Van AI tot Data mining

dr. Walter Kosters, Universiteit Leiden

Gouda — woensdag 17 oktober 2007

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:

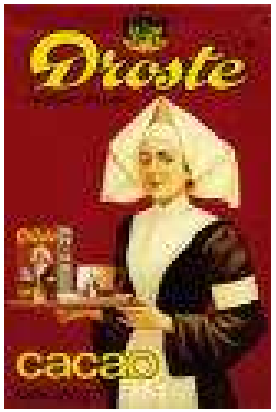
- Boodschappenmandjes: wat en hoe kopen we?
- Bio-informatica: DNA?

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:

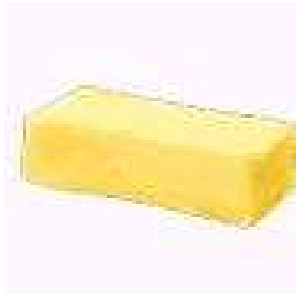
1. data selectie
2. opschonen: de-duplicatie en domein-consistentie
3. verrijking: data-fusie
4. coderen
5. Data mining — het echte gebeuren
6. rapporteren



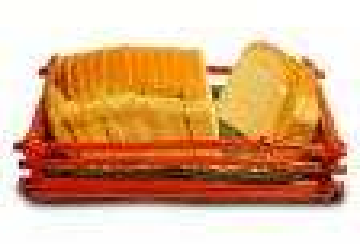
Er zijn *veel* Data mining technieken, of beter gezegd: algoritmen, die we kunnen gebruiken om data te "minen":

- (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

Associatieregels



$\Rightarrow$



Klanten en producten

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.

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: $\{2, 7\} \Rightarrow \{5\}$.

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.

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. En $100 \leftrightarrow 120$ verschilt van $1 \leftrightarrow 21$.

Hoe dan ook, er zijn veel mogelijkheden!

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.

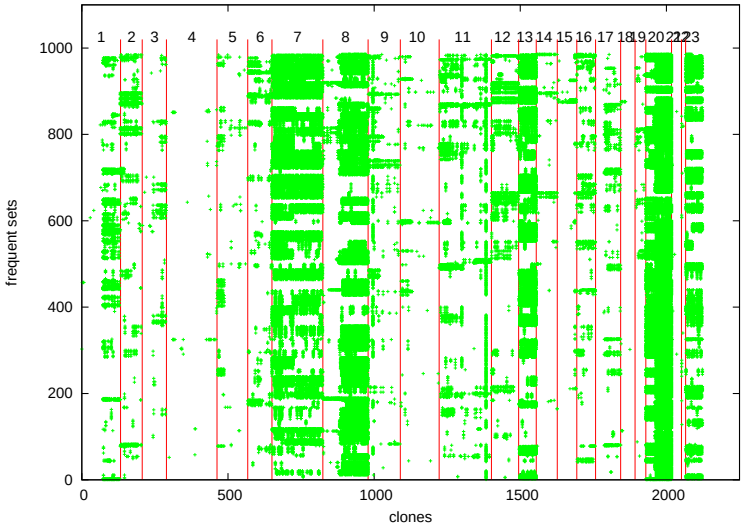
Bio-informatica

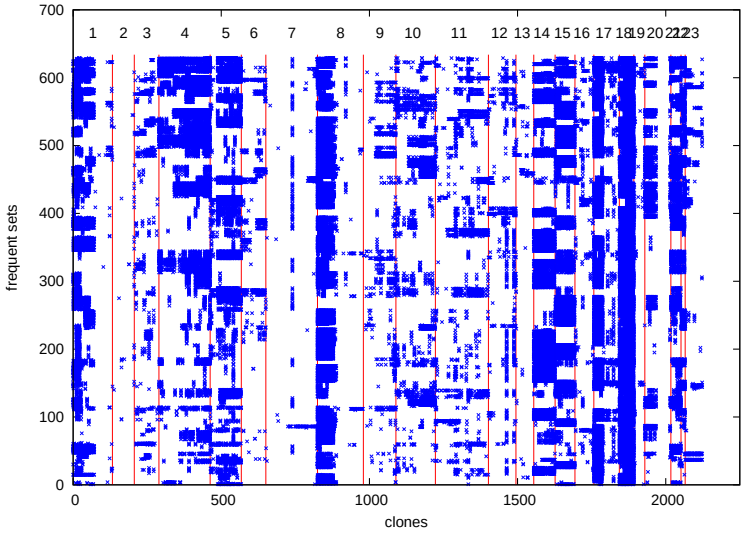
Veel vragen uit de Bio-informatica kunnen als Data mining problemen worden beschouwd. Een enkel voorbeeld.

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.





Slotopmerkingen

We zien succesvolle inzet van Data mining.

Aandachtspunten zijn onder meer:

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

Enkele recente afstudeeronderwerpen (bachelor, master):

⋈ ⋈ Poker; Rummy; Nurikabe; Sudoku; Netflix; Sokoban

⋈ A Model of Evolution Focussing on Neural Networks

⋈ Particle Swarm Optimization: Finding Optimal Poker Strategies

⋈ A Way to Improve the Accuracy of Mining Fused Data

⋈ Making the Right Offer to the Right Customer

⋈ Crime Data Mining

⋈ Unique Factors in the Human Genome

⋈ Automated Generation and Analysis of Logical Puzzles, The Flats Puzzle

⋈ Unravelling the Genetic Structure of NP-completeness

⋈ Computer game development for speech therapy support

⋈ ⋈ Liquid State Machines; Stambomen; Eten; Diagnoses