Data mining

Van boodschappenmandjes tot bio-informatica

Walter Kosters

Informatica, Universiteit Leiden

donderdag 6 april 2006

http://www.liacs.nl/home/kosters/

Wat is Data mining?

Data mining probeert interessante en (on)verwachte patronen te vinden in grote hoeveelheden (on)geordende data.

Bijvoorbeeld:

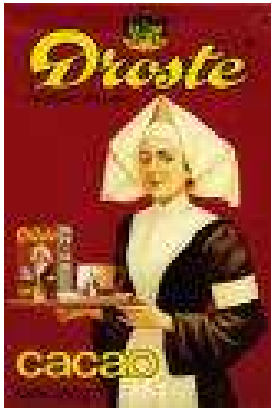
- Boodschappenmandjes: wat en hoe kopen we?
- Bio-informatica: DNA?

Problemen (selectie):

- resultaten: zowel verwacht als onverwacht
- bewegende doelen: steeds andere eisen
- data vergaren: wat/hoeveel kan/mag/is er?
- afstandsbegrip: wie lijkt op wat?

Het Data mining proces — of beter het KDD proces, voor Knowledge Discovery in Databases — wordt vaak als volgt opgedeeld:

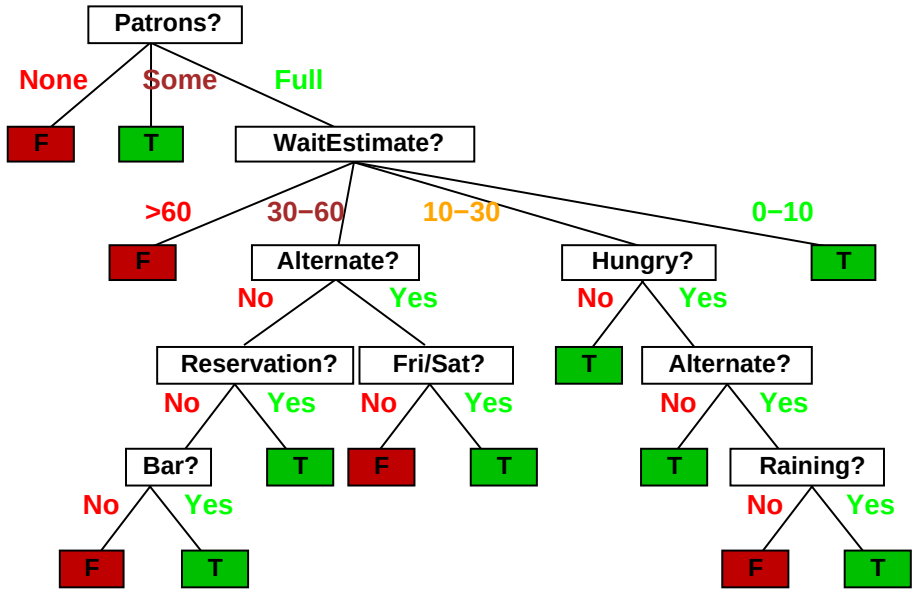
1. data selectie
2. opschonen: de-duplicatie en domein-consistentie
3. verrijking: data-fusie
4. coderen
5. Data mining — het echte gebeuren
6. rapporteren



Er zijn *veel* Data mining technieken, of beter gezegd: algoritmen, die we kunnen gebruiken om data te "minen":

- (niet vergeten:) statistische methoden
- allerlei "machine learning" technieken uit de AI: evolutionaire algoritmen, neurale netwerken, Bayesiaanse netwerken, ...
- allerlei technieken voor clustering en classificatie: beslissingsbomen, ...
- associatieregels
- ad-hoc methoden

Als voorbeeld iets over beslissingsbomen (= decision trees). Meestal wordt Quinlan's algoritme J48 gebruikt om uit een aantal voorbeelden zo'n boom te maken. Het berust op Shannon's informatietheorie. Interne knopen vragen naar een attribuutwaarde, bladeren geven de classificatie.



