Nagyméretű adathalmazok előfeldolgozása

Jámbor Attila

Budapesti Műszaki és Gazdaságtudományi Egyetem
Számítástudományi és Információelméleti Tanszék

March 5, 2011

Tartalom

Bevezetés

A valós adatok rendszerint

- zajosak,
- hiányosak,
- inkonzisztensek,
- hatalmas méretűek
- és több, heterogén forrásból származnak.

„Minőségi adatbányászathoz minőségi adatokra van szükség.”

Bevezetés

A gyenge minőségű adatok előfordulásának okai:

- rögzítéskor még nem érhetők el az adatok
- egy adat nem tűnik fontosnak, ezért nem rögzítjük
- hibás működés a rögzítés vagy tárolás során
- hibásan működő adatgyűjtés
- hálózati hiba továbbításkor
- inkonzisztens formátumok (pl. dátum)
- duplikált adatok

Tartalom

Attribútum típusok

Attribútum típusok:

- kategória típusú
- sorrend típusú
- intervallum típusú
- arány skálájú

Attribútum típusok

A *kategória típusú* attribútumnál az attribútum értékei között csak azonosságot tudunk vizsgálni.
Mindössze annyit tudunk mondani, hogy $a = b$ vagy $a \neq b$.

A kategória típusú attribútum egy speciális esete a *bináris attribútum*, ahol az attribútum csak két értéket vehet fel.

A *sorrend típusú* attribútumoknál az értékeket sorba tudjuk rendezni, azaz az attribútum értékén teljes rendezést tudunk megadni.
Ha tehát $a \neq b$, akkor még azt is tudjuk, hogy $a < b$ vagy $a > b$.

Attribútum típusok

Ha az eddigiek mellett meg tudunk adni egy, az adatokon értelmezett $+$ függvényt, akkor *intervallum típusú* attribútumról beszélünk.

Ha egy intervallum típusú attribútumnál meg lehet adni zérus értéket, akkor az attribútum *arány skálájú*. Az arány skálájú attribútumok megadására rendszerint valós számokat használunk, így szokás őket *valós* attribútumoknak is hívni.

Attribútum típusok

Példák különböző attribútum típusokra.

- Kategória: szemszín
- Bináris: nem
- Sorrend: legmagasabb iskolai végzettség
- Intervallum: születési év
- Arány skálájú: testmagasság

Hasonlósági mértékek

Az adatbányászat során szükségünk lesz arra, hogy attribútumokkal leírt elemek között hasonlóságot definiáljunk. Minél több közös attribútummal rendelkezik két elem, annál hasonlóbbak egymáshoz.

A gyakorlatban a hasonlóság helyett a *különbözőséget* mérjük.

Tulajdonságok:

- Az x és y elem különbözősége: $d(x, y)$,
- $d(x, x) == 0$,
- $d(x, y) = d(y, x)$,
- a különbözőségre teljesül a háromszög egyenlőtlenség, azaz $d(x, y) < d(x, z) + d(z, y)$,
- a $d(x, y)$ különbözőséget az x és y elemek távolságának is nevezik.

Bináris attribútumok különbözősége

m db bináris attribútummal leírt x és y elemek különbözősége a következő:

	1	0	$\sum$
1	q	r	q+r
0	s	t	s+t
$\sum$	q+s	r+t	m

Invariáns hasonlóság: $d(x, y) = \frac{r + s}{m}$

Variáns hasonlóság (*Jaccard-koefficiens komplementere*):

$d(x, y) = 1 - \frac{q}{m - t} = \frac{r + s}{m - t}$

Kategória típusú attribútumok különbözősége

A különbözőség mértéke a nemegyezések relatív száma:

$d(x, y) = \frac{u}{m}$, ahol m a kategória típusú attribútumok száma, u pedig a nem egyező attribútumok száma.

A kategória típusú attribútumokra létezik a *Jaccard-koefficiens komplementere*.

Sorrend típusú attribútumok különbözősége

Sorrend típusú attribútumok esetén az egyes attribútumértékeket egész számokkal helyettesítik, majd ezeken alkalmazzák valamelyik intervallum típusú hasonlóságot.

Ha több sorrend típusú attribútumunk van, ahol a lehetséges állapotok száma eltérő, akkor célszerű mindegyiket a $[0, 1]$ intervallumba normalizálni.

Intervallum típusú attribútumok különbözősége

Az m db intervallum típusú attribútummal (*általában valós számokkal*) leírt elemre tekinthetünk úgy, mint egy vektorra az m-dimenziós vektortérben.

Az x és y elemek különbözőségén a vektoraik különbségének normáját értjük, azaz $d(\overrightarrow{x}, \overrightarrow{y}) = \left\| \overrightarrow{x} - \overrightarrow{y} \right\|$.

Euklideszi-norma: $L_2(\overrightarrow{z}) = \sqrt{|z_1|^2 + |z_2|^2 + \cdots + |z_m|^2}$

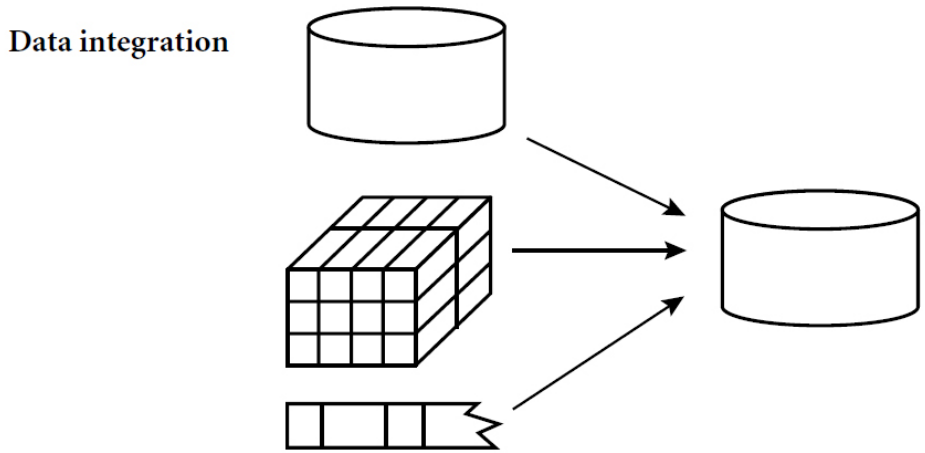
Minkowski-norma: $L_p(\overrightarrow{z}) = \left(|z_1|^p + |z_2|^p + \cdots + |z_m|^p\right)^{1/p}$

Ha bizonyos attribútumoknak nagyobb szerepet szánunk:
$L_2(\overrightarrow{z}) = \sqrt{w_1 \cdot |z_1|^2 + w_2 \cdot |z_2|^2 + \cdots + w_m \cdot |z_m|^2}$, ahol w_i az i-edik attribútum súlya és $\sum_{i=1}^{m} w_i = 1$.

