ADATALAPÚ MEGOLDÁSOK

Adatállományok típusai, adatminőség, adatok előfeldolgozása

Adatállományok jellemzői

- **Dimenzió**: az adatállomány objektumainak attribútumszáma
  dimenzió probléma → dimenziócsökkentés
- **Ritkaság**: nem 0 adatok elfordulásának száma
  (sok esetben a bejegyzések kevesebb mint 1%-a!)
- **Felbontás**: az adatok tulajdonságai gyakran eltérőek különböző felbontás mellett, sőt az adatállományban található mintázatok szintén függenek a felbontás szintjétől

Adatállományok típusai

- Számos típus
- Leggyakoribb típusok:
  - Rekord adatok
  - Gráfalapú adatok
  - Rendezett adatok

Rekord adatok

- Az adatállomány rekordok (adatobjektumok) gyűjteménye
- Rekord = adatmezők (attribútumok) rögzített halmaza
- Legegyszerűbb esetben:
  - Nincs explicit kapcsolat rekord és adatmező között
  - Minden rekord ugyanazzal az attribútumhalmazzal rendelkezik
  - Tárolása: pl. relációs adatbázis
- Tranzakciós adatok
- Adatmátrix
- Ritka adatmátrix

Tid	Térítés	Családi állapot	Adóköteles jövedelem	Késedelmes adós
1	igen	egyedülálló	125000	nem
2	nem	házas	100000	nem
3	nem	egyedülálló	70000	nem
4	igen	házas	120000	nem
5	nem	elvált	95000	igen
6	nem	házas	60000	nem
7	igen	elvált	220000	nem
8	nem	egyedülálló	85000	igen
9	nem	házas	75000	nem
10	nem	egyedülálló	90000	igen

Tranzakciós adatok

- Minden rekordban (tranzakcióban) tételek halmaza
- Példa: Élelmiszerbolt
  - Tranzakció = megvásárolt termékek
  - Tétel = egyes termékek
- Ezt vásárlói kosár adattípusnak nevezzük
- Az attribútumok legtöbbször binárisak
- De lehetnek diszkrétek vagy folytonosak is (vásárolt áru mennyisége, elköltött pénzösszeg, stb.)

TID	TÉTELEK
1	kenyér, üdítöital, tej
2	sör, kenyér
3	sör, üdítöital, pelenka, tej
4	sör, kenyér, pelenka, tej
5	üdítöital, pelenka, tej

Adatmátrix

- Az adatállományban az adatobjektumok mind ugyanazzal a rögzített numerikus attribútumhalmazzal rendelkeznek

→ Adatobjektumok tekinthetők pontoknak (vektoroknak) egy többdimenziós térben. Minden dimenzió megfelel egy attribútumnak.

Projection of x Load	Projection of y Load	Distance	Load	Thickness
10.23	5.27	15.22	27	1.2
12.65	6.25	16.22	22	1.1
13.54	7.23	17.34	23	1.2
14.27	8.43	18.45	25	0.9

→ $m \times n$ (adat)mátrix, ahol a sorok felelnek meg az objektumoknak és az oszlopok az attribútumoknak

- Rekord típus változata, de mivel numerikus adatokból áll, ezért az adatok transzformálására és manipulálására standard mátrixműveletek alkalmazhatóak.

Ritka adatmátrix

- Speciális esete az adatmátrixnak
- Az attribútumok egyforma típusúak és aszimmetrikusak
- A tranzakciós adatok például egy olyan ritka mátrix, ahol az elemek 0 és 1 értékűek lehetnek
- Másik példa: dokumentum adatok
  Ábrázoljuk a dokumentumokat kifejezésvektorként
  → dokumentum-kifejezés mátrix
- Ritka adatmátrixokból csak a nem nulla elemeket tároljuk