Het recursieve algoritme werkt als volgt. Stel we zitten in een knoop, en hebben de verzameling V van alle voorbeelden die in die knoop terecht komen. Dan:

1. Geen voorbeelden meer ($V = \emptyset$)? Geef defaultwaarde, bijvoorbeeld "majority-value" van ouderknoop.

2. Alle voorbeelden dezelfde classificatie? Geef die.

3. Geen attributen meer? "Majority-value" van V.

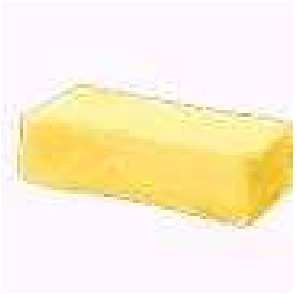
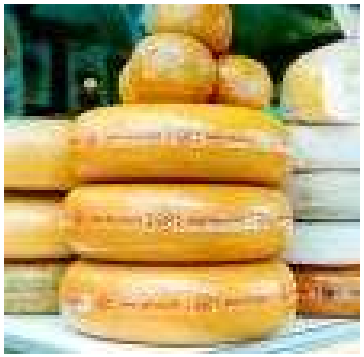
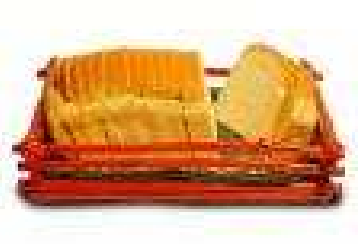
4. Bepaal "beste" attribuut (entropie), splits daarop, en ga recursief verder met de kinderen:

$$\sum_{i=1}^{\nu} \frac{p_i + n_i}{p + n} \cdot \{ - \frac{p_i}{p_i + n_i} \log_2 \frac{p_i}{p_i + n_i} - \frac{n_i}{p_i + n_i} \log_2 \frac{n_i}{p_i + n_i} \}$$

Ter inspiratie bekijken we nu drie thema's:

- associatieregels
- boodschappenmandjes
- bio-informatica

Associatieregels



$\Rightarrow$


We zijn geinteresseerd in verbanden tussen verzamelingen "items": producten, of moleculen, of woorden, of bezoeken aan websites. Regels zijn: "als je boter koopt, koop je meestal ook brood".

Stel we hebben een database met records = transacties = klanten, waarbij ieder record uit een aantal items = producten bestaat. De support van een stel artikelen is het aantal klanten dat al die artikelen koopt, meestal als percentage van het totaal. Een stel artikelen met hoge support (boven een zekere drempel) heet frequent.

Bijvoorbeeld: 20% van de klanten in een supermarkt kopen brood, boter en kaas — en wellicht meer producten.

Stel dat van de α klanten die A kopen, er β ook B kopen. We zeggen dan dat de associatieregel $A \Rightarrow B$ een betrouwbaarheid heeft van β/α.

Bijvoorbeeld: het zou kunnen zijn dat 80% van de klanten die boter en kaas kopen, ook brood aanschaffen.

We zijn benieuwd naar regels $A \Rightarrow B$ met zowel hoge betrouwbaarheid als hoge support, dat wil zeggen hoge support voor A en B samen (20% in ons voorbeeld).

Er is veel nagedacht over associatieregels:

- geef efficiënte algoritmen om ze te vinden
- hoe vind je vervolgens de interessante?
- en hoe ga je om met niet 0/1-en?

Voor dat laatste gebruikt men fuzzy logica, met in plaats van alleen 0/1 (niet-kopen vs. kopen) ook tussenliggende waarden. Voorbeeld: iemands leeftijd kan 0,3 “jong” zijn.

product = item klant = transactie	1	2	3	4	5	6	7	8	9
1	1	1	0	0	1	1	1	1	0
2	1	0	1	0	0	0	0	1	1
3	0	1	1	0	1	0	1	0	0
4	1	0	1	0	1	1	0	1	1
5	0	0	0	0	1	0	0	0	0
6	0	1	0	0	1	0	1	0	0

Zelfs voor deze kleine database is het moeilijk om te zien dat $\{2, 5, 7\}$ het enige drietal is dat "gekocht" wordt door minstens 50% (een vaste drempelwaarde) van de klanten.

