Кластеровање вискодимензионих података

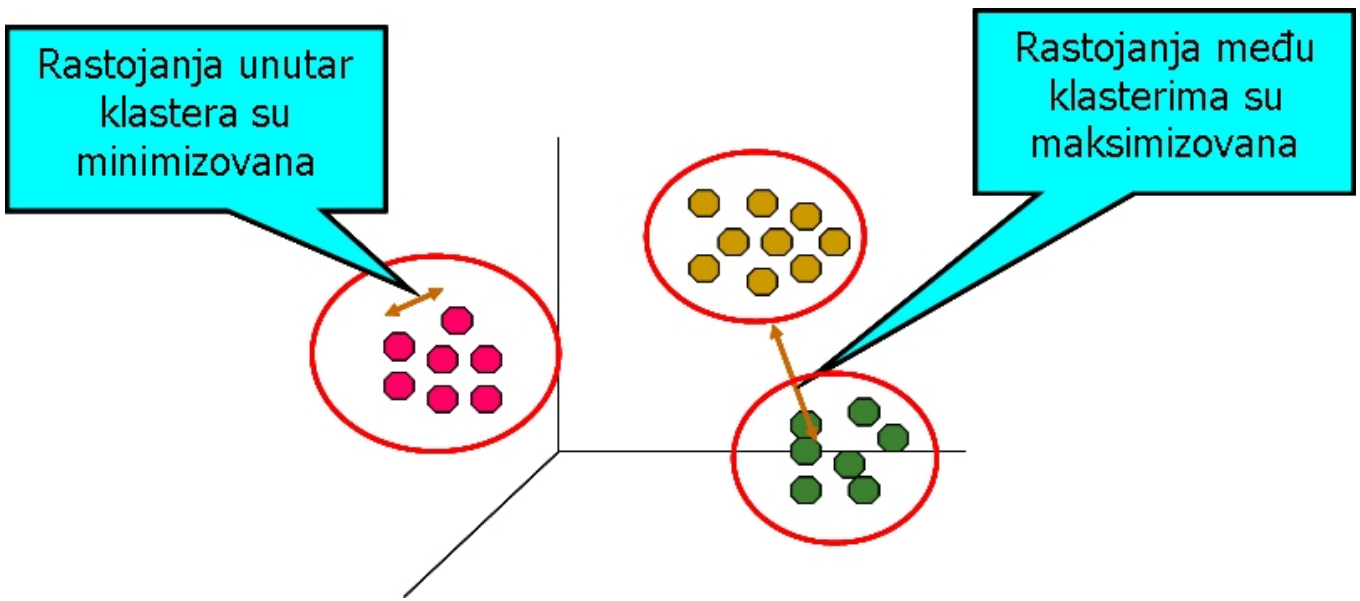
Ненад Митић

Математички факултет
nenad@matf.bg.ac.rs

Увод

Кластеровање:

Извршити поделу датог скупа објеката $X = \{x_1, x_2, ..., x_n\}$ на групе (подскупове) тако да је објекат x_i који припада групи G_l сличнији по неком критеријуму објектима x_j који припадају (истој) групи G_l него неком објекту x_k који припада некој другој групи G_m. Свака од група G се назива кластер, а целокупан поступак поделе улазног скупа кластеровање.



Uvod

- Објекти су обично представљени као тачка у вишедимензионалном простору
- Скупови података који се обрађују постају све сложенији са великим бројем објеката који имају велики број димензија
- Јавља се проблем скалабилности постојећих алгоритама - потребна је или њихова модификација или развој нових алгоритама
- Проблем квалитета и брзине кластеровања
- Постојеће (класичне) мере сличности постају неупотребљиве. Велики број димензија може да доведе до тога да су сви објекти међусобно једнако удаљени

Проблеми при кластеровању високодимензионих података

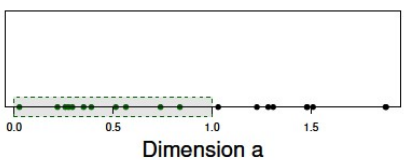
При кластеровању високодимензионих података могу да се јаве следећи проблеми:

- Проклетство димензионалности
- Ефекат концентрације норми
- Проблем релевантности (локалних) особина/атрибута

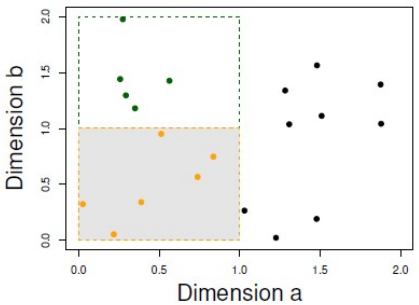
Проклетство димензионалности

- Термин је првобитно коришћен 1961. године да би описала немогућност оптимизације функција са више променљивих методом грубе силе на дискретну вишедимензиону решетку
- У каснијем периоду је почео да се користи за означавање било ког проблема у анализи података у коме се користи велики број атрибута (променљивих)
- Већи број атрибута - више информација али и више проблема у обради
- Проблем је и што објекати садрже велики број 'празних' атрибута
- Ефекат повећања димензионалности на мере сличности/различитости