Tartalom

Integráció

Az *integráció* során összegyűjtjük a különböző forrásokból származó adatokat egy közös helyre, például egy adattárházba (data warehouse)



Lehetséges adatforrások:

- Adatbázisok
- Adatkockák
- Fájlok

Integráció

Lehetséges nehézségek:

- Entitások azonosítása (Entity identification proglem)
  - Eltérő attribútumnevek (*customer_id* vs. *cust_id*)
  - Eltérő attribútumértékek (*„Jámbor Attila”* vs. *„Jámbor A.”*)
  - Megoldás lehet a metaadatok vizsgálata (attribútumok értelmezése, típusa, értékkészlete; null elemek kezelése)
- Redundancia (*éves fizetés* vs. *havi fizetés*)
- Értékkonfliktus (Data value conflict)
  - Eltérő reprezentáció, skálázás, kódolás
  - Eltérő mértékek (*kilométer* vs. *mérföld*)
  - Eltérő tartalom (*szálloda*)

Tartalom

Transzformáció

A *transzformáció* során az adatainkat olyan formára hozzuk, hogy azok megfelelőek legyenek az adatbányász algoritmusok számára.

Lépései:

- Értékek kisimítása (Smoothing)
- Aggregálás
- Általánosítás (Generalization)
- Új attribútumok létrehozása
- Adatok elrontása
- Normalizálás

Data transformation $-2, 32, 100, 59, 48 \longrightarrow -0.02, 0.32, 1.00, 0.59, 0.48$

Transzformáció

- Értékek kisimítása: Zaj és kiugró értékek eltávolítása. Lehet dobozolás, regresszió, klaszterezés.
- Aggregálás: Több adat helyettesítése eggyel (*havi fizetés* $\rightarrow$ *éves fizetés*)
- Általánosítás: Az alacsony szintű értékeket magasabb szintűekkel helyettesítjük (*város* $\rightarrow$ *ország*, *életkor* $\rightarrow$ *fiatal/öreg*)
- Új attribútumok létrehozása: Új attribútumokat hozunk létre, hogy növeljük az eredmények érthetőségét, az algoritmus sebességét (*szélesség/magasság* $\rightarrow$ *terület*)
- Adatok elrontása:
  - Megvizsgáljuk, hogy az adatbányász módszerünk mennyire érzékeny a zajra
  - Publikussá szeretnénk tenni az adathalmazt azok pontos jelentése nélkül
- Normalizálás: Ez hasznos lehet osztályozási feladatoknál vagy távolságszámításnál.

Normalizálás

A normalizálás során az attribútum értékkészletét egy másik (rendszerint egységnyi) tartományra transzformáljuk.

Típusai:

- Min-max normalizálás

$$v' = \frac{v - min_A}{max_A - min_A}(new\_max_A - new\_min_A) + new\_min_A$$

- Zérus pont normalizálás (z-score normalization)

$$v' = \frac{v - \overline{A}}{\sigma_A},$$

ahol $\overline{A}$ az A átlaga, σ_A pedig A szórása.

- Decimális skálázás (Normalization by decimal scaling)

$$v' = \frac{v}{10^j},$$

ahol j a legkisebb egész szám, amire $Max(|v'|) < 1$.

Tartalom

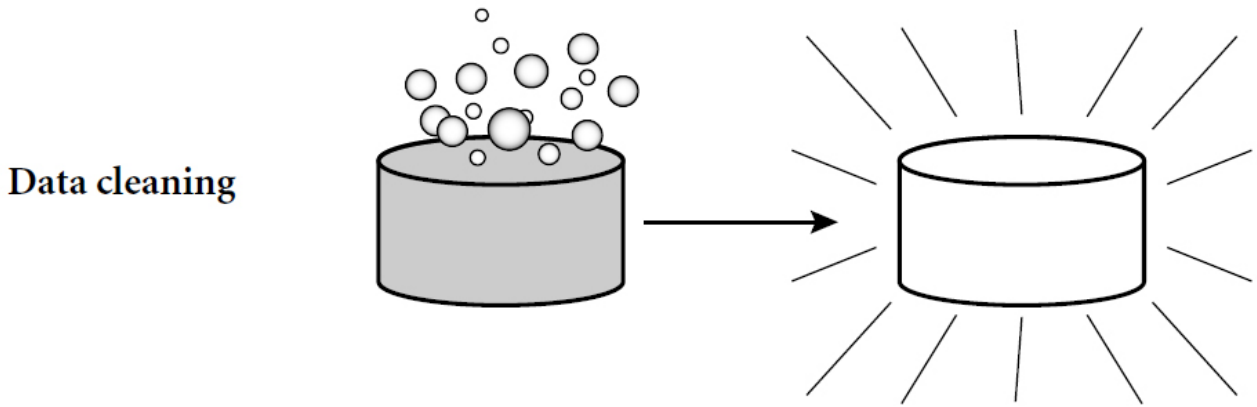
Adatok tisztítása

Az adatok *tisztítása* révén eltávolítjuk a zajt és kijavítjuk az inkonzisztens állapotot.

A *piszkos* (dirty) adaton végzett adatbányászat eredménye megbízhatatlan a felhasználók számára.

Lépései:

- Hiányzó adatok feltöltése
- Zaj kezelése
- Inkonzisztencia feloldása



Adatok tisztítása

- Hiányzó adatok feltöltése
  - Figyelmen kívül hagyás/törlés: rendszerint, ha az osztályozó attribútum hiányzik
  - Feltöltés kézzel: időigényes
  - Globális konstans használata: *„Unknown”* vagy ∞^-
  - Ismert elemek átlagával való feltöltés
  - Azonos osztályban levő rekordok átlagával való feltöltés (*credit_risk*)
  - Legvalószínűbb értékkel való feltöltés: az ismert attribútumok felhasználásával döntési fákat vagy dedukciót használva
  - Több új elem létrehozás: kategória típusú attribútumoknál

Adatok tisztítása

- Zaj kezelése: a zaj egy véletlen hiba vagy eltérés a mért értékekben.
  - Dobozolás (binning): részletesebben a diszkretizálásnál
  - Regresszió
  - Klaszterezés
- Inkonzisztencia feloldása
  - Egyediség szabály (unique rule)
  - Folytonossági szabály (consecutive rule)
  - Null szabály (null rule): megmondja, hogy mely attribútumok vehetnek fel null értéket, és hogyan kell értelmezni őket

Tartalom

Diszkretizálás (Data discretization)

A diszkretizálás (*kvantálás*) során az kiválasztott attribútum lehetséges értékeinek számát csökkentjük (*GPS adatok*).