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Gráfalapú adatok

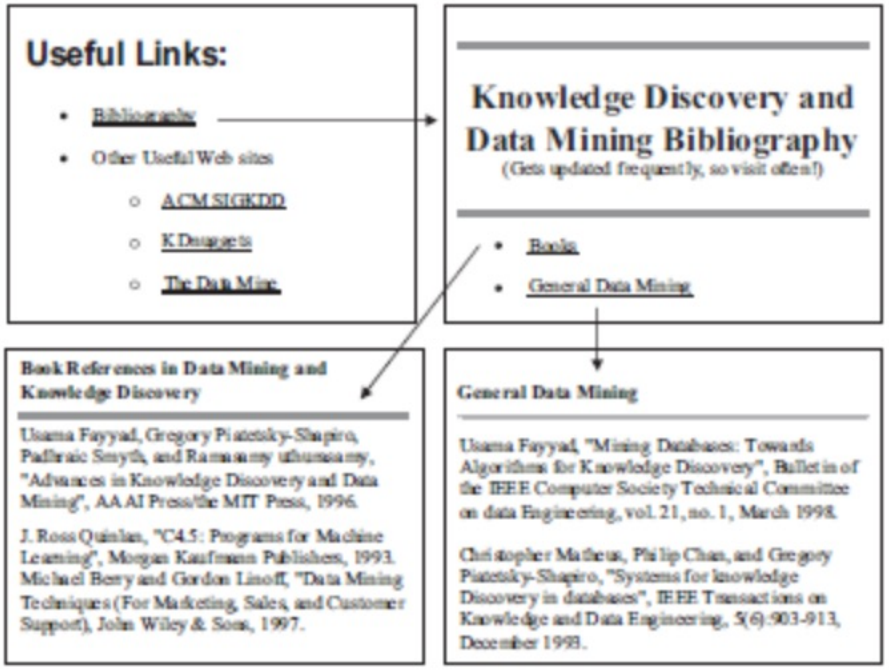
- Az adatoknak kényelmes és hatásos reprezentációja lehet egy gráf

- Két eset:
  1. a gráf az adatobjektumok közötti kapcsolatokat tartalmazza
  2. magukat az adatobjektumokat reprezentáljuk gráfokkal

Adatok objektumok közötti kapcsolatokkal

- Objektumok közti kapcsolatok gyakran hordoznak fontos információt → gráf reprezentáció
- Adatobjektumok = gráf csúcsai
- Objektumok közti kapcsolatok = gráf élei, élek súlya, iránya
- Példa: weboldalak

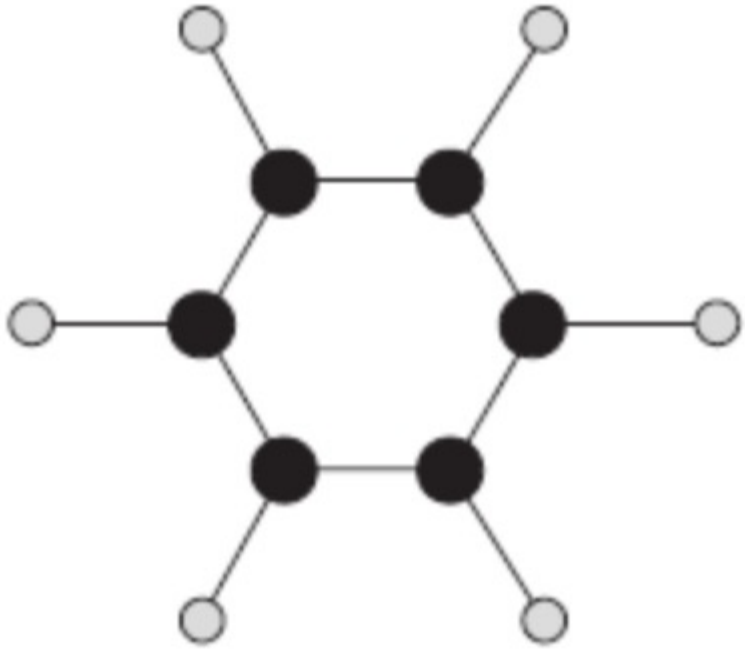
Tartalmaznak különféle tartalmakat és hivatkozásokat más weboldalakra.
A keresőmotorok figyelembe veszik az oldalak közti hivatkozásokat is.