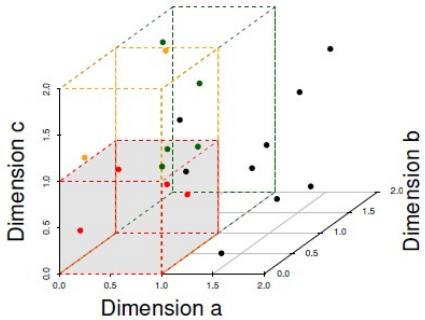
Проклетство димензионалности



(a) 11 Objects in One Unit Bin



(b) 6 Objects in One Unit Bin



(c) 4 Objects in One Unit Bin

Подаци у једној димензији су релативно густо спаковани. Повећање броја димензија развлачи и удаљава тачке, доводећи до тога да су тачке у таквом простору релативно ретке

Ефекат концентрације норми

- На пример, ако се користи Еуклидско растојање као мера, показује се да таква мера има друге карактеристике у односу на њену примену на објекте у просторима мале димензије. Под одређеним предусловима релативна растојања од било које тачке до њеног најближег и најдаљег суседа имају тенденцију да буду скоро идентична.
- Када релативне варијансе растојања (релативне у односу на средњу вредност растојања) тачака конвергирају ка нули за веће димензије, тада заједно са повећањем броја димензија релативна разлика растојања до најближе и најдаље тачке такође тежи ка нули
- Феномен хаба: за високодимензионе податке важи да се релативно мали број тачака (хабови) врло често налази између к-најближих суседа других тачака.

Проблем релевантности атрибута

- Чест случај је да се у великом броју димензија јављају атрибути чија вредност има само локално значење за један (мањи) број објеката
- Такви атрибути су некорисни и могу да сметају нпр. у случају димензионе редукције
- Могуће је да су у неким случајевима показатељи неких скривених кластера

Димензиона редукција

- Технике димензионе редукције нису ефикасне у случају већег броја ирелевантних атрибута
- Проблем настаје код интерпретације оригиналаних домена после редукције и када су објекти повезани на различите начине у различитим подскуповима димензија
- Избор карактеристика (атрибута) је проблематичан када су кластери у појединим потпросторима

Алгоритми кластеровања високодимензионих података

Алгоритми се могу груписати према методама које су коришћене, сложености, просторним захтевима, скалабилности

- Засновани на хијерархији (BIRCH, CURE, CHAMELEON, ROCK, TwoStep)
- Засновани на подели (K-sredina, K-modalni, CLARA, CLARANS, FCM)
- Засновани на густини (DBSCAN, DBCLASD, DENCLUE, SNN, Hibridni(BRIDGE, CUBN))
- Засновани на мрежи GRIDCLUS, BANG, WaveCLuster, STING, CLIQUE, ENCLUS, MAFIA, OPTIGRID (O-Cluster, CBF), PROCLUS, ORCLUS, GDILC
- Засновани на моделу GAUSSIAN (EM), COOLCAT, STUCCO, SOM

BIRCH

- Birch: Balanced Iterative Reducing and Clustering using Hierarchies (Zhang, Ramakrishnan & Livny, SIGMOD'96)
- Инкрементално конструише CF (Clustering Feature) дрво
  - Фаза 1: скенира базу за формирање иницијалног CF дрвета у меморији. CF дрво се користи за компресију података на више нивоа ради очувања наслеђене кластер-структуре података
  - Фаза 2: користи произвољни алгоритам кластеровања листова у CF дрвету
- Линеарно скалирање - у једном пролазу се налази добро кластеровање које се побољшава новим итеративним кластеровањима
- Недостатак - ради само са нумеричким подацима, и осетљив је на редослед слогова у улазним подацима

TwoStep

- Заснован на CF дрвету
- Ради са нумеричким и категоричким подацима
- У Фази 1 се формира CF дрво
- У Фази 2 користи сакупљајуће хијерархијско кластеровање
- Детаљи: SPSS Algorithmic Guide