A folyamat során az értékkészletet intervallumokra osztjuk, és az egyes intervallumokba eső értékeket az intervallum „címkéjével” helyettesítjük, amely csökkentjük és egyszerűsítjük az eredeti adathalamazt.

A diszkretizálás hatásaként az adatbányászat

- felbontása, részletessége csökken,
- eredménye tömörebbé, áttekinthetőbbé válik,
- sebessége, hatékonysága nő.

Diszkretizálás

A felhasznált információt tekintve a diszkretizálás lehet

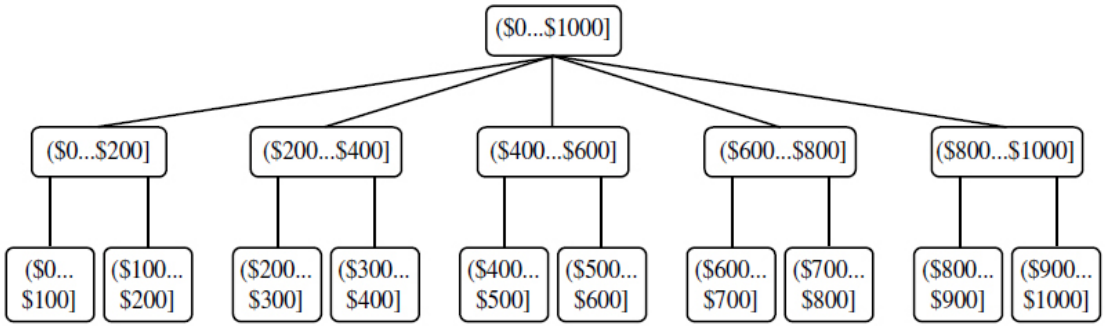
- felügyelt (supervised): figyelembe vesz bizonyos osztály-információkat (class information)
- nem felügyelt (unsupervised): nem vesz figyelembe osztály-információkat.

Irányát tekintve pedig lehet

- fentről lefelé (top-down)
- lentről felfelé (bottom-up)

Diszkretizálás

A diszkretizálás történhet rekurzív módon is, amikor is egy hierarchikus felbontását végezzük ez az attribútumértékeknek. A felbontásból képzett fát fogalmi hierarchiának (concept hierarchy) nevezzük.



A fogalmi hierarchiákban az alacsonyabb szintű fogalmakat magasabb szintű fogalmakkal helyettesítjük. Pl. az életkor megadása helyett csak annyit mondunk, hogy valaki *fiatal*, *középkorú* vagy *öreg*.

Számos diszkretizálási algoritmus esetén a fogalmi hierarchiák automatikusan generálhatók.

Diszkretizálás

Diszkretizálási algoritmusok:

- Binning, hisztogram analízis
- Entrópia alapú diszkretizálás
- χ^2-összevonás
- Klaszter analízis
- Intuitív partícionálás

Binning, histogram analízis (Histogram analysis)

Tulajdonságok:

- Ládákat alakítunk ki
- Az attribútumértékeket a ládák átlagával vagy mediánjával helyettesítjük
- A ládák kialakítása lehet egyenlő nagyságú vagy egyenlő gyakoriságú
- Fentről lefelé típusú, nem felügyelt technika
- Leállási feltétel
  - Minimum szélesség
  - Maximum ládaszám

Entrópia alapú diszkretizálás (Entropy-based discretization)

Tulajdonságok:

- Azt vizsgálja, hogy az egyes felbontások után hogyan változik meg az elemek entrópiája
- Minden iterációban azt az elemet választja vágási pontnak, amely mentén az entrópia változás minimális
- Fentről lefelé típusú, felügyelt technika

Entrópia alapú diszkretizálás

Működés:

- Az adatokat a D halmaz jelölve. Az attribútumok között $\exists A, C$, ahol A a diszkretizálandó attribútum, C pedig egy osztályattribútum (class-label attribute). $C = \{c_1, c_2, \ldots, c_m\}$.
- Kezdetben minden $a \in A$ értéket lehetséges vágási pontnak (split-point) tekintünk. Ha egy $a \in A$ érték vágási pont, akkor a D halmaz felbontható D_1 és D_2 diszjunkt halmazokra. Ekkor

$$D_1 = \{d \in D \mid d.A \leq a\}$$

és

$$D_2 = \{d \in D \mid d.A > a\}.$$

Entrópia alapú diszkretizálás

- Egy felbontás ideális, ha a C attribútum értékeit is diszjunkt módon bontja fel. Egy felbontás minőségét az alábbiak szerint tudjuk mérni:

$$Q_a(D) = \frac{|D_1|}{|D|} Entropy(D_1) + \frac{|D_2|}{|D|} Entropy(D_2),$$

  ahol $|D|$ jelenti a D adathalmaz elemszámát.
- Az entrópia a következőképpen számolható:

$$Entropy(D_1) = -\sum_{i=1}^{m} p_i log_2(p_i),$$

  ahol p_i jelenti a c_i attribútum relatív gyakoriságát a D_1-beli elemek között.

Entrópia alapú diszkretizálás

- Az összes a attribútum közül azt választjuk ki vágási pontnak, amelyre a $Q_a(D)$ érték minimális. Ekkor a D halmazt felbontjuk D_1 és D_2 halmazokra, majd ezt rekurzívan megismételjük.
- Leállási feltétel:
  - $Q_a < \varepsilon, \forall a \in A$
  - A részhalmazok száma meghalad egy küszöbértéket

χ^2-összevonás (Interval merging by χ^2 analysis)

Tulajdonságok:

- Azt vizsgálja, hogy az egyes szomszédos intervallumok mennyire hasonlítanak egymásra
- Minden iterációban azt a két szomszédos intervallumot vonja össze, amelyek a legjobban hasonlítanak egymásra
- Lentről felfelé típusú, felügyelt technika

χ^2-összevonás

Működés:

- Kezdetben minden bejegyzés külön intervallumnak tekintendő.
- Ha K db intervallumunk van, akkor minden $(k, k+1), 0 < k < K$ intervallumpárra kiszámoljuk a χ^2 értéket, majd összevonjuk azt a két intervallumot, amelyre χ^2 minimális volt.
- Ha az A attribútumot szeretnénk diszkretizálni, ahol $A = \{a_1, a_2, \ldots, a_m\}$, akkor χ^2 a következőképpen számolható:

$$\chi^2_k = \sum_{i=k}^{k+1} \sum_{j=1}^{m} \frac{(o_{ij} - e_{ij})^2}{e_{ij}},$$

ahol o_{ij} jelenti a_j relatív gyakoriságát az i. intervallumban, míg e_{ij} jelenti a_j elvárt gyakoriságát az i. intervallumban.

