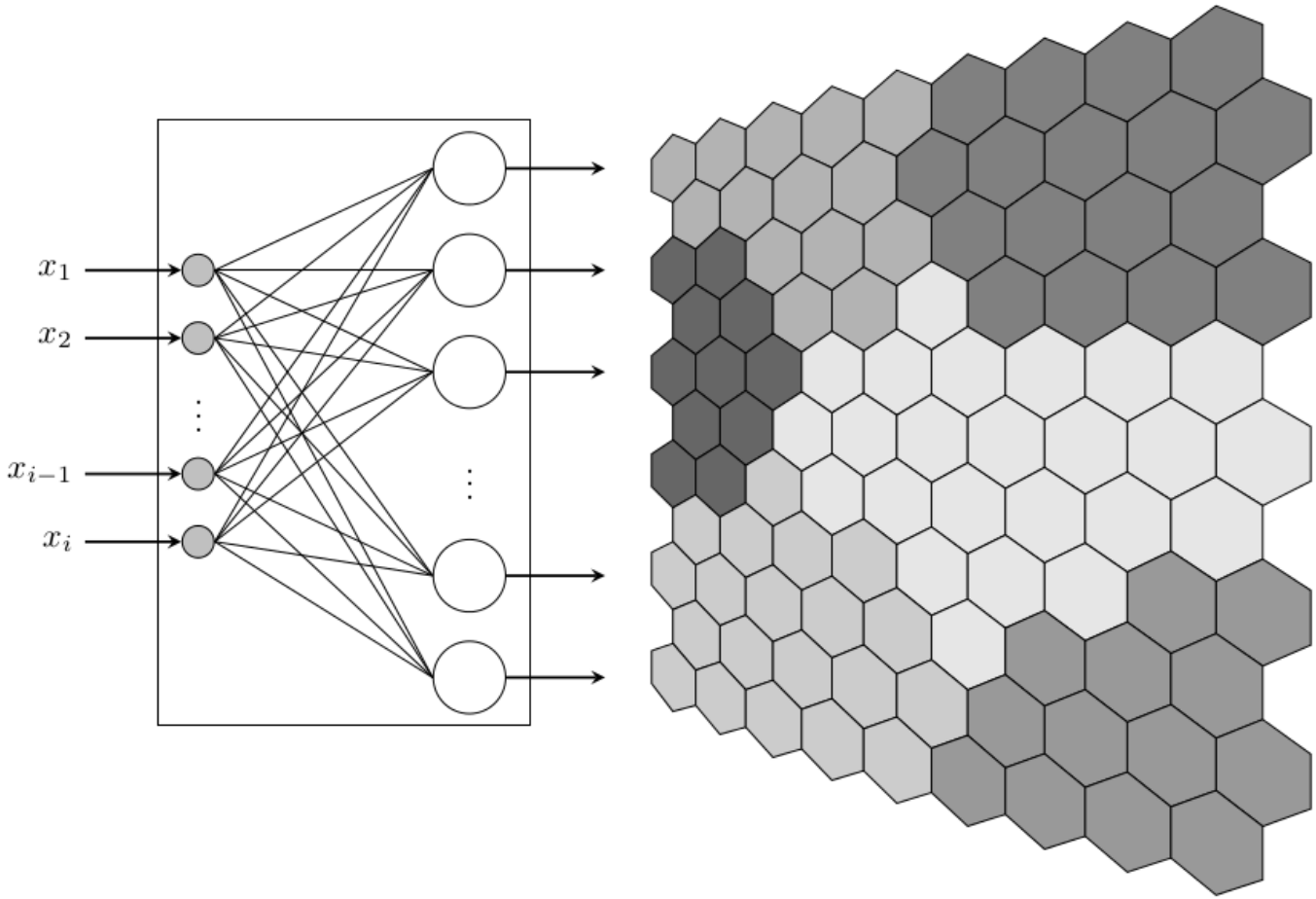
Основни SOM алгоритам кластеровања

```
Inicijalizovati centroide

repeat

    Izabrati sledeci objekat
    Odrediti najblizi centroid izabranom objektu
    Azurirati centroid i susedne centroide
        (centroide koji su u blizini)

until Centroidi se ne menjaju /dostignut je prag
Dodeliti svaki objekat najblizem centroidu
    i vratiti centroide i klastere
```

Основни SOM алгоритам - кораци

- Инцијализација
  - случајан избор центроида у интервалу посматраних вредности
  - случајан избор тачака за центроиде
- Избор објекта
  - ако је број објеката јако велики, не користе се сви
- Одређивање најближег центроида
  - метрике растојања (еуклидско/косинусно растојање)
- Ажурирање центроида
- Терминирање