CURE

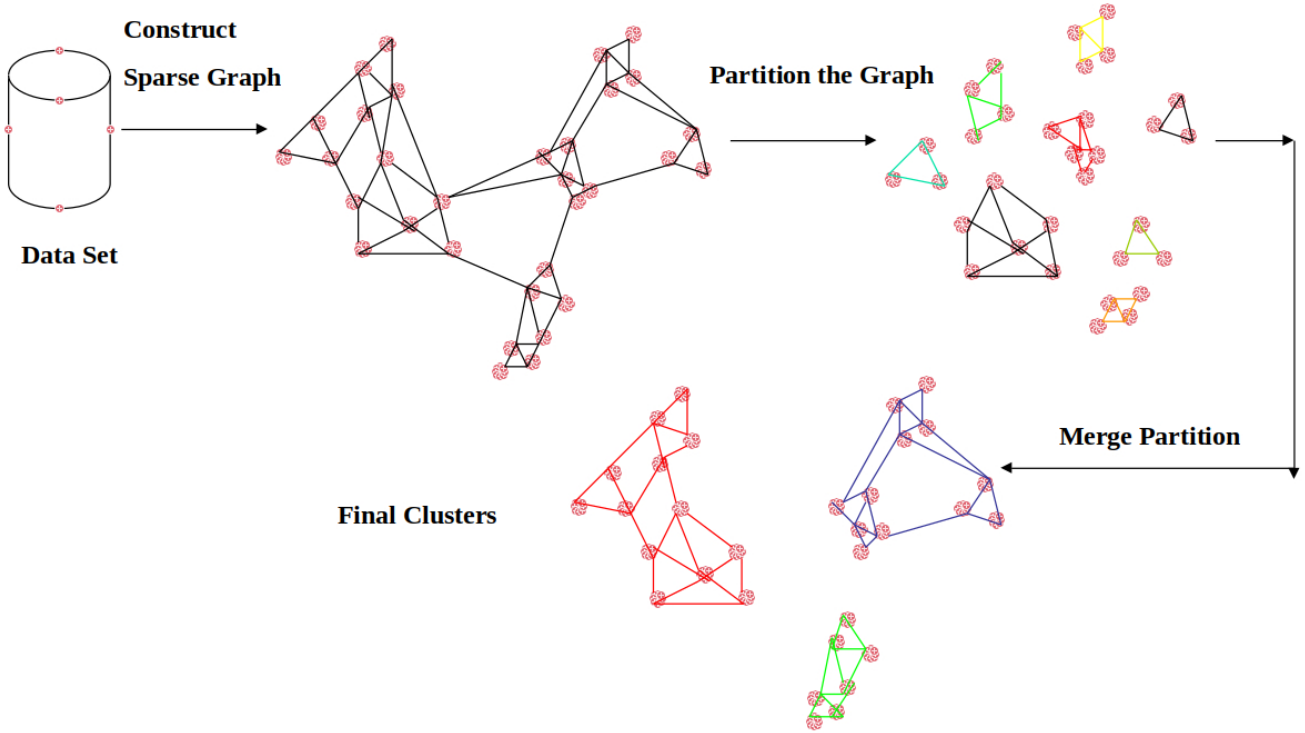
CURE – Clustering Using REpresentatives

- Сваки кластер се представља одређеним бројем фиксних тачака које су генерисане издвајањем добро-расутих тачака из кластера скупљајући их према центру кластера за одређену величину скалирања.
- Постојање више од једне репрезентативне тачке омогућава CURE алгоритму добро прилагођавање скуповима објеката несферичног облика, а скупљање помаже да се пригуше ефекти постојања елемената ван граница
- CURE користи комбинацију случајног узорковања и партиционисања за проширење на велике базе података
- Извучен случајни узорак из скупа података се партиционише у првом кораку, при чему је свака партиција је делимично кластерована. У наредном кораку се делимични кластери групишу у финалне кластере

CHAMELEON

- За дату матрицу сличности у првој фази се конструише репрезентација података у облику великог броја релативно малих ретких графова употребом методе к-најближих суседа
- У другој фази се користи сакупљајуће хијерархијско кластеровање за спајање два најближа поткластера узимајући у обзиро њихову повезаност и близину
- Разлика у односу на CURE: CURE игнорише информације о међусобној повезаности објеката
- Разлика у односу на ROCK: ROCK игнорише информације о блискости два кластера

CHAMELEON


Construct
Sparse Graph
Data Set
Partition the Graph
Merge Partition
Final Clusters

ROCK

ROCK –Robust Clustering using linKs

- Сакупљајући хијерархијски алгоритам који користи повезаност за спајање кластера
- Користи меру засновану на повезаности између две тачке и између два кластера
- Не користи растојање (тачака, кластера) као меру
- Традиционалне мере не дају увек добар резултат

ROCK

На пример, употреба Жакардове мере на скуповима:

- C1. <a, b, c, d, e>: {a, b, c}, {a, b, d}, {a, b, e}, {a, c, d}, {a, c, e}, {a, d, e}, {b, c, d}, {b, c, e}, {b, d, e}, {c, d, e}
- C2. <a, b, f, g>: {a, b, f}, {a, b, g}, {a, f, g}, {b, f, g}

даје погрешан резултат

- C1: 0.2 ({a, b, c}, {b, d, e}) до 0.5 ({a, b, c}, {a, b, d})
- C1 & C2: може да буде највише 0.5 ({a, b, c}, {a, b, f})
- Нпр. за $S_1 = \{a, b, c\}$, $S_2 = \{c, d, e\}$ сличност је 0.2

ROCK

Ако се као јачина везе рачуна број заједничких суседа

- $S_1 = \{a, b, c\}$, $S_2 = \{c, d, e\}$, $S_3 = \{a, b, f\}$
- веза(S_1,S_2) = 4 јер имају 4 заједничка суседа {a, c, d}, {a, c, e}, {b, c, d}, {b, c, e}
- веза(S_1,S_3) = 3 јер имају 3 заједничка суседа {a, b, d}, {a, b, e}, {a, b, g}

даје бољу основу за спајање кластера у односу на Жакардов коефицијент

K-средина

- Бира се K репрезентативних тачака за почетне центроиде
- Свака тачка се придружује најближем (у односу на коришћену меру) центроиду
- Када се кластер формира, ажурирају се вредности центроида
- Итеративно се понављају претходна два корака све док има промена или се задовољи критеријум конвергенције (у пракси када мање од 1% тачака мења кластер)
- Детаљно је обрађиван на ИП1

Варијанте K-средина

Варијанте K-средина се разликују по

- начину избора различитих иницијалних репрезентативних тачака (K-медоид, K-медијана, K-модална вредност) за Бира се K репрезентативних тачака за почетне центроиде
- избору боље процене инцијалних центроида (интелигентне K-средине, Генетске K-средине)
- примени неког облика трансформације (тежинске K-средине, кернел K-средине)

K-медоид

Отпорнији на елементе ван граница и шум у односу на K-средина
Укупна грешка - збир свих различитости између сваког објекта п и његовог репрезетнативног објекта

```
/* K-medoid algoritam */
Bira se $K$ reprezentativnih tacaka za pocetne centroide
repeat
    Dodeli svaku tacku klasteru sa najblizim reprezentivnim objektom
    Izabrati slucajnu nereprezentativnu tacku x
    Odrediti ukupnu gresku S pri zameni reprezentaivnog objekta m sa x
    Ako je S<0 zameniti m sa x i formirati novi skup reprezentativnih objekata
until dok se ne dostigne kriterijum konvergencije
```