Adatok gráfobjektumokkal

- Az adatobjektumoknak struktúrája van
- Al-objektumokat is tartalmazhatnak
- Kapcsolatok vannak
- Példa: kémiai vegyület

  Csúcsok az atomok
  Élek a kémiai kötések



Rendezett adatok

- Egyes attribútumok kapcsolatai között szerepel azok tér- vagy időbeli rendezettsége is
- Ilyenek lehetnek:
  - Szekvenciális adatok
  - Sorrendi adatok
  - Idősor adatok
  - Térbeli adatok

Szekvenciális adatok

- Más néven időbeli adatok
- A rekord adatok olyan kiterjesztése, ahol minden rekordhoz egy időpont van hozzárendelve
- Időbeli lefolyású cselekvés mintázatok ismerhetők fel
  Példa: a cukorkák eladásának csúcspontja Halloween előtt van; aki DVD lejátszót vesz, az DVD-ket is vesz az azt követő időszakban

Időpont	Vásárló	Vásárolt tételek
t1	C1	A, B
t2	C3	A, C
t2	C1	C, D
t3	C2	A, D
t4	C2	E
t5	C1	A, E

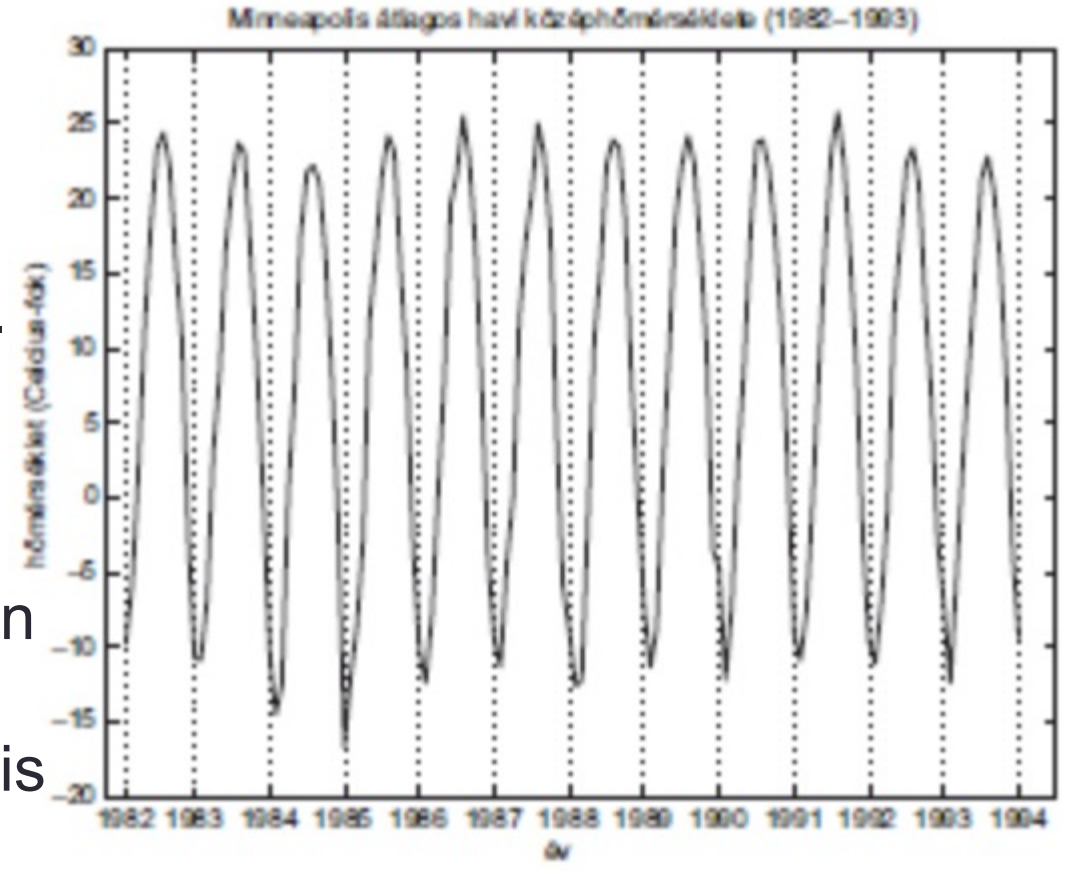
Vásárló	Időpont és vásárolt tételek
C1	(t1: A,B) (t2:C,D) (t5:A,E)
C2	(t3: A, D) (t4: E)
C3	(t2: A, C)