χ^2-összevonás

- $$e_{ij} = \frac{(|D_i|) \times (|\{d, \text{ahol } d \in D, d.A = a_j\}|)}{|D|},$$

  ahol D_i az i. intervallum, D pedig a teljes adathalmaz.
- Leállási feltétel:
  - χ^2 elért egy küszöbértéket (alul-, túldiszkretizálás)
  - Intervallumok száma egy küszöb alá csökkent

Klaszter analízis (Cluster analysis)

Tulajdonságok:

- Az A attribútum értékeit klaszterekre osztja
- Figyelembe veszi az A attribútum eloszlását
- Létezik fentről lefelé és lentről felfelé típusa is
- Részletesebben később (Zsolnai Károly)

Intuitív partícionálás (Discretization by intuitive partitioning)

Tulajdonságok:

- Az *A* attribútum elemeit úgy partícionálja, hogy a határvonalak „barátiak” legyenek
- A 3-4-5 szabályt alkalmazza
- Felülről lefelé típusú

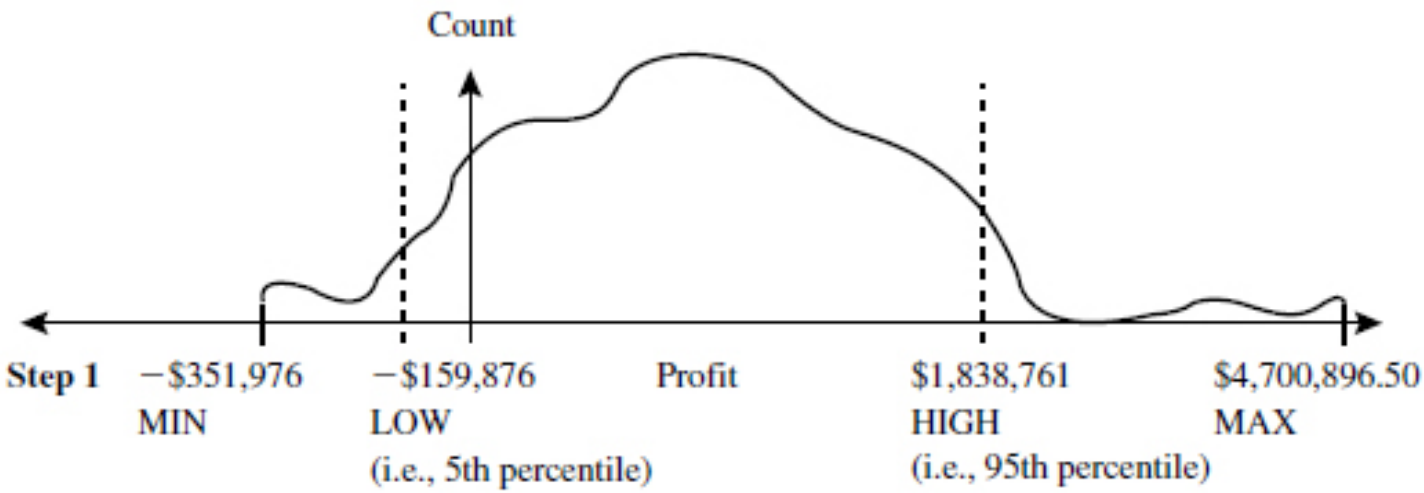
Intuitív partícionálás

3-4-5 szabály (3-4-5 rule):

- Egy intervallumot az alapján oszt fel egyenlő nagyságú részekre, hogy a legnagyobb helyiértéken mekkora az eltérés az intervallum kezdő és végpontja között
- Ha 3, 6, 7, vagy 9 az eltérés, akkor 3 részintervallumra osztja az intervallumot
- Ha 2, 4 vagy 8 az eltérés, akkor 4 részintervallumra osztja az intervallumot
- Ha 1, 5 vagy 10 az eltérés, akkor 5 részintervallumra osztja az intervallumot