K-медијана

- Бира се K репрезентативних тачака за почетне центроиде
- Свака тачка се придружује најближем (у односу на коришћену меру) центроиду
- Када се кластер формира, ажурирају се вредности центроида.
- Као центар се бира медијана - циљ је минимизација збира растојања између сваке тачке и њој најближег центра. Као мера се користи L_1 норма
- Итеративно се понављају претходна два корака све док има промена или се задовољи критеријум конвергенције

K-модално кластеровање

- Користи се за кластеровање података са категоричким атрибутима
- Као мера користи се L_0 - пребројавање броја истих/различитих тачака

Интелигентне K-средине

Заснован је на принципу да је интересантност тачака сразмерна удаљености од центроида

```
/* Intel. K-sredina algoritam */
Izracunati centar gravitacije c_g tacaka datog skupa
repeat
    Odrediti centroid c najdalje od c_g
    Formirati klaster S_iter od tacaka koje su blize c nego c_g
    Azurirati centroid s_g klastera S_iter
    Dodeliti c_g=s_g
    Odbaciti male klastere (ako postoje) koristeci zadati prag
until dok se ne dostigne kriterijum konvergencije
```

Генетско K-средине кластеровање

- За избегавање конвергенције ка локалном минимуму корсти се генетски алгоритам
- Развијен је Генетски K-средина алгоритам

Кернел K-средине кластеровање

- Подаци се иницијално пресликавају у вишедимензиони кернел простор користећи кернел функцију
- Спектрално кластеровање се може посматрати као варијанта кернел K-средина кластеровања

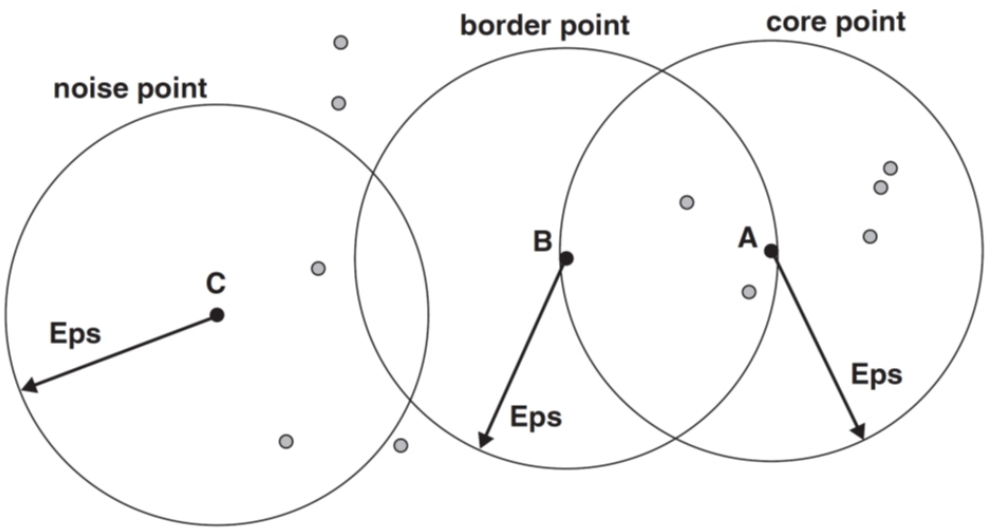
Тежинско K-средине кластеровање

- Класичан алгоритам K-средина се модификује увођењем тежина сваког астрибута
- Као мера за растојање користи се модификована варијанта Еуклидске мере са тежинама

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
- Детаљно је обрађиван на ИП1

DBCLASD

DBCLASD (Distribution Based Clustering of Large Spatial Databases)

- Инкрементални алгоритам - додељивање тачака кластеру је засновано само на до сада обрађеним тачкама без разматрања комплетног кластера или комплетне базе
- Почетни кластер се постепено повећава његовим суседним тачкама све док је растојање од кластера до тачке која је најближи сусед у оквиру задате расподеле растојања

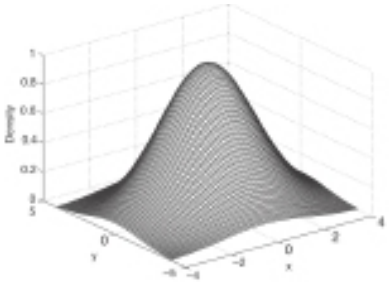
DENCLUE

DENCLUE (DENsiti CLUstEring) је приступ кластеровању заснован на густини, моделовањем укупне густине скупа тачака као збира функција повезаних са сваком тачком

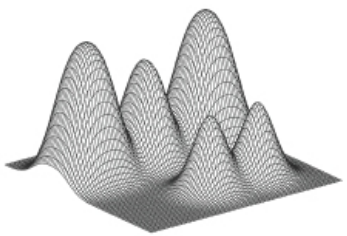
- Резултујућа функција укупне густине имаће локалне максимуме густине који се користе за директно дефинисање кластера
- За сваку тачку се налази најближи максимум функције, а скуп свих тачака повезаних са отако дређеним максимумом постаје кластер
- Ако је густина на локалном максимуму мања од прага минималне густине, тачка је шум и одбацује се
- Ако се локални максимум може повезати са другим локалним максимумом преко тачака података, а густина у свакој тачки путање је изнад прага минималне густине, тада се кластери који су додељени овим локалним максимумима спајају

DENCLUE - илустрација

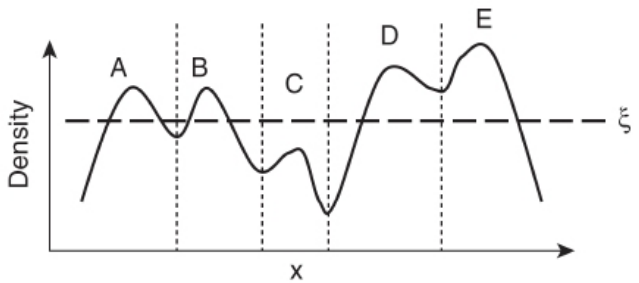
$$K(y) = e^{-distance(\mathbf{x},\mathbf{y})^2/2\sigma^2}$$



Set of 12 points.