Frequente itemsets leveren associatieregels: $\{2, 7\} \Rightarrow \{5\}$.

De meeste snelle implementaties zijn gebaseerd op:

een deel van een iets frequents is zelf ook frequent!

Dit heet de Apriori-eigenschap, omdat het wordt gebruikt in Apriori, een van de meest bekende Data mining algoritmen.

Kleine frequente verzamelingen worden samengevoegd tot grotere. De databases worden opgeslagen in FP-bomen.

Boodschappenmandjes

Winkels analyseren klantgedrag: boodschappenmandjes.
Toepassingen zijn bijvoorbeeld:

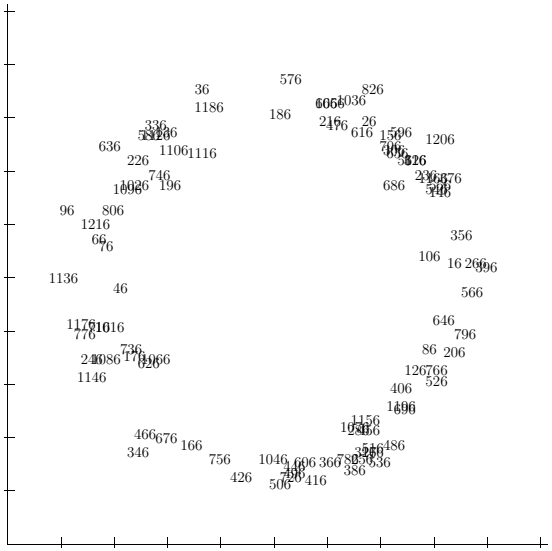
- direct marketing
  "Welke klant doen we een speciale aanbieding?"

- Wat verkopen we?
  "Welke producten zetten we in een kleine winkel, zeg op een station?"

- clustering en classificatie
  "Kunnen we klanten groeperen of herkennen aan hun koopgedrag?" (Ja!)

Een wat algemener doel is het *begrijpen* van klanten.

Een grote winkelketen wilde inzicht in hun “positionering”, gebaseerd op weekverkopen. Dit resulteert in plaatjes als:



Hier zien we 100 winkels.

Om te clusteren heb je een afstandsbegrip nodig. Stel dat twee winkels de afgelopen week dit verkocht hebben:

wijn	brood	kaas	tomaat	banaan
7	120	34	0	40
10	100	21	0	0

Hun afstand kan dan zijn: $3 + 20 + 13 + 0 + 40 = 76$, of $3 + 20 + 13 = 36$, of $3/10 + 20/120 + 13/34$, of ...
Het lijkt redelijk te normaliseren voor de totale verkoop bij een winkel.

Hoe dan ook, er zijn veel mogelijkheden!