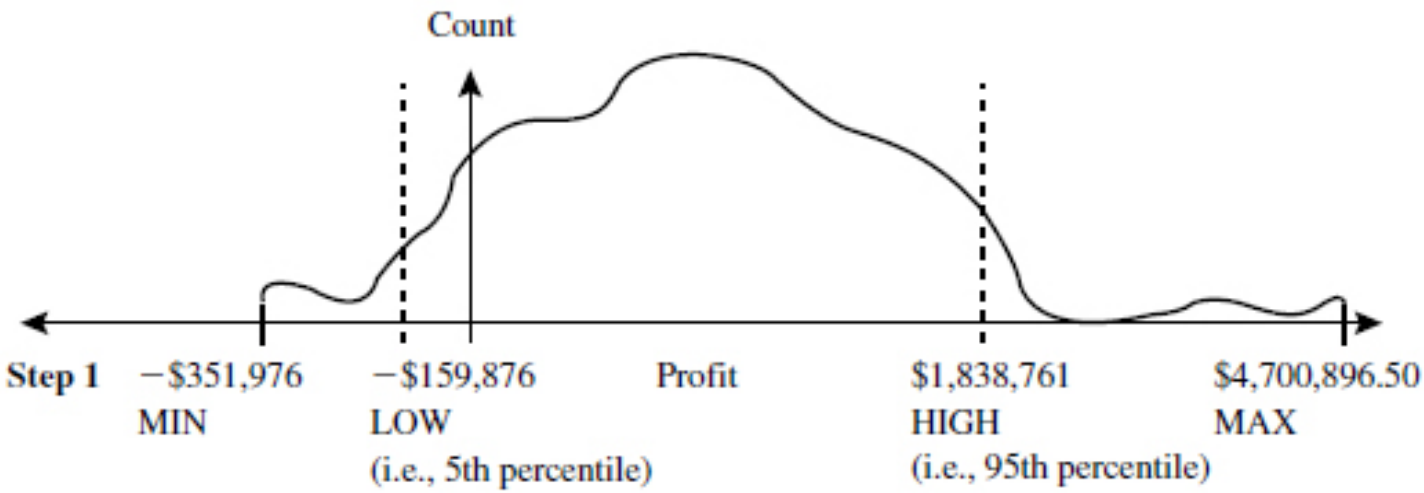
Intuitív partícionálás

Példa



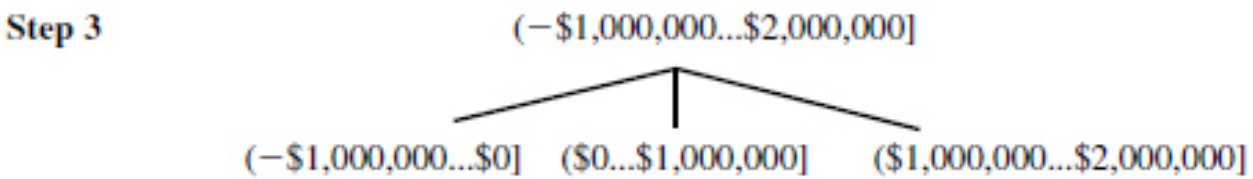
Intuitív partícionálás

Példa

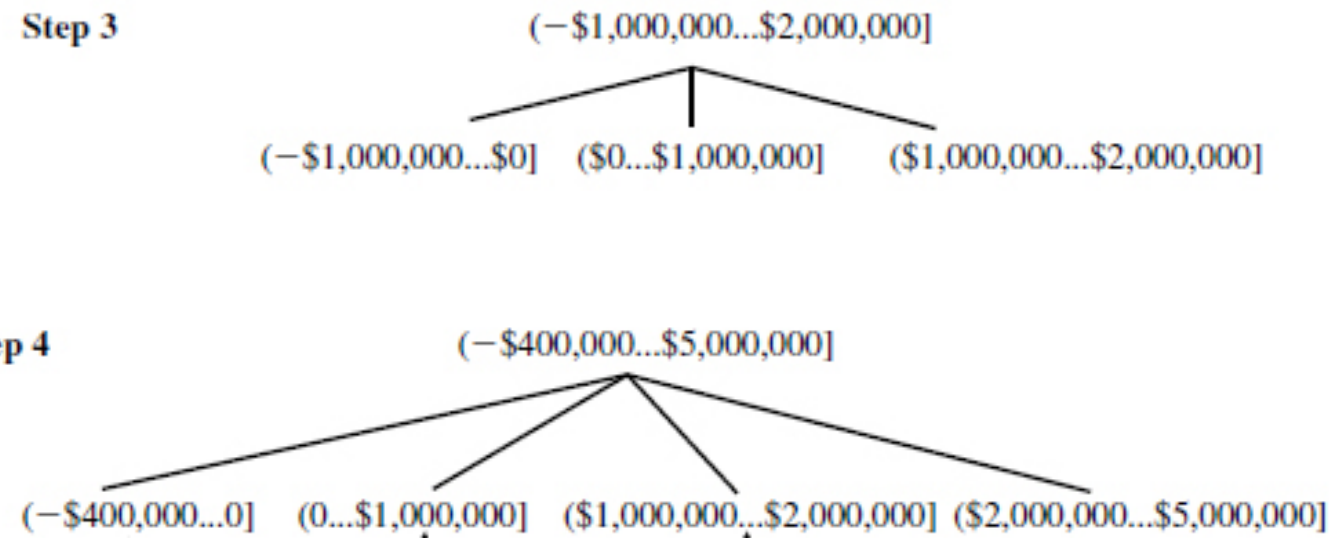


Step 2 msd = 1,000,000 LOW′ = −$1,000,000 HIGH′ = $2,000,000

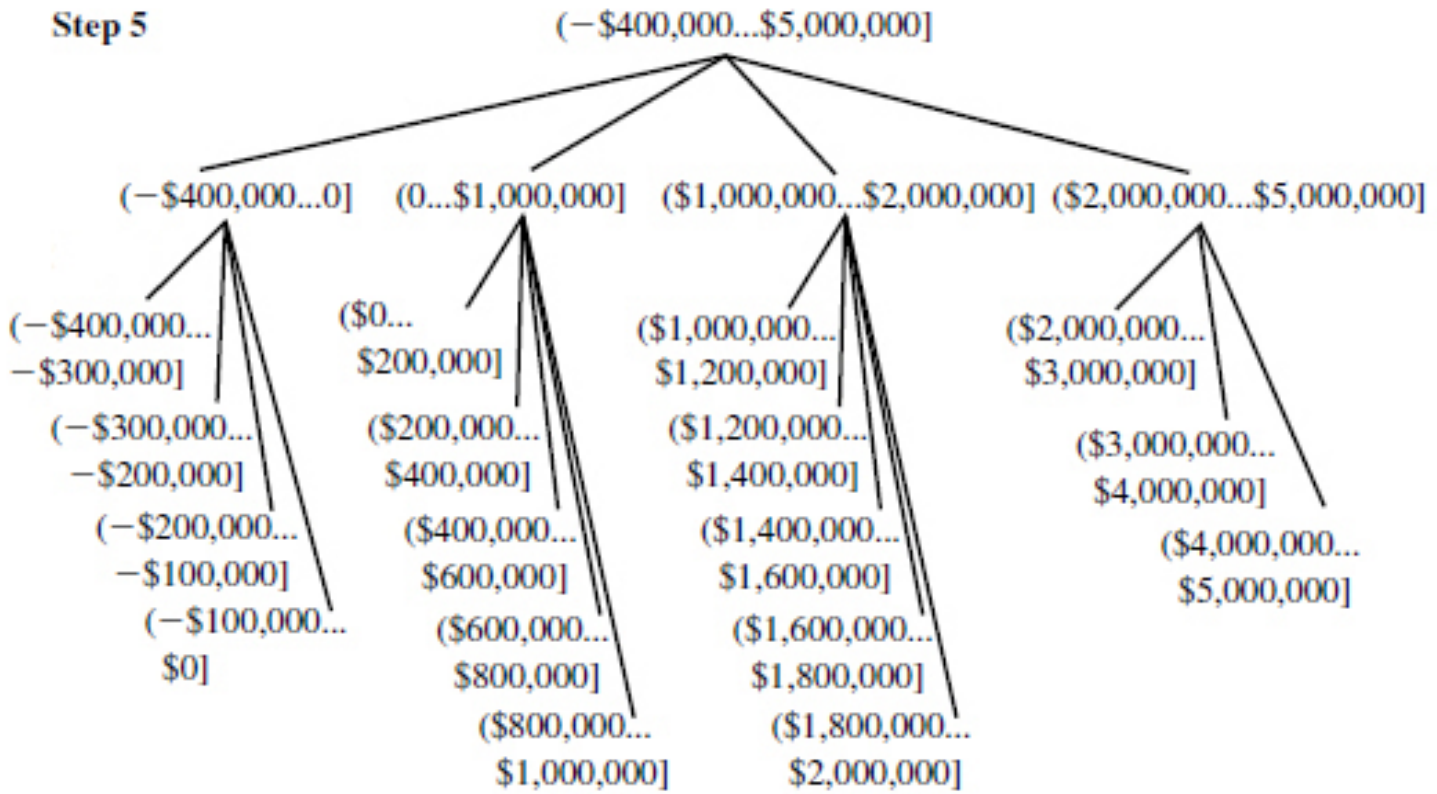
Intuitív partícionálás



Intuitív partícionálás



Intuitív partícionálás



Tartalom

Az adatmennyiség csökkentése

Nagyobb méretű adathalmazon az adatbányászat eredménye pontosabb, ugyanakkor lassabb is.