Sorrendi adatok

- Önálló entitások sorozatai alkotják (például szó- vagy betűsorozat)
- Hasonló a szekvenciális adatokhoz, viszont itt nincsenek időbélyegek, helyettük sorozatban elfoglalt pozíciók vannak
- Például növények és állatok genetikai infóit ábrázolhatjuk a génekként ismert nukleotidok sorozataként

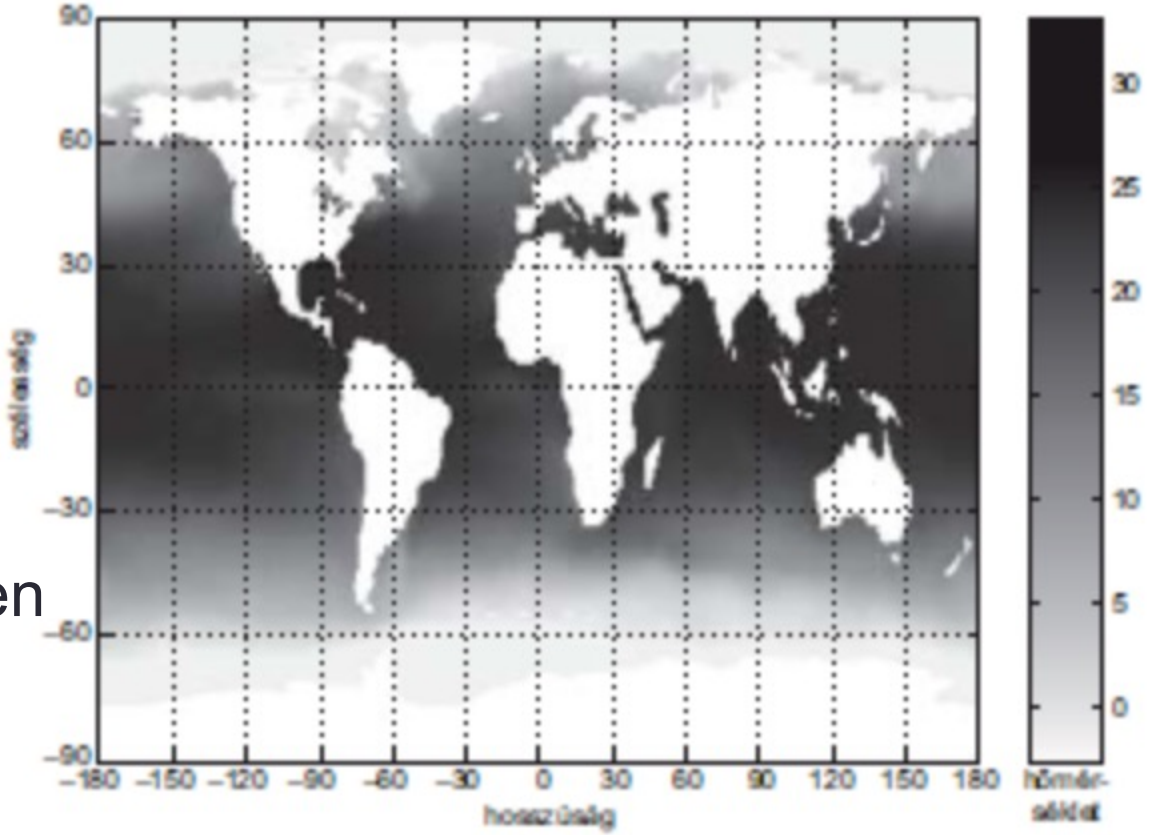
Idősor adatok

- Szekvenciális adatok speciális esete
- Minden rekord egy idősor, azaz időben elvégzett mérések egy sorozata
- Példa: Minneapolis átlagos havi közép-hőmérsékletének idősora 1982-től 1994-ig. Fontos figyelembe venni az időbeli autokorrelációt, azaz, hogy ha két mérés időben közel van egymáshoz, akkor a mérések értékei is



Térbeli adatok

- Adatobjektumok rendelkezhetnek térbeli attribútumokkal, mint helyzet, terület, stb.
- Ilyenek az időjárás adatok (csapadék, hőmérséklet, nyomás), melyeket számos földrajzi helyen gyűjtenek
- Térbeli adatok fontos jellemzője a térbeli autokorreláció, azaz fizikailag közel lévő objektumok jellemzően más szempontokból is hasonlóak



Nem rekord típusú adatok kezelése

- Legtöbb adatalapú módszert, algoritmust rekord adatokhoz, vagy azok valamilyen változatához, például tranzakciós adatokhoz vagy adatmátrixokhoz tervezték
- Mi van akkor ha az adatok nem rekord típusúak?

→Az adatobjektumokból kinyerjük a jellemzőiket