SOM



Ажурирање центроида

- Нека су $m_1, ..., m_k$ центроиди
- Нека је $p(t)$ текући објекат у тренутку t и нека је њему најближи центроид m_j
- У тренутку $t+1$ j-ти центроид се ажурира

$$m_j(t+1) = m_j(t) + h_j(t)(p(t) - m_j(t))$$

- $h(t)$ одређује ефекат разлике и обично се бира

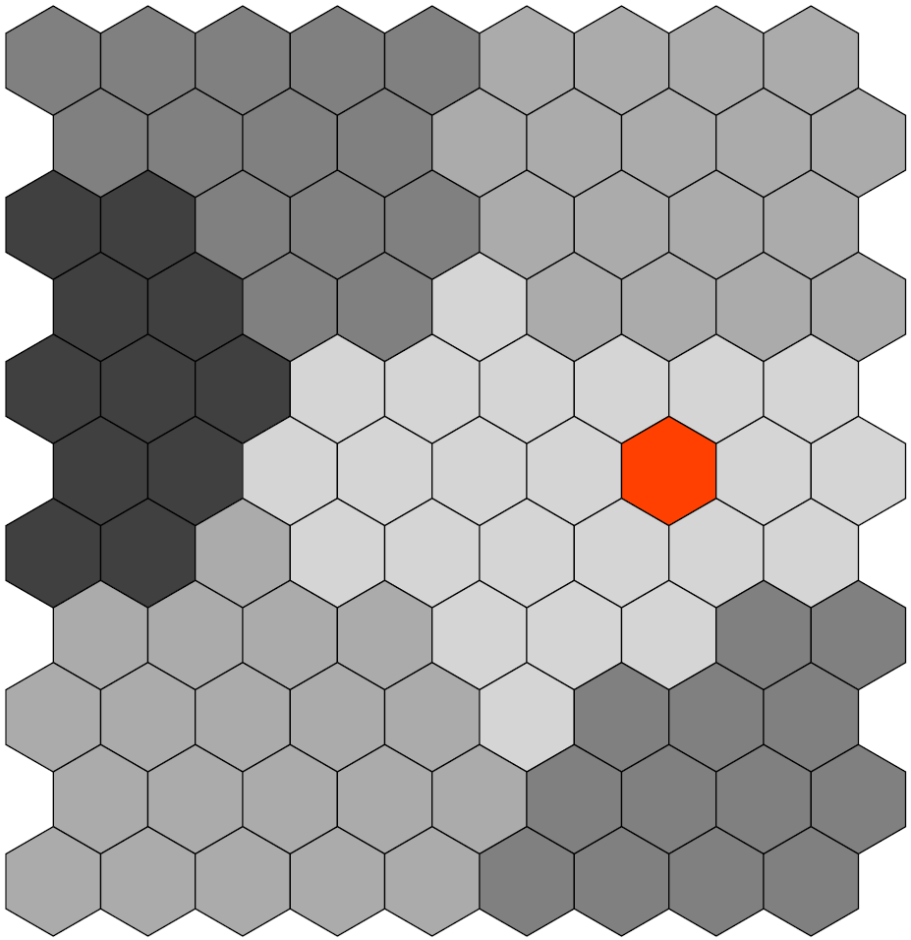
$$h_j(t) = \alpha(t) e^{\frac{-dist(r_j, r_k)^2}{2\sigma^2(t)}} \quad (Gausova \quad funkcija)$$

или

$$h_j(t) = \begin{cases} \alpha(t) & \text{ako} \quad dist(r_j, r_k) \leq \text{prag} \\ 0 & \text{inace} \end{cases}$$

где је $0 < \alpha(t) < 1$, $r_k = (x_k, y_k)$су координате центроида, а $dist(r_j, r_k)$ је Еуклидско растојање између два центроида

SOM



Предности и ограничења SOM

- Предности
  - Суседни кластери су више у релацији од несуседних
  - Погодно за визуелизацију
  - SOM одређује структуру
- Недостаци
  - Потребан одабир параметара, функције за рачунање суседства и избор центроида
  - SOM кластер не одговара природном кластеру (може да садржи више природних кластера али и један природни кластер може да се разбије на више SOM кластера)
  - Недостаје специфична функција објекта којом може да се изрази поступак
  - Нема гаранције за конвергенцију, мада у пракси често конвергира