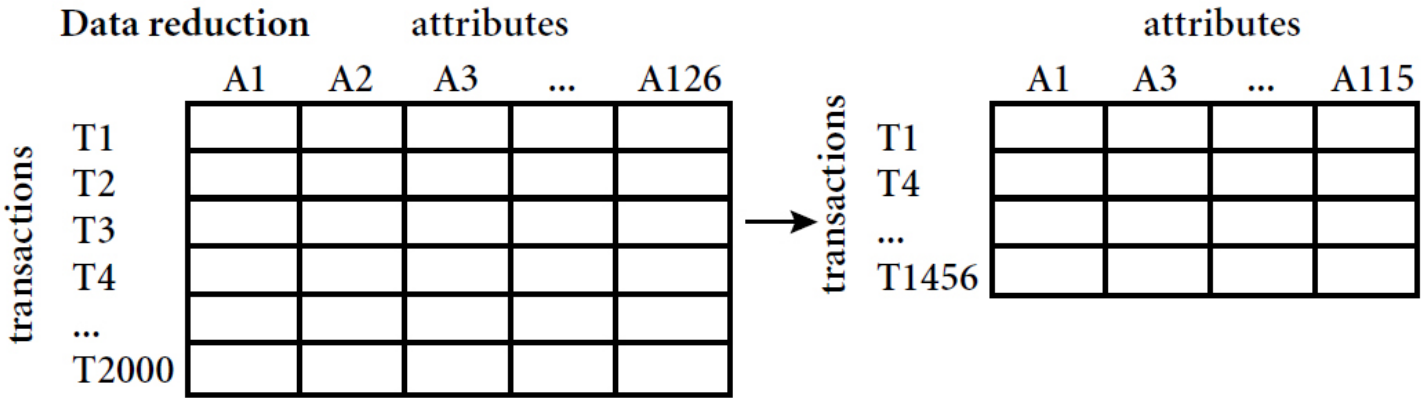
A feldolgozáshoz az adatokat kezelhető méretűre kell *csökkenteni*.

Feltétel, hogy a csökkentett adathalmaznak ugyan azt az analitikus eredményt kell szolgáltatnia, mint az eredetinek.

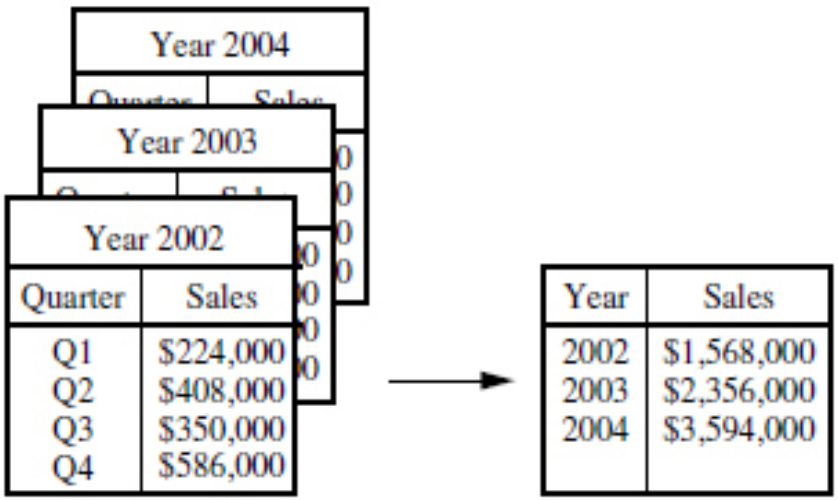
Az adatmennyiség csökkentése

Típusai:

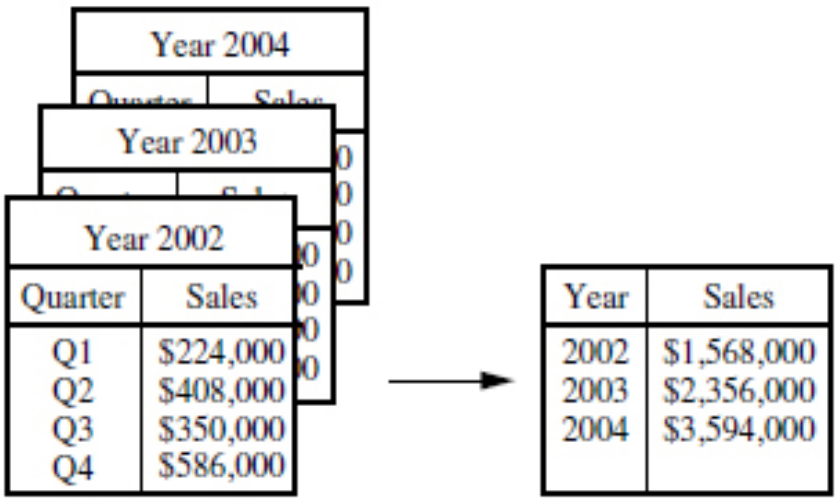
- Adatkocka aggregálás
- Attribútum részhalmaz kiválasztás
- Dimenziócsökkentés
- Mintaszámcsökkentés

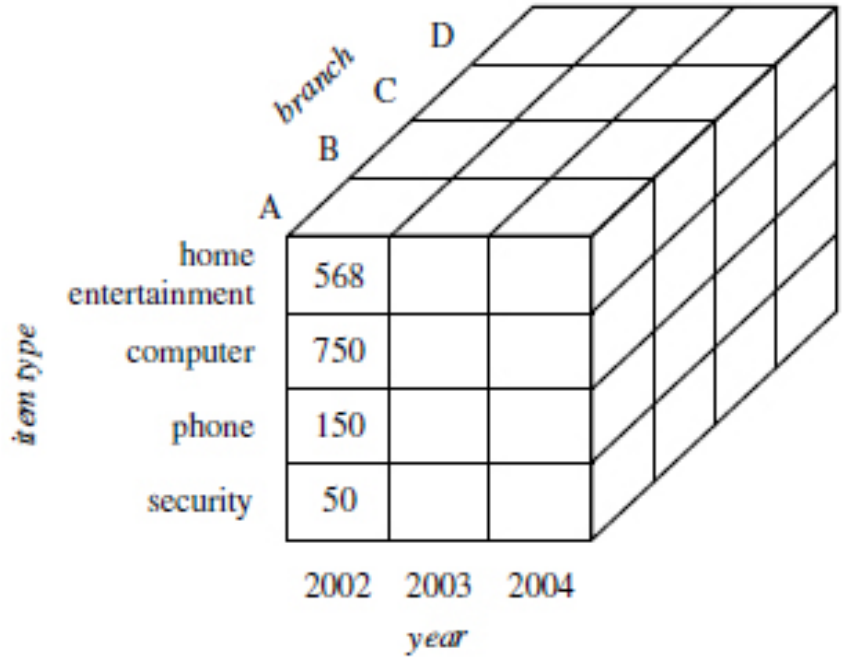


Adatkocka aggregálás (Data cube aggregation)



Adatkocka aggregálás (Data cube aggregation)





Attribútum részhalmaz kiválasztás (Attribute subset selection)

Az eredeti attribútumhalmaz egy részét megtartjuk, a másik részét pedig elvetjük.
A vizsgálat szempontjából irreleváns attribútumok az adatbányász algoritmust lassítják, illetve zavarják.

