Data Mining – er der guld i virksomhedens data?

Af Karsten Boye Rasmussen

Resumé

Data mining omhandler nyere metoder til analyse af store mængder af virksomhedens data. I denne oversigtsartikel omtales flere af de udbredte metoder anvendt indenfor data mining. Speciel opmærksomhed henledes på objektet for data mining: virksomhedens data skabt i virksomhedens processer. Da stadig flere processer frembringer stadig flere data, producerer virksomheden en stigende strøm af data; specielt udløser anvendelsen af Internet en eksplosion af datamængden. Analysen af de store datamængder kræver i sig selv et højt stade af informationsteknologi. Samtidig medfører behovet for frembringelse af velegnede data til data mining, at etablering af et data warehouse i virksomheden bliver et krav for med sikkerhed at kunne forsyne virksomhedens data mining med integrerede og valide data.

Artiklen illustrerer, hvorledes data warehouse og data mining er elementer i virksomhedens aktive akkvisition af data/information/viden, aktive produktion og den aktive udbredelse af virksomhedens viden. Med anvendelse af data warehouse og data mining foretager virksomheden en bevægelse fra passiv opsamling og passiv udbredelse af viden – virksomheden som pulterrum for viden – til virksomheden som videnspumpe[1].

Data mining - begreb og historie

Den engelske metafor "data mining" skaber billeder af en eksotisk minedrift: guldgravning. Den ordrette metafor er således ikke korrekt. Det er ikke data der søges, men noget der kan sammenlignes med guld. Metaforen fremgår også af valget af firma- og produktnavne som SPSS Clementine[2] og Kdnuggets[3]. Specielt det sidste navn refererer til at finde store klumper af guld (viden) og fremhæver gennem denne type minedrift en individbaseret metode med anvendelse af kløgt - og måske også en del held. Heroverfor kunne stilles billeder af andre fremfindingsmetoder for samme værdifulde metal, fx udvaskning af guldstøv, som ville fremstille et mere møjsommeligt, men mekaniserbart arbejde. Billedet af det mere prosaiske og gentagne arbejde passer i virkeligheden langt bedre på processen omkring data mining, specielt understreges det særdeles omfattende forberedende arbejde fremfor heldet.

Det er karakteristisk at data mining processen omfatter et stort arbejde, fordi miningen bygger på analyse af store samlinger af data - ofte endda gigantiske mængder (Giga-bytes). Billedet på at finde nogle enkelte data i de meget store samlinger er holdbart for nogle tilgange, men i mange andre tilfælde angår miningen at der foretages en gruppering af datamængden. Data mining er som begreb noget nyere, men tæt knyttet til begrebet "Knowledge Discovery in Databases" og her

understreges størrelsen også ved at referere til "Very Large Databases"[4].

Størrelsen af datamængden er dog ofte vanskelig at illustrere, og på trods af understregningen af de meget store datamængder illustreres fundene oftest gennem enkeltobservationer eller kortvarige forløb (se eksempelvis Mitchell (1999) og Berry & Linoff (1997, s. 151)), fx det mønster at en bankkunde gentagne gange hæver større beløb på kontoen og derefter afslutter ved at lukke kontoen. Hvis lukningen kan forudsiges udfra mønsteret, burde banken reagere. Mere komplicerede mønstre kan fremfindes ved analyse af de store datamængder, men netop kompleksiteten gør det vanskeligt at illustrere. Omvendt forekommer nogle mønstre sjældent. Således kræves der (heldigvis) en stor mængde data for at kunne finde eksempler på forsikringssvindel. Data mining dækker også værktøj til at finde sådanne sjældne mønstre og udtage mistænkelige forhold til yderligere bearbejdning.

Objektet for data mining er store datamængder som samles i et data warehouse.

Data warehouse

Begrebet "data warehouse" – som her anvendes i den fordanskede form data varehus[5] – tillægges almindeligvis W.H. Inmon (1996). Data varehuset er en metode til at samle, validere, integrere, bevare og tilgængeliggøre virksomhedens information. Den statiske betragtning af data varehuset udfordres af andre anskuelser:

> *"Data Warehousing is a process of fulfilling Decision Support enterprise needs through the availability of information" (Welbrock, 1998).*

Her lægges vægten først på, at der er tale om en proces. Dernæst på at det er virksomhedens krav, der opfyldes. Denne betragtning hos Welbrock illustrerer, at data varehus processen ikke blot er en teknisk disciplin, men at teknikken er underlagt forretningens mål.