Overall density—surface plot.



DENCLUE - кораци

- У првом кораку препроцесирања формира се мрежа за податке дељењем минималног граничног хиперправоугаоника у d-димензионалне хиперправоугаонике дужине ивице 2σ
- Одређују се мрежне ћелије које садрже тачке. Ћелије су нумерисане на једној ивици граничног хиперправоугаоника
- Вредности нумерације се чувају у дрвету претраживања ради ради касније обраде.
- За сваку сачувану мрежну ћелију чува се број тачака, сума тачака у ћелији и везе са суседним ћелијама
- У другом кораку се врши кластеровање узимајући у обзир само ћелије са великим бројем тачака и ћелије које су повезане са њима
- За сваку тачку k, функција локалне густине израчунава се узимајући у обзир само оне тачке које су из мрежних ћелија које су у њеној близини
- DENCLUE спаја густе ћелије које се могу спојити путем тачака које имају густину већу од задатог параметра ξ
- Крајњи резултат су само кластри чија је густина већа или једнака ξ

SNN

SNN (Shared Near Neighbour Graph)

- Иницијално, алгоритам одређује K најближих суседа сваке тачке у улазном скупу
- Сличност парова тачака се рачуна ка број заједничких најближих суседа обе тачке у пару
- Сличност пара је 0 ако друга тачка није у скупу најближих суседа првој тачки
- Густина сваке тачке је једнака броју тачака унутар растојања ε
- Тачке су класификоване као тачке језгра ако је густина тачака $\geq$ већа од задатог прага
- Кластери се формирају од тачака у језгру, при чему истом кластеру припадају тачке које су на растојању $\leq \varepsilon$
- Тачка је шум ако је на растојању већем од ε од било које тачке језгра
- Преостале тачке се доделе кластеру који садржи најсличније тачке језгра

Хибридни алгоритми

BRIDGE - комбинација K-средина и приступа заснованог на густини

- Изврши се алг. K-средина и свакој тачки додели идентификација кластера
- За сваки кластер K-средина изврши се DBSCAN са параметрима наслеђеним од K-средина
- Помоћу DBSCAN се одреде тачке које јесу и нису су шум
- Изврши се K-средина са првобитним центроидима уз искључивање тачака које су шум

CUBN - комбинација K-средина и приступа заснованог на растојању

- У првом кораку се одређују тачке на граници
- Методом најближих суседа се кластерују тачке на граници. Добија се скуп кластера са граничним тачкама
- Методом најближих суседа се кластерују унутрашње тачке и тачке из скупа кластера са граничним тачкама

GRIDCLUS

- Алгоритам дели простор података помоћу координатне мреже у облику решетке сачињене од хиперкоцки
- Тачке се смештају у неку од хиперкоцки у зависности од вредности њихових атрибута
- Кластеровање се врши методом тражења суседа (хиперкоцки које се додирују)

BANG

- Проширење GRIDCLUS алгоритма - простор се дели у хијерархијски скуп блокова
- Блокови су соротирани у опадајућем редоследу према густини
- Блокови са највећом густином постају центри кластера
- Остали блокови се итеративно кластерују према њиховој густини тако што формирају нови кластер или се додељују постојећим

WaveCLuster

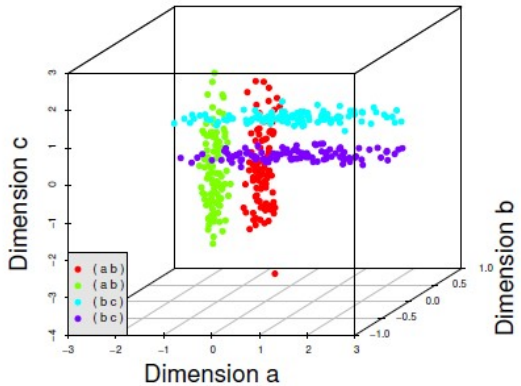
- Може да кластерује, поред осталих, просторне и мултимедијалне податке
- Користи се трансформација улазних података таласићима и начажење кластера у тако трансформисаним подацима
- Алгоритам је отпоран на шум, може да открије кластере произвољног облика са различитим нивоом детаља
- Може да ради са великом количином улазних података