→Ezeket felhasználva minden objektumhoz létrehozunk egy hozzá tartozó rekordot

Adatminőség

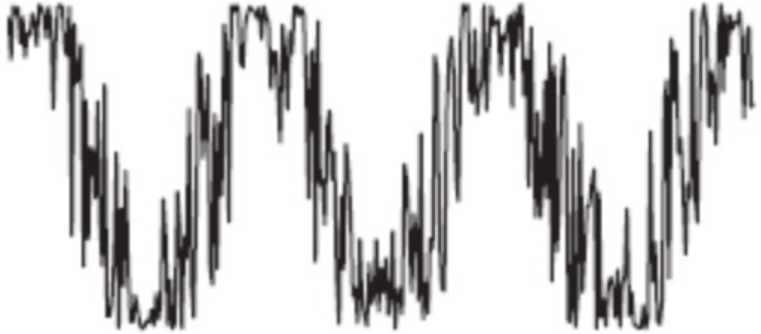
- Statisztikában általában egy előre meghatározott adatminőségi szint érhető el
  - kísérlettervezéssel és
  - kérdőívtervezéssel
- Viszont! Adatalapú megoldások és adatbányászati megoldások esetén nincs mindig lehetőség az adatminőségi problémák megelőzésére!
- Ezért fontos
  - az adatminőségi problémák felismerése és javítása
  - valamint az alacsony adatminőséget toleráló algoritmusok használata

Mérési és adatgyűjtési hibák

- A mérési hiba: olyan probléma, amely a mérési folyamat eredményeként merül fel
- A feljegyzett érték valamilyen mértékben eltér a valóstól
- Folytonos attribútumok esetén a mért és valós érték numerikus különbségét nevezzük hibának
- Az adatgyűjtési hiba: adatobjektumok, attribútumértékek kihagyása vagy adatobjektumok helytelen felvétele
- A hibák lehetnek szisztematikusak vagy véletlenszerűek

Zaj és technikai hibák

- A zaj a mérési hibák véletlen komponense
- Magában foglalhatja egy érték torzulását vagy hibás objektumok felvételét



- Determinisztikusabb jelenségek is okozhatnak adathibákat
- Példa: egy csík egy fotósorozat minden fotóján ugyanazon a helyen
- Az adatok ilyen determinisztikus torzulását technikai hibáknak nevezzük