En overflod af data – brug og gem

Anskuelsen af en organisation som en informationsbearbejdende enhed implicerer, at organisationens objekt hovedsageligt er information eller tidligere i processen data. Organisationens oplysninger, spørgsmål og beslutninger bevæges fra arbejdssted til arbejdssted og nu som strømme af digitaliseret information. Kundeorienterede transaktioner er gennem flere årtier blevet registreret i virksomhedens informationssystemer som operative data, nu opbygges også spor af den interne behandling i virksomheden.

Den traditionelle før-elektroniske opbevaring har store fysiske omkostninger, fx vil der ved opbevaring af papirdokumenter (fx fakturaer, ordresedler etc.) skulle anvendes ganske store ressourcer i form af ringbind, arkivbokse, hyldemetre, reoler, rum og bygninger. Desuden er søgbarheden af et dokument ofte begrænset til en enkelt indgang (fx fakturanummer), som direkte er en nøgle for den fysiske placering. Store omkostninger sammenholdt med vanskelig anvendelse gør kassation særdeles attraktiv i det forrige århundredes traditionelle informationsverden. Men med virksomhedens automatiske registrering af elektroniske spor og opbevaring i digitaliseret form er omkostningerne ved bevaring særdeles små, og mulighederne for fremfinding næsten ubegrænset fleksible ved passende opbygning af registrene. Dermed bliver bevaring langt mere attraktiv end kassation (som i øvrigt også kan være fejlbehæftet, ved at noget bevaringsværdigt fejlagtigt kasseres). Omkostningen ved opbevaringen er hovedsageligt sikkerhedsproblemer, herunder bl.a. problemet med forsat at kunne tilgå informationen – "digital information lasts forever - or five years, whichever comes first!" (Rothenberg, 1995). Det er sjældent nedbrydning af mediet, der er problemet med digital information. Problemet findes i den teknologiske forældelse; når der ikke længere findes den gamle teknologi (fx 5 år gamle), der kan læse de gemte medier.

Mennesker har gennem århundreder bekymret sig over den stigende mængde af information i samfundet. Et nyere eksempel findes hos Brandi og Hildebrandt (2000), der citeres for at udtrykke at viden fordobles på 2 år, og at i 2020 vil fordoblingen være sket på bare 70 dage. Men der er forskel på om der med korte mellemrum dukker en ny Einstein op, og på at virksomhedens viden om kundernes adfærd automatregistreres i stigende omfang. Den sidste type viden angår kun virksomheden og kunden og forandrer ikke umiddelbart vores verdenssyn. Selvom lagringskapaciteten forøges, og databaserne forøges, så er den menneskelige administration af systemerne netop ikke lineært afhængig af antallet af registreringer. Der behøves ikke tre gange så mange ansatte for at administrere en vækst på 3 gange i lagringskapacitet (således som det fremsættes i en notits i Computerworld, 24. april 2001). Udskiftes en PC, er lagringskapaciteten for det meste fordoblet. Pladsen bliver hurtig brugt, men der skal alligevel ikke to personer til at betjene den nye maskine.

En yderligere grund til opbevaringen kan være, at de elektroniske registreringer er lovgivningsmæssigt fastsatte, fx opbevaring af regnskabsoplysninger. Desuden kan informationerne være nødvendige for fastlæggelse af ansvar. Manglende opfyldelse af kvalitetsmål for et produkt eller en ydelse vil gennem informationssystemerne kunne spores til leverandører og deres underleverandører. Virksomheden og dens kontakter har således en interesse i disse registreringer. Angår registreringen ansatte eller kunder, kan registreringen få karakter af overvågning. Det henleder opmærksomheden på, at handlingerne også har en etisk dimension. Men man burde snarere udtrykke, at den digitale registrering ikke fører til ændring af det forhold, at enhver handling også har en etisk dimension.

Den stigende opsamling af information, den øgede lagringskapacitet og tendensen til opbevaring snarere end kassation introducerer informationssamfundets opgør med den materielle samfunds fysiske "brug-og-smid-væk" mentalitet, der erstattes med en omfattende "gem-og-brug" af information. Udover den digitaliserede informations perfekte egenskaber mht. opbevaring og distribution, så har information som materiale et interessant værdiforløb. Når viden udbredes, opnår den en højere værdi, når viden slet ikke deles, er den uden social værdi. Virksomheden vil derfor på den ene side opbevare sin information og på den anden side have en stor interesse i at udnytte og udbrede denne information og skabe forretningsmæssig fordel gennem en vidensproduktion, der vanskeligt kan eftergøres.