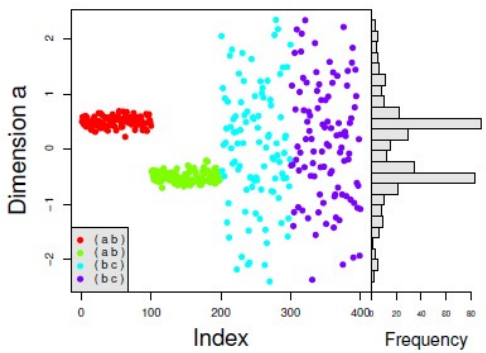
STING

STING - STatistical INformation Grid-based clustering method

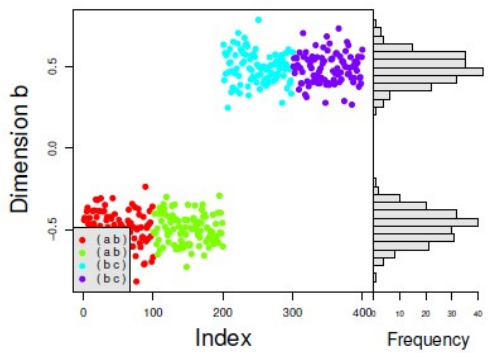
- Кластерује просторне базе података и обезбеђује регион-оријентисане упите.
- STING дели просторно окружење у правоугаоне ћелије и чува ћелије у хијерархијској мрежној структури која има облик дрвета
- STING одржава статистичке податке за сваку ћелију у хијерхијском дрвету на основу које се могу једноставно израчунати статистички параметри за родитељ и дете ћелије