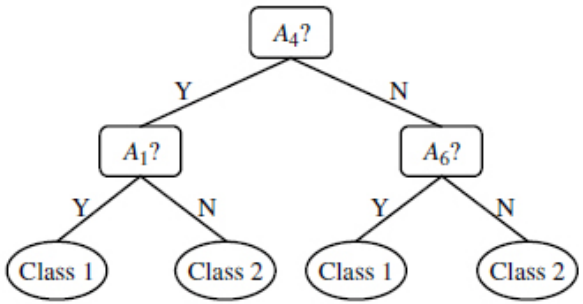
Tulajdonságok:

- Eltávolítja az irreleváns attribútumokat
- Cél megtalálni a legszűkebb részhalmazát az attribútumoknak, amely még azonos eredményhez vezet, mint az eredeti halmaz
- Segít egyszerűsíteni, ezáltal megérteni az algoritmust
- m db attribútum esetén 2^m részhalmaz létezik
- Rendszerint heurisztikán alapuló mohó algoritmust használnak

Attribútum részhalmaz kiválasztás

Típusai:

- Iteratívan növekvő halmaz (stepwise forward): Üres halmazból indul. Lépésenként a legjobb attribútumot adja hozzá.
- Iteratívan csökkenő halmaz (stepwise backward): A teljes halmazból indul. Lépésenként törli a legrosszabb attribútumot.
- Döntési fa indukció (decision tree induction): Döntési fát építünk. A fában szereplő attribútumokat relevánsnak, a többit irrelevánsnak tekintjük.

Forward selection	Backward elimination	Decision tree induction
Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ Initial reduced set: $\{\}$ $\Rightarrow \{A_1\}$ $\Rightarrow \{A_1, A_4\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_3, A_4, A_5, A_6\}$ $\Rightarrow \{A_1, A_4, A_5, A_6\}$ => Reduced attribute set: $\{A_1, A_4, A_6\}$	Initial attribute set: $\{A_1, A_2, A_3, A_4, A_5, A_6\}$  => Reduced attribute set: $\{A_1, A_4, A_6\}$

Dimenziócsökkentés (Dimensionality reduction)

A *dimenziócsökkentés* során a tárolt adatokat kódoljuk vagy transzformáljuk, hogy tárolásuk hatékonyabb legyen.

Típusai:

- Veszteségmentes: az eredeti adathalmaz visszaállítható
- Veszteséges: az eredeti adathalmaz csak közelíthető

Főkomponens analízis (Principal components analysis)

Egy lehetséges veszteséges dimenziócsökkentő eljárás a *főkomponens analízis*.

Lépései:

- Tfh. az adataink m db. attribútummal vannak leírva, amelyek így $m - dimenzis$ vektoroknak tekinthetőek.
- Megkeressük az $m - dimenzis$ tér m db ortogonális egységvektorát.
- Ezeket „fontosság” szerint csökkenő sorrendbe rendezzük.
- Az egységvektorok közül $k \leq m$ db-ot megtartunk, a többit elvetjük.
- Ezzel egy közelítését adtuk meg az adathalmaznak, ugyanis a legkevésbé fontos attribútumokat hagytuk el.

Mintaszámcsökkentés (Numerosity reduction)

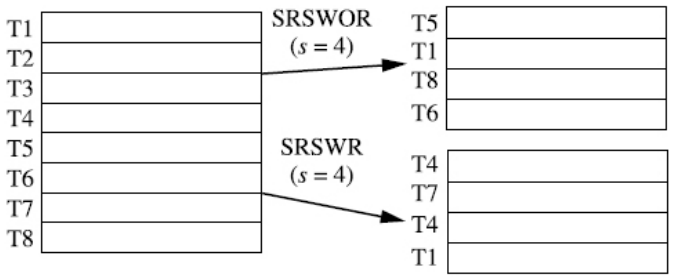
Az eredeti adathalmazt egy kevesebb mintát tartalmazóval helyettesítjük.

Lehetséges típusa a mintavételezés, amely során véletlenszerűen választunk elemeket az eredeti halmazból.

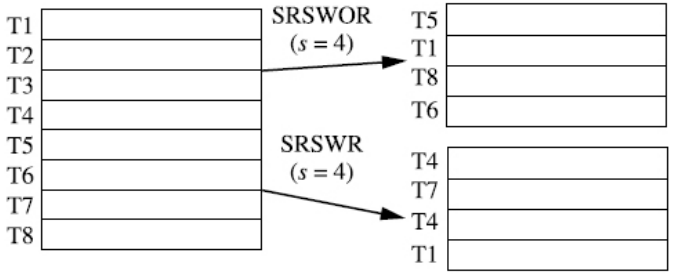
Mintavételezés típusai:

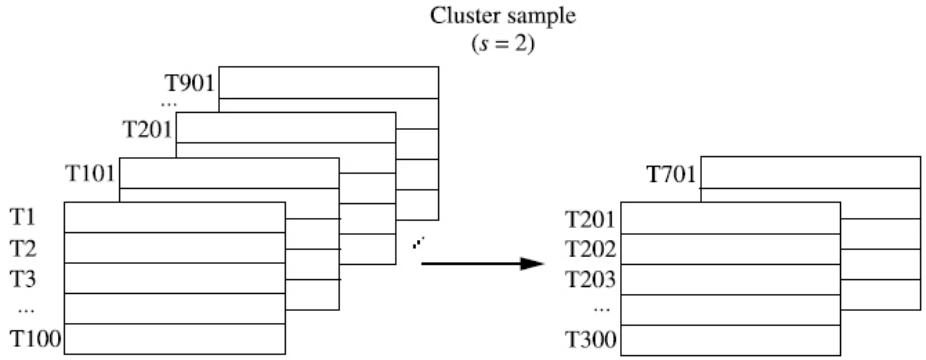
- Visszatevéses/visszatevés nélküli véletlen választás (Simple random sample with/without replacement)
- Klaszter mintavételezés (Cluster sample)
- Rétegzett mintavételezés (Stratified sample)