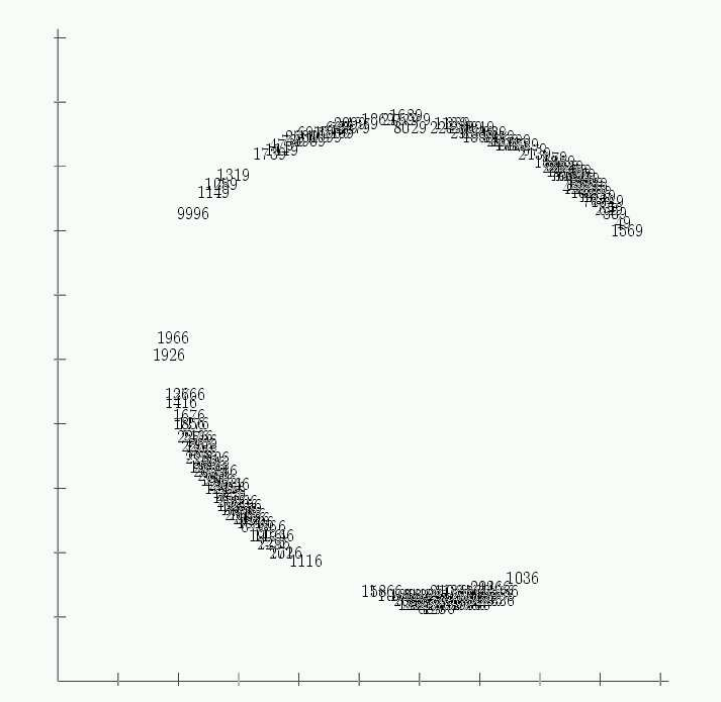
Clustering technieken kunnen

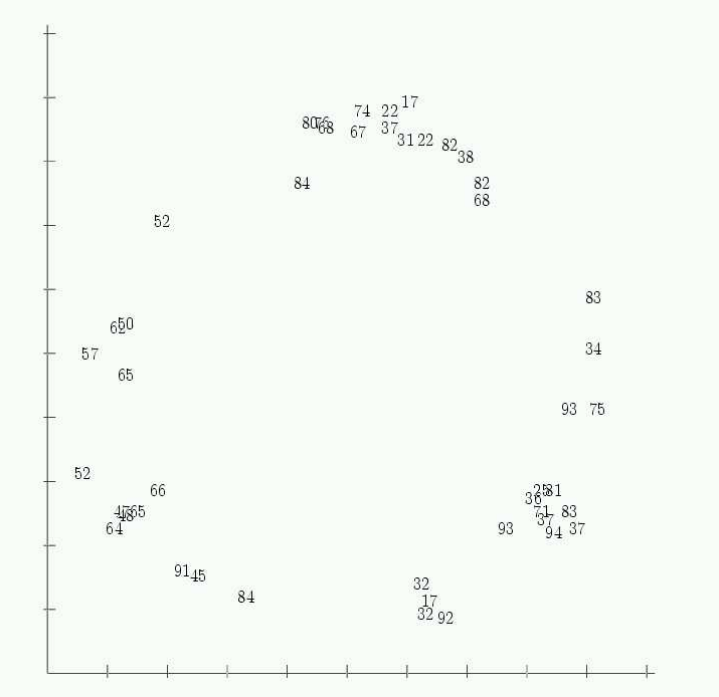
- hierarchisch zijn: voeg kleinere clusters samen tot grotere
- niet-hierarchisch zijn: verbeter bestaande clustering

Wij gebruiken AI-technieken als Self Organizing Maps (Kohonen's SOMs), gebaseerd op neurale netwerken.

Er zijn random aspecten — kansen spelen een rol.

Links zien we duidelijk verschil tussen winkels met en zonder restaurant. Recht gebruiken we postcode, en zien we "eetgewoontes" terug.





In een supermarkt zijn er *veel* klanten die een *grote* keuze hebben. Associatieregels kunnen eenvoudig worden toegepast.

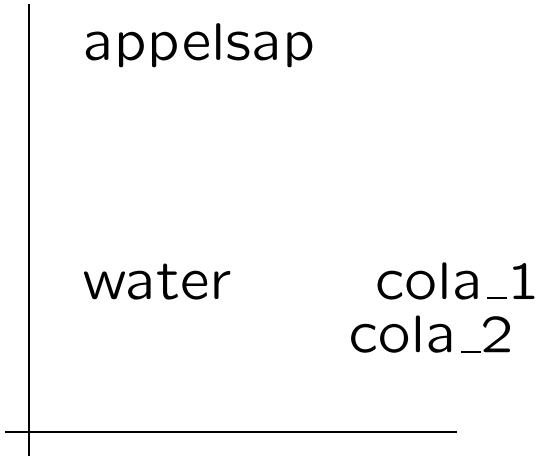
"Onderzoek" onthulde dat vloei en tabak vaak samen worden verkocht (wat een verrassing!), maar ook dat speciale shag speciale vloei vereist.

Er is veel interesse in hierarchieën: merken $\leftrightarrow$ algemene categorieën.

Het luiers—bier verhaal is een sprookje.

Een laatste toepassing in een supermarkt is het volgende. We willen dranken zodanig groeperen dat producten dicht bij elkaar komen als ze elkaar kunnen "vervangen".