Yderligere oplysninger fra Internet

Med fremkomsten af elektronisk handel over Internet er der sket en mangedobling af muligheden for at opsamle informationsspor. Med Internet-logning og "clickstream analysis" kan den enkelte forbrugers vej mod beslutningen om køb fastlægges. Internet-logning består primært af identifikation (oplysning om klientens navn eller IP (Internet Protocol) adresse), hvilket element eller side brugerens browser har bedt om, hvilken returkode denne transaktion havde (primært om det var succesfuldt, koden 200 betyder "OK"), samt antallet af bytes der returneredes (Kendall, 2000, s. 101).

Derudover kan en mere avanceret logning vise, hvorfra brugeren kom til denne side ("referrer tracking"). Logningen kan dermed for det første fastslå (for-)brugerens indgang til siden, og desuden kan den foretagne rute på virksomhedens eget web-sted opsamles. Hvis man mangler fantasi, vil man stråle over at kunne benytte disse oplysninger til en minutiøs fastlæggelse af den interne rute, der er gået forud for salget (det positive mønster). En mere kreativ tilgang vil indse at logningen også vil sætte os i stand til at analysere adfærden blandt de der ikke købte (det negative mønster). Med forbedringer baseret på den viden kan ikke-køberne blive til kommende kunder.

Klient/IP-adresse		Tidspunkt	Web-side eller element	Retur	Bytes
133.225.107.171	- -	[04/Jan/2001:06:29:24 -0700]	"GET /home HTTP/1.0"	301	330
133.225.107.171	- -	[04/Jan/2001:06:29:24 -0700]	"GET /home/ HTTP/1.0"	304	-
133.225.107.171	- -	[04/Jan/2001:06:29:32 -0700]	"GET /home/pubs.html HTTP/1.0"	200	1204
133.225.107.171	- -	[04/Jan/2001:06:29:37 -0700]	"GET /home/iq.html HTTP/1.0"	200	2516
133.225.107.171	- -	[04/Jan/2001:06:29:37 -0700]	"GET /home/getacro.gif HTTP/1.0"	304	-

For begge typer brugere kan vi gennem "cookies" genkende brugeren på web-stedet. En cookie er en lille tekst-fil der lægges på brugerens PC. Brugerens IP-adresse er sjældent en statisk oplysning; ved opkobling til en internetudbyder tildeles brugeren ofte dynamisk et IP-nummer. Den samme bruger kan således have ændret IP-adresse ved næste besøg. Men gennem anvendelsen af cookie-filen kan web-stedet fastslå og genkende brugerens[6] tidligere besøg på stedet, og oplysninger om brugeren i form af foretrukne sider og indkøbsvaner kan benyttes til at fastlægge et individualiseret web-sted (Laudon, 2000; Kimball, 2000). Dette illustreres fx i startbilledet fra Amazon.com, der indeholder meddelelsen:

> *"Hello Karsten, here are our recommandations for you"*

Personaliseringen kan også typisk finde sted gennem udvælgelse af de viste bannerreklamer ("target marketing"). Internetlogning er et bastant eksempel på, hvorledes der kan opsamles minutiøse oplysninger for hver enkelt bruger. Herigennem åbnes mulighed for personificering af behandlingen af kunden – og desuden må der overvejes etiske og sikkerhedsmæssige perspektiver omkring persondata.

Den løbende opsamling fra Internet vil bidrage med gigantiske mængder af data til virksomheden.

Validering af data samlinger

Data varehuset udtrækker oplysninger fra virksomhedens operationelle data. Hermed menes både det traditionelle transaktionssystem, men også de oplysninger der indsamles fra Internet. Da mange informationssystemer i virksomheden er utilstrækkeligt integrerede, vil en efterfølgende integration af oplysninger kræve et omfattende valideringsarbejde. Valideringen vil bestå i en omfattende liste af procedurer der må gennemløbes; spændende fra "skrubning" af data (Welbroch, 1998, s. 155) hvor der sikres konsistens - fx således at samme felttype (fx køn) angives på ensartet vis (fx som "M" og "K") i samtlige tabeller - til mere komplicerede afgørelser, hvor der gennem inddragelse af adskillige felter i flere tabeller sikres, at samling af oplysninger angår samme individ (hvor der kan have været uens registreringer), og til etablering af entydighed blandt modstridende oplysninger (fx en adresse). Foretages aggregeringer er det afgørende, at de aggregerede oplysninger valideres mod detailoplysningerne. Valideringerne vil ofte være meget omfattende og langt mere tidskrævende end anvendelsen af de senere forskellige analysemetoder (Weis & Indurkhya, 1998, s. 5).