Mintavételezés

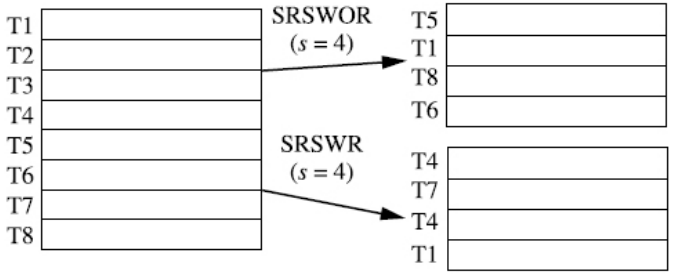


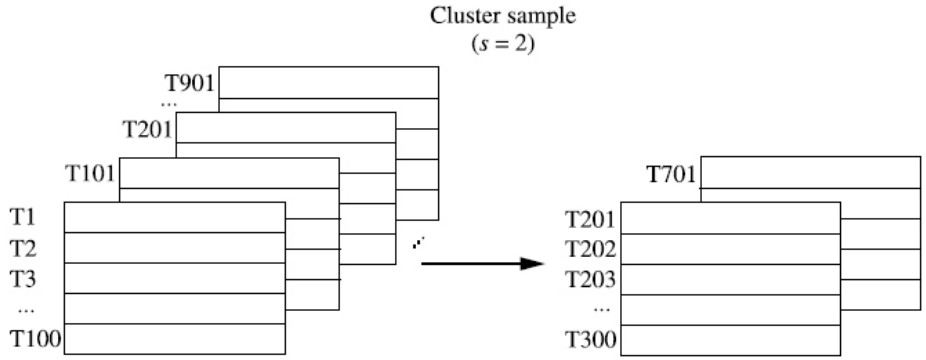
Mintavételezés





Mintavételezés





Stratified sample
(according to *age*)

T38	youth
T256	youth
T307	youth
T391	youth
T96	middle_aged
T117	middle_aged
T138	middle_aged
T263	middle_aged
T290	middle_aged
T308	middle_aged
T326	middle_aged
T387	middle_aged
T69	senior
T284	senior

T38	youth
T391	youth
T117	middle_aged
T138	middle_aged
T290	middle_aged
T326	middle_aged
T69	senior

Mintavételezés

Mennyi mintát vegyünk, hogy torzításmentesen reprezentáljuk az eredeti adathalmazt?

- Tfh. az elemek halmazából az x elem előfordulásának valószínűsége p és m mintát vettünk.
- A mintavételezés hibázik, amennyiben x relatív gyakorisága eltér p-től:

$$hiba(x) = P(|rel.gyakorisag\,(x) - p| \geq \varepsilon)$$

- Jelölje X_i azt a vv.-t, amely 1, ha x-et választottuk az i-dik húzásnál, különben 0.
- Jelölje Y azt a vv.-t, amely $Y = \sum_{i=1}^{m} X_i$. Mivel a húzások egymástól függetlenek, ezért Y eloszlása m, p paraméterű binomiális eloszlás.

Mintavételezés

$$hiba(x) = P\left(\left|\frac{Y}{m} - p\right| \geq \varepsilon\right)$$

Mintavételezés

$$hiba(x) = P\left(\left|\frac{Y}{m} - p\right| \geq \varepsilon\right) = P\left(|Y - m \cdot p| \geq m \cdot \varepsilon\right)$$

Mintavételezés

$$hiba(x) = P\left(\left|\frac{Y}{m} - p\right| \geq \varepsilon\right) = P\left(|Y - m \cdot p| \geq m \cdot \varepsilon\right) =$$

$$P\left(|Y - E[Y]| \geq m \cdot \varepsilon\right)$$

Mintavételezés

$$hiba(x) = P\left(\left|\frac{Y}{m} - p\right| \geq \varepsilon\right) = P\left(|Y - m \cdot p| \geq m \cdot \varepsilon\right) =$$

$$P\left(|Y - E[Y]| \geq m \cdot \varepsilon\right) =$$

$$P\left(Y \geq m \cdot (E[Y] + \varepsilon)\right) + P\left(Y \leq m \cdot (E[Y] - \varepsilon)\right)$$

Mintavételezés

$$hiba(x) = P\left(\left|\frac{Y}{m} - p\right| \geq \varepsilon\right) = P\left(|Y - m \cdot p| \geq m \cdot \varepsilon\right) =$$

$$P\left(|Y - E[Y]| \geq m \cdot \varepsilon\right) =$$

$$P\left(Y \geq m \cdot (E[Y] + \varepsilon)\right) + P\left(Y \leq m \cdot (E[Y] - \varepsilon)\right)$$

Csernov-korlát:

$P\left(Y \geq m \cdot (E[Y] + \varepsilon)\right) \leq e^{-2\varepsilon^2 m}$ és

$P\left(Y \leq m \cdot (E[Y] - \varepsilon)\right) \leq e^{-2\varepsilon^2 m}$,

Mintavételezés

$$hiba(x) = P\left(\left|\frac{Y}{m} - p\right| \geq \varepsilon\right) = P\left(|Y - m \cdot p| \geq m \cdot \varepsilon\right) =$$

$$P\left(|Y - E[Y]| \geq m \cdot \varepsilon\right) =$$

$$P\left(Y \geq m \cdot (E[Y] + \varepsilon)\right) + P\left(Y \leq m \cdot (E[Y] - \varepsilon)\right)$$

Csernov-korlát:
$P\left(Y \geq m \cdot (E[Y] + \varepsilon)\right) \leq e^{-2\varepsilon^2 m}$ és

$P\left(Y \leq m \cdot (E[Y] - \varepsilon)\right) \leq e^{-2\varepsilon^2 m}$, amiből megkapjuk, hogy:

$$hiba(x) \quad \leq \quad 2 \cdot e^{-2\varepsilon^2 m}$$

Mintavételezés

$$hiba(x) = P\left(\left|\frac{Y}{m} - p\right| \geq \varepsilon\right) = P\left(|Y - m \cdot p| \geq m \cdot \varepsilon\right) =$$

$$P\left(|Y - E[Y]| \geq m \cdot \varepsilon\right) =$$

$$P\left(Y \geq m \cdot (E[Y] + \varepsilon)\right) + P\left(Y \leq m \cdot (E[Y] - \varepsilon)\right)$$

Csernov-korlát:
$P\left(Y \geq m \cdot (E[Y] + \varepsilon)\right) \leq e^{-2\varepsilon^2 m}$ és

$P\left(Y \leq m \cdot (E[Y] - \varepsilon)\right) \leq e^{-2\varepsilon^2 m}$, amiből megkapjuk, hogy:

$$\begin{aligned} hiba(x) &\leq 2 \cdot e^{-2\varepsilon^2 m} \\ m &\geq \frac{1}{2\varepsilon^2} ln \frac{2}{hiba(x)} \end{aligned}$$

Mintavételezés

ϵ	δ	$\lvert\mathcal{M}\rvert$
0.05	0.01	1060
0.01	0.01	27000
0.01	0.001	38000
0.01	0.0001	50000
0.001	0.01	2700000
0.001	0.001	3800000
0.001	0.0001	5000000