Deze informatie haal je uit de boodschappenmandjes: zulke producten worden zelden samen verkocht. Technieken als boven leiden tot plaatjes als:



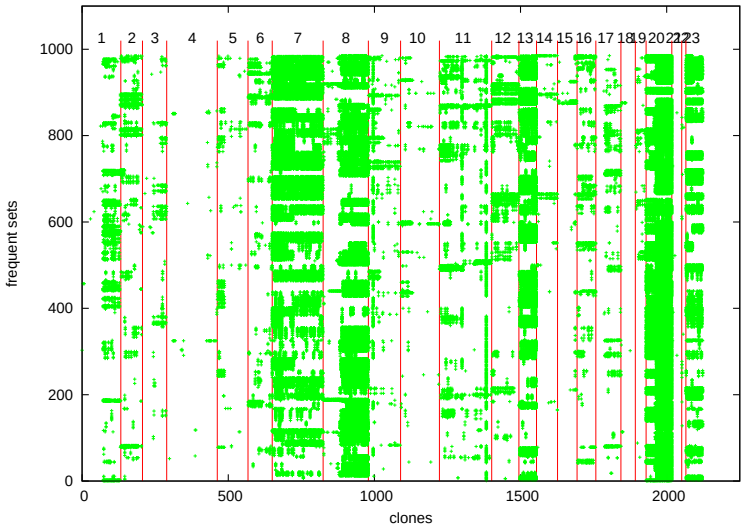
Bio-informatica

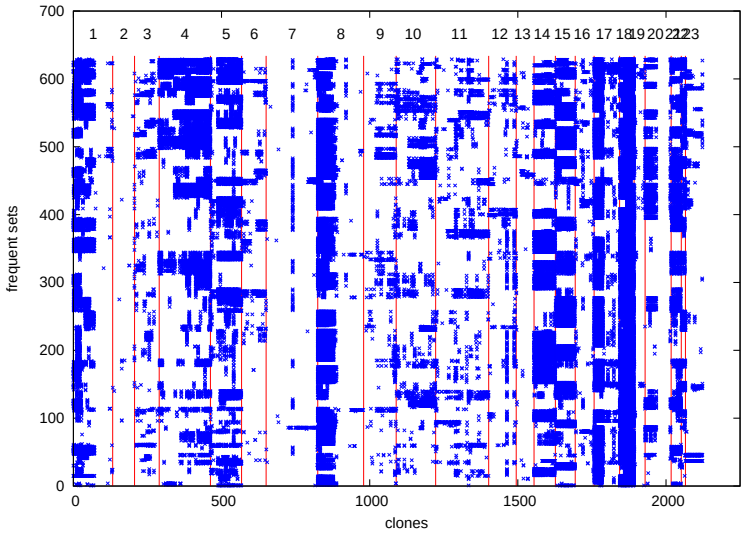
Veel vragen uit de Bio-informatica kunnen als Data mining problemen worden beschouwd. Een enkel voorbeeld.

DNA kan gezien worden als een rijtje letters uit het alfabet $\{A, C, G, T\}$: $AGGTCAAT \ldots TT$. Menselijk DNA bevat ongeveer 3.000.000.000 letters — het menselijk genoom, verdeeld over zo'n 20^+ chromosomen.

Stel dat we voor zo'n 2.000^+ speciale stukjes DNA (clones geheten), netjes verspreid over het menselijk genoom, weten hoeveel er aanwezig is in 150 patiënten. Hier is "hoeveel" een getal tussen -5 ("verlies") en $+5$ ("versterking"). Hoe "mine" je deze berg aan data?

Links: versterkingen (985 frequente "duos"'); rechts: verliezen (629 frequente "duos"'). Verticale lijnen zijn de chromosoomgrenzen. Drempelwaarde: 100.