Кластеровање по подпросторима

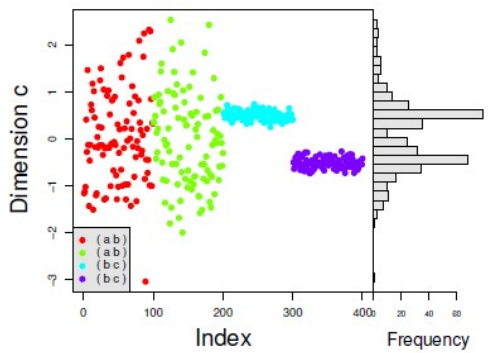




(a) Dimension a

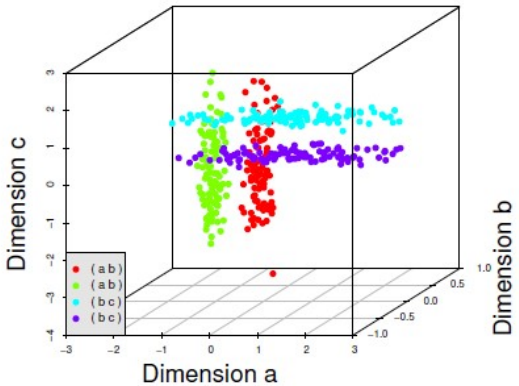


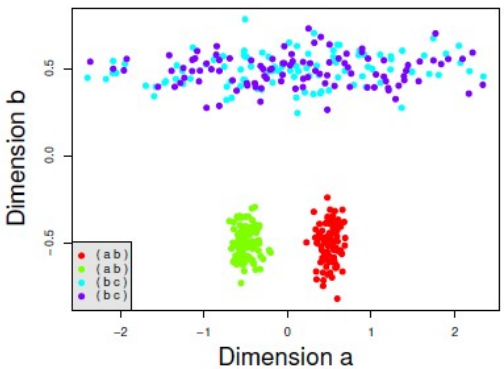
(b) Dimension b



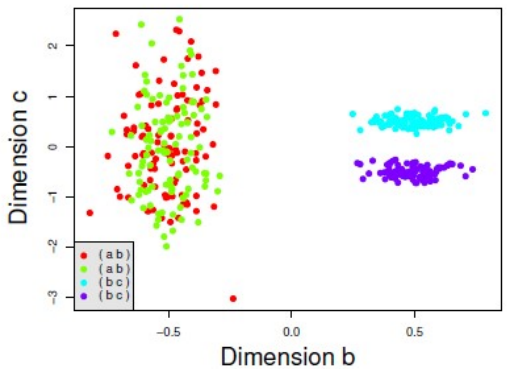
(c) Dimension c

Кластеровање по подпросторима

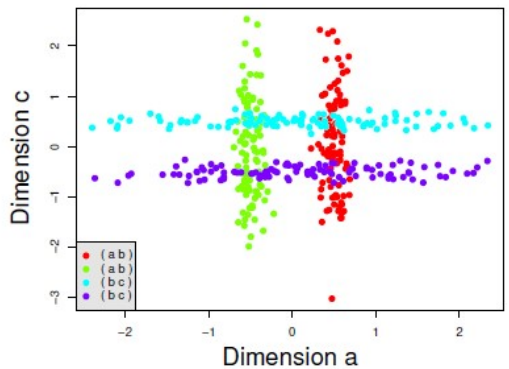




(a) Dims a & b



(b) Dims b & c



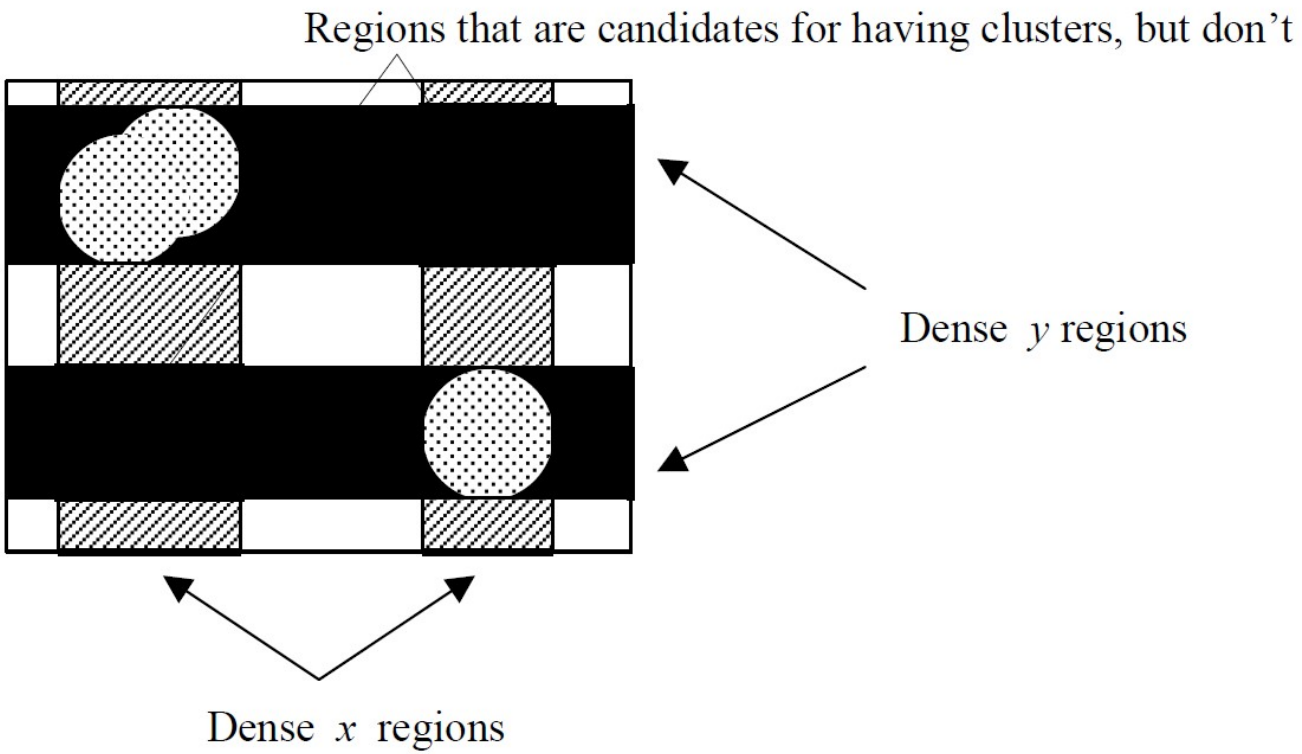
(c) Dims a & c

CLIQUE

CLIQUE (CLustering In QUEst)

- Један пд првих алгоритама који проналази кластере унутар потпростора података
- Хибридни алгоритам заснован на густини и мрежи који врши кластеровање по потпросторима високодимензионих нумеричких података
- Одређује кластере угнеждене у потпросторима без велике интервенције корисника
- Користи APRIORI технику за кластеровање потпростора
- Пронађени густи потпростори се сортирају по покривености, где је покривеност дефинисана као део скупа података који покривају густе јединице у потпростору
- Алгоритам затим проналази густе суседне ћелије у сваком од изабраних потпростора користећи претрагу у дубину. Кластери настају комбиновањем таквих мрежних ћелија користећи похлепну стратегију
- CLIQUE је у стању да пронађе различите врсте и облике кластера, као и било који број кластера у било којем броју димензија

CLIQUE


Regions that are candidates for having clusters, but don't
Dense y regions
Dense x regions

ENCLUS

ENCLUS (ENtropy-based CLUStering)

- Модификација CLIQUE алгоритма која не мери директно густину или покривеност, већ ентропију
- Потпростор са кластерима обично има нижу ентропију од потпростора без кластера
- Ентропија се смањује како се повећава густина ћелија
- Користи АПРИОРИ принцип и приступ одоздо према горе као и CLIQUE за одређивање значајних потпростора
- Може да пронађе кластере произвољног облика који су угнеждени у потпросторима оригиналног скупа података

MAFIA

MAFIA (Merging of Adaptive Finite IntervAls)

- Модификација CLIQUE алгоритма која не користи фиксну величину ћелија решетке са једнаким бројем делова у свакој од димензија, већ конструише адаптивну мрежу ћелија за сваку димензију
- Циљ модификација је побољшање кластеровања у потпросторима и омогућавање паралелизације процеса кластеровања ради обраде великих скупова података
- Алгоритам формира хистограм за одређивање минималног броја ћелија неке димензије, а затим комбинује суседне ћелије сличне густине да би се формирале веће
- Када су ћелије дефинисане, MAFIA наставља слично као CLIQUE, користећи АПРИОРИ приступ формира кластере у простору веће димензије користећи густе ћелије
- MAFIA проналази произвољан број кластера произвољног облика у потпросторима различите величине

OPTIGRID

OptiGrid (OPTimal GRID-Clustering)

- Користи процену густине ради одређивања ценатара кластера као да је кластеровање рађено DENCLUE алгопритмом
- Кластер је регион са концентрисаном густином центриран око јаког атрактора густине или локалног максимума функције густине, при чему је густина већа од задатог прага
- Рекурзивном поделом простора атрибута у вишедимензионе решетке формира се оптимална мрежа партиција конструисањем хиперравни
- Конструисане хиперравни деле простор на области мале густине и чувају просторе високе густине (кластере), и посебно центре кластера
- Хиперравни се одређују коришћењем скупа скупих договорених линеарних пројекција простора атрибута

OPTIGRID варијанте

O-Cluster (Orthogonal partitioning CLUSTERing)

- Уклања ограничење OPTIGRID-а које се односи на скалабилност у односу на величину меморије
- Користи технику случајног узорковања и бафер релативно мале величине
- Користи стратегију деобе на делове паралелно са координатним осама ради налажења густих региона
- Поред хиперравни користи и статистички тест ради провере квалитета хиперравни које секу простор