Pontosság, torzítás, helyesség

- A statisztikában és a kísérleti tudományokban a mérési folyamat és az eredményéül előálló adatok minősége a pontossággal és a torzítással mérhető
- Pontosság: Az (ugyanazon a mennyiségen végzett) ismételt mérések közelsége egymáshoz
- Torzítás: A mérések szisztematikus ingadozása a mért mennyiségtől
- Helyesség: A mérések értékének közelsége a mért mennyiség valódi értékéhez

  → A helyesség a pontosság és a torzítás függvénye

Kiugró értékek

- 1. olyan adatobjektumok, amelyek jellemzői bizonyos értelemben különböznek az adatállomány legtöbb adatobjektumáétól
- 2. olyan attribútumértékek, amelyek szokatlanok ezen attribútum tipikus értékeit tekintve
- Azaz, beszélhetünk rendellenes objektumokról és értékekről
- A kiugró értékek szabályos adatobjektumok vagy értékek is lehetnek

  → a zajjal ellentétben a kiugró értékek egyes esetekben fontosak lehetnek (csalás és hálózati behatolás észlelése)

Hiányzó értékek

- Egy objektum egy vagy több attribútumának értéke hiányzik
- A hiányzó adatok kezelésére több stratégia létezik
  - Adatobjektumok vagy attribútumok törlése
  - Hiányzó értékek becslése
  - A hiányzó értékek figyelmen kívül hagyása az elemzés során

- További mérési és adatgyűjtési hibák:
  - Inkonzisztens értékek
  - Duplikált adatok

Adatok előfeldolgozása

- Aggregálás
- Mintavétel
- Dimenziócsökkentés
- Jellemzők részhalmazainak kiválasztása
- Új jellemzők létrehozása
- Diszkretizálás és binarizálás
- Változó transzformáció

- Két nagy csoport: (i) az adatobjektumok és attribútumok kiválasztása az elemzéshez, (ii) illetve az attribútumok létrehozása és cseréje

Aggregálás

- Az aggregálás kettő vagy több objektum egyetlen objektummá egyesítését jelenti

→Az adatok redukciójának eredményeként létrejövő kisebb adatállományok kevesebb memóriát és feldolgozási időt igényelnek

→Az aggregálás megváltoztathatja a hatáskört és a skálát azzal, hogy az alacsony szintű nézet helyett az adatok egy magas szintű nézetét biztosítja

- Objektumok és attribútumok csoportjainak viselkedése gyakran stabilabb, mint az egyedi objektumoké és attribútumoké
- Aggregált mennyiségeknek, mint az átlagok és az összegek, kisebb az ingadozása, mint az aggregált egyedi objektumoknak

Aggregálás

Tranzakció azonosító	Tétel	Bolt helye	Dátum	Ár	...
⋮	⋮	⋮	⋮	⋮	
101123	Karóra	Chicago	09/06/04	$25,99	...
101123	Elem	Chicago	09/06/04	$5,99	...
101124	Cipő	Minneapolis	09/06/04	$75,00	...
⋮	⋮	⋮	⋮	⋮	

- Azonban az aggregálás egyik hátránya annak a veszélye, hogy fontos részleteket veszítünk el

Mintavétel

- Az adatok egy részhalmazának kiválasztása az elemzéshez
- Statisztika esetén: a teljes adathalmaz előállítása túl költséges vagy időigényes lenne
- Adatbányászat esetén: túl költséges vagy időigényes lenne az összes adat feldolgozása

→ lehetővé válik egy jobb, de költségesebb algo. használata

- Hatékony mintavétel alapelve: egy mintát használva majdnem olyan jó eredményt kapunk, mintha a teljes adatállományt használnánk, amennyiben a minta reprezentatív
- Reprezentatív minta: közelítőleg olyan (számunkra fontos) tulajdonsággal bír, mint az eredeti adatállomány

Mintavételi megközelítések

- Egyszerű véletlen mintavétel: minden egyes objektum kiválasztásának megegyezik az esélye
  - visszatevés nélküli mintavétel
  - visszatevéses mintavétel
- Rétegzett mintavétel: objektumok előre meghatározott csoportjaiból indul ki