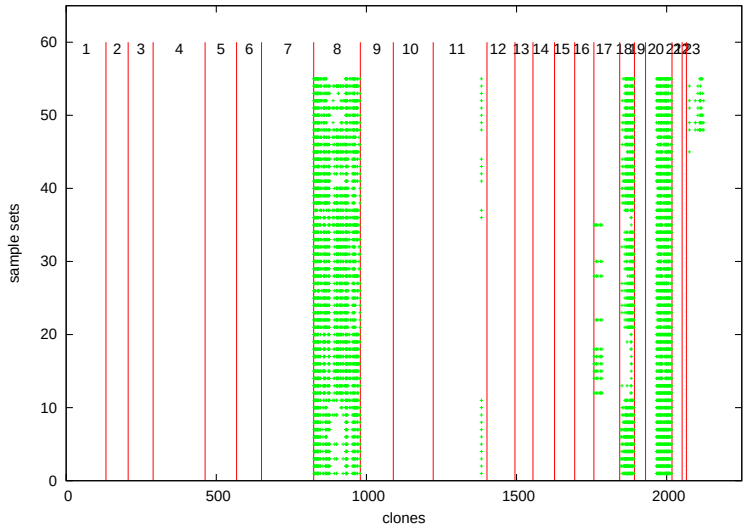


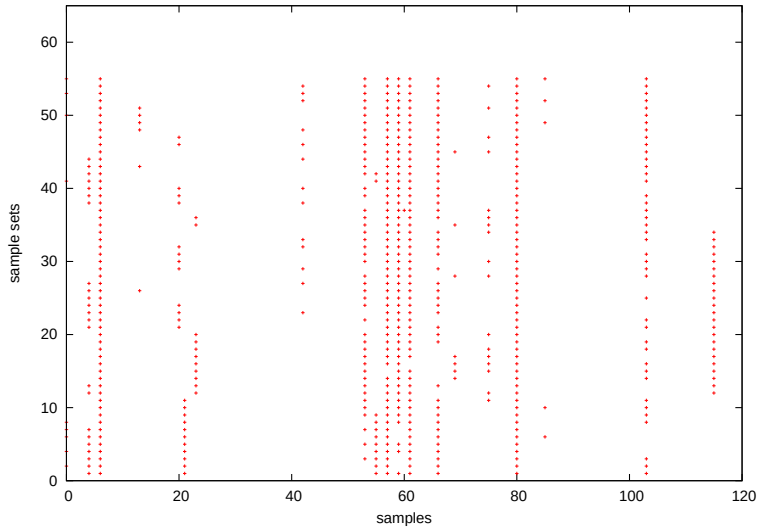


In zo'n plot zien we talrijke groepen patiënten, waarbij elke groep zich ongeveer hetzelfde gedraagt op de 2000^+ clones.

Een tweetal patiënten is frequent als beiden op minstens 100 dezelfde clones een versterking, respectievelijk een verlies hebben.

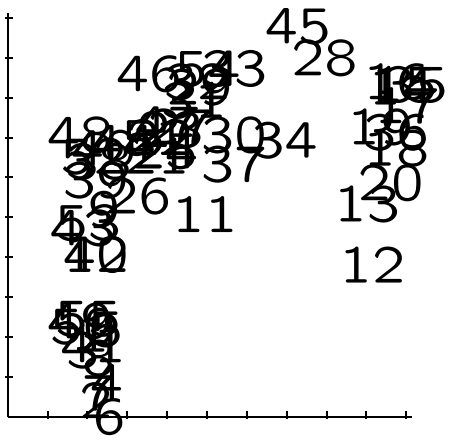
Links: de 55 “maximale” 10-itemsets (met steeds 10 patiënten), gecombineerde versterkingen en verliezen, met drempelwaarde 100; rechts: de patiënten uit die verzamelingen.





De verzamelingen hebben veel gemeen: goed of slecht?

Deze 55 verzamelingen kun je inbedden in het eenheidsvierkant:



De Euclidische afstand lijkt op de afstand die je krijgt door versterkingen en verliezen “kwadratisch op te tellen”.

En verder?

- beoordeel grenzen versterking/verlies (0.225? fuzzy?)
- maten voor interessantheid
- visualisaties (SOMs, . . .)
- verwerk “aaneengesloten zijn” in support
- voeg gegevens als leeftijd toe
- en tijdsafhankelijkheid
- tool

We zien succesvolle inzet van Data mining.

Aandachtspunten zijn onder meer:

- privacy
- presentatie van resultaten
- toepasbaarheid van mooie algoritmen
- grootschaligheid