Valideringsarbejdet skaber intern konsistens i data varehuset. Populariserede fremstillinger kalder dette etablering af "En Sandhed". De totalitære overtoner i udtrykket medfører hos forfatteren her en under-

stregning af, at konsistens kan opnås, uden at der er tale om sandhed.

Forekomsten af manglende data udgør et omfattende problem ved data mining. Med inddragelsen af et stort antal variable vil forekomsten af records med oplysninger i samtlige variable blive stadig mindre. Udelukkelse af records med manglende data er derfor uholdbart og det bliver nødvendigt at beslutte og behandle forskellige typer af manglende data[7].

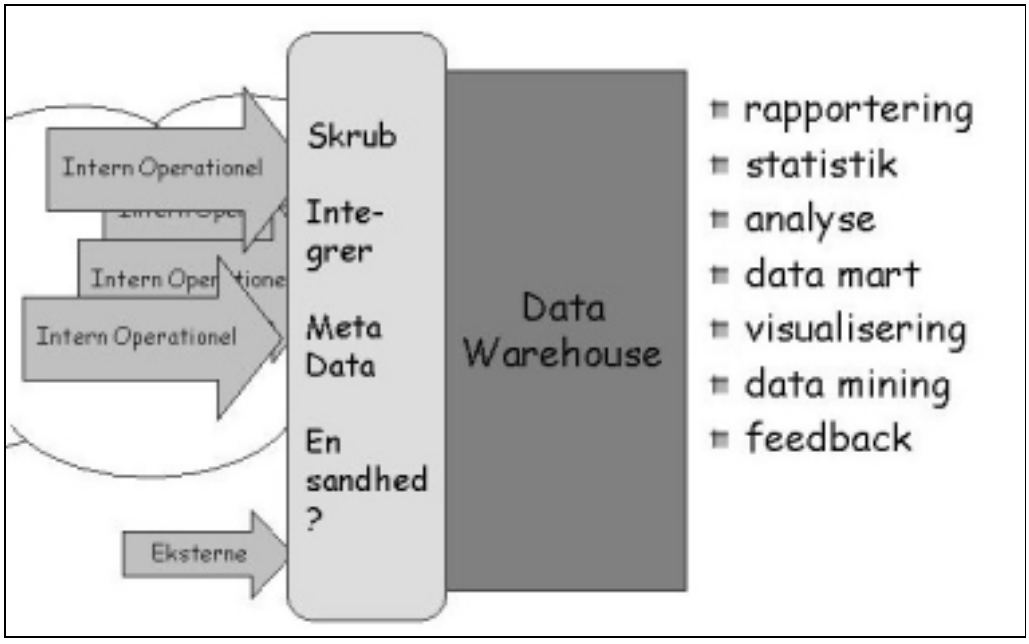
Inddragelse af historiske data

Det er kendetegnende for data varehuset, at det er et blivende opbevaringssted for data. Mens de operationelle systemers arkitektur primært vil være optimeret med henblik på hurtige svartider - "performance" (Kimball, 1996, s.2) - og derfor indrettes til kun at angå og opbevare direkte produktionsrelevante data, vil data varehuset være optimeret med henblik på at kunne forsyne beslutningssystemerne med de relevante data.

En forespørgsel til det operationelle system kan fx være: "Har Hansen betalt sin faktura nr. 794837-2?". På et eller andet tidspunkt er svaret antagelig "Ja". Det operationelle system vil dermed levere forskelligt svar til forskellige tidspunkter; det operationelle system producerer et øjebliksbillede af et system i stadig bevægelse. Heroverfor kan en forespørgsel til et data varehus være: "Hvor stor en andel fakturaer var ubetalte i mere end 30 dage i 1999?". Svaret bør være uforanderligt – i betydningen uafhængigt af tidspunktet det stilles på - og det naturlige følgespørgsmål "Og hvordan så i 1998?" illustrerer, at data varehuset rummer en tidsserie og dermed omfatter historiske data.

Tilgængeliggørelse – metadata

Formålet med opsamling og integration af virksomhedens data er ikke at skabe en ressource for senere historieskrivning, men at foretage en umiddelbar anvendelse. For at kunne anvende de store datasamlinger optimalt, må data ligeledes dokumenteres optimalt, og dokumentationen kan selv betragtes som data, altså metadata (Rasmussen, 2000). Den næsten øjeblikkeli-



ge anvendelse sikrer mod ophobning af udokumenterede data. Frem for at bygge midlertidige uformelle "underground" datasamlinger er data varehuset en kvalitetssikrende proces der forhindrer, at metadata blot findes i hovedet på enkeltindivider. I stedet gøres metadata direkte tilgængelige, og