→ különböző előfordulási gyakoriságok összeegyeztetése

  - minden csoportból egyenlő számú objektumot veszünk
  - az egyes csoportokból kiválasztott objektumok száma arányos az adott csoport méretével

Mintavétel és információveszteség

- A nagyobb mintanagyság növeli annak valószínűségét, hogy a minta reprezentatív lesz, ugyanakkor megszünteti a mintavétellel elért nyereség nagy részét is. Viszont kis mintanagyság mellett mintázatokat hagyhatunk ki vagy hibás mintázatokat észlelhetünk.

→Adaptív vagy progresszív mintavételi sémák alkalmazása

Dimenziócsökkentés

- Számos adatbányászati algoritmus jobban működik, ha a dimenziószám -- az adatok attribútumszáma – kisebb
- Kiküszöbölhetők a lényegtelen jellemzők és csökkenthető a zaj
- Érthetőbb modellhez vezethet
- Adatok könnyebb ábrázolását teszi lehetővé
- Algoritmusok számára szükséges idő és memóriamennyiség is csökken
- (Részben) kiküszöbölhető a dimenzió probléma

A dimenzió probléma

- A dimenziószám növekedésével az adatok egyre ritkábban helyezkednek el az általuk kitöltött térben
- Osztályozásnál: nem lesz elég adatobjektum ahhoz, hogy létrehozzunk egy olyan modellt, amely minden lehetséges objektumot megbízhatóan besorol egy osztályba
- Klaszterezésnél: a sűrűség és a pontok közötti távolság definíciói, amelyek ennél a módszernél kritikus fontosságúak, veszítenek jelentőségükből

→ csökken az osztályozás pontossága és gyenge minőségű klaszterek jönnek létre

A dimenziócsökkentés lineáris algebrai módszerei

- Főkomponens analízis (PCA -- Principal Component Analysis): lineáris algebrai módszer, amely olyan új attribútumokat (főkomponenseket) tár fel, melyek
  - az eredeti attribútumok lineáris kombinációi,
  - ortogonálisak (merőlegesek) egymásra,
  - az adatokban fellelhető ingadozást maximálisan kifejezik.

- Szinguláris felbontás (SVD -- Singular Value Decomposition)

Jellemzők részhalmazainak kiválasztása

- A jellemzőknek csak egy részhalmazát használjuk (dimenzió csökkentésének másik módja)
- Felesleges jellemző: egy vagy több más attribútumban fellelhető információ nagy részének vagy egészének másolatai
- Lényegtelen jellemző: semmi olyan információt nem tartalmaznak, amely hasznos lenne az elvégzendő adatbányászati feladathoz
- A felesleges és lényegtelen jellemzők csökkenthetik az osztályozás pontosságát és a feltárt klaszterek minőségét

→ Célszerű ezek elhagyása, azaz jellemzők egy részhalmazának kiválasztása

Kiválasztási stratégiák

- Beágyazott megközelítések: az algoritmus futása során maga dönti el, hogy mely attribútumokat használja, és melyeket hagyja figyelmen kívül (pl. döntési fák)
- Szűrő (filter) megközelítések: az adatbányászati algoritmus futása előtt megtörténik egy olyan módszer alkalmazásával, amely független az adatbányászati feladattól (pl. minimálisan korreláló attribútum párok)
- Borító (wrapper) megközelítések: az adatbányászati algoritmust fekete dobozként használva az attribútumok legjobb részhalmazának megtalálása

Jellemzők létrehozása

- Új attribútumhalmaz létrehozása az eredeti attribútumokból, amely sokkal hatékonyabban adja vissza az adatállományban lévő fontos információkat

→ lehetővé teszi a dimenziócsökkentés összes említett előnyének kihasználását

- Jellemzők kinyerése
- Az adatok leképezése egy új térre
- Jellemzők konstrukciója