CBF (Cell-Based Filtering)

- Фокусира се на скалабилност мрежне структуре
- Може да ради са великим датотекама у меморији
- Омогућава ефикасно уношење и дохватање кластера из мрежне структуре
- Дефинише јасне критеријуме за конструкицју хиперравни које деле простор

PROCLUS

PROCLUS (PROjected CLUStering)

- Варијанта K-медоид алгоритма у кластеровању по потпросторима
- Алгоритам се састоји из три фазе
  - У првој фази се користи похлепни алгоритам (случајан избор) за одабир скупа потенцијалних ценатара кластера
  - Фаза итерације за сваки од к текућих медоида се одређује потпростор минимизацијом стандардне девијације растојања тачака у суседству медоида у односу на медоид у свим димензијама
  - Тачке се додељују релевантном медоиду и чине потпростор сваког од медоида
  - Кластери се побољшавају заменом лоших медоида све док се побољшава квалитет кластеровања
  - У фази постпроцесирања се уклањају тачке које представљају шум

ORCLUS и GDILC

ORCLUS –ORiented projected CLUSter generation

- Приступ личи на K-средина
- Дијагонализује матрице коваријанси сваког кластера и на основу њих налази информације о пројекцијама у потпросторима

GDILC –Grid-based Density-IsoLine Clustering

- Врши кластеровање употребом density-isoline слике
- Претпоставља се да су сви улазни подаци нормализовани
- Имплицитно се рачуна density-isoline слика, која представља контуру густине тачака
- Кластери се одређују на основу израчунате слике

Алгоритми засновани на моделу

Алгоритми ове групе покушавају да оптимизују поклапање између посматраних података и неког математичког модела израчунавањем вероватноће. Ове методе се заснивају на претпоставци да подаци задовољавају мешавину основних дистрибуција вероватноће. У пракси, сваки кластер се може математички представити параметарском дистрибуцијом вероватноће, као што су нпр. Гаусова или Поасонова расподела. Такви типови дистрибуција могу да моделирају елипсоидне кластере.

На основу тога, проблем кластеровања улазних података се преводи у проблем процене параметара, тако да се тачке које највероватније имају исту дистрибуцију групишу у (појединачне) кластере

Поступак:

- За дате дистрибуције, обично истог типа али са различитим параметрима, случајно се бира једна од њих и генерише објекат који јој припада
- Процес се понавља m пута, где је m број објеката

EM

У општем случају није познато које тачке су генерисане којом расподелом. Због тога се израчунава вероватноћа да свака тачка припада свакој дистрибуцији и затим користи ове вероватноће да се израчуна нова процена за параметре (који максимизују вероватноћу). Итерација се наставља док се параметри према процени не мењају или се мењају веома мало.

EM се примењује када је расподела Гаусова (нормална) или Пуасонова. EM је сличан алгоритму K-средина. Алгоритам K-средина за Еуклидске податке је специјалан случај EM алгоритма са сферичном нормалном расподела са једнаким матрицама коваријанси, али са различитим срединама.

EM алгоритам конвергира ка локалном максимуму, крајње решење јако зависи од почетног избора параметара, што доводи до процене субоптималног максимума вероватноће.

EM

```
/* EM algoritam */
Izdvojiti inicijalni skup parametara modela
/* Nacin izbora slican kao kod K-sredina
  (slucajno ili na neki drugi nacin) */
repeat
    Korak ocekivanja:
    za svaki objekat izracunati verovatnocu
    da svaki objekat pripada svakoj distribuciji

    Korak maksimizacije:
    s obzirom na dobijene verovatnoce u koraku
    ocekivanja odrediti novu procenu parametara
    koji maksimizuju ocekivanu verovatnocu
until parametri se ne menjaju
    /* Alternativno, staje se ako je promena
    parametara manja od zadatog praga */
```

COOLCAT

- Кластерује категоричке податке користећи ентропију
- Користи објекте као центре кластера
- Циљ кластеровања је минимизација ентропије у подели тачака по кластерима
- За почетне кластере узима се скуп од k максимално различитих објеката, тако да су парови ентропија кластера максимални
- Сви преостали објекти смештају у један кластер тако да је у сваком кораку повећање ентропије резултујућег кластеровања минимално
- У итеративном поступку се тачке распоређују у кластере израчунавањем очекиване ентропије која настаје након постављања објекта у сваки од кластера и избор кластера за који је та очекивана ентропија минимална

STUCCO

- STUCCO - Search and Testing for Understandable Consistent Contrasts
- Уводи концепт контраст скупова
- Формира се дрво претраживања које се користи за израчунавање свих могућих канонички (лексикографских) комбинација вредност атрибута
- За сваку комбинацију (групу) претражује се база и одређује подршка те комбинације, да ли ту грану треба поткресати као и да ли грана има наследнике
- Значајност контраст скупова се одређује тестирањем нулте хипотезе да је подршка контраст скупа једнака у свим групама или, алтернативно, да је независна од припадности групи
- Приказују се контраст скупови који имају неочекиване вредности вероватноће

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
- Обрађено у курсу ИП1

CLUTO

CLUTO - Software for Clustering High-Dimensional Datasets
http://glaros.dtc.umn.edu/gkhome/cluto/cluto/overview

time ./vcluster -rclassfile=klase_15.txt -clustfile=def_15_i1.txt -crfun=i1 -clmethod=direct

-agglofrom=30 broads1.cluto_cista_matrica.csv 15 >izvestaj.15_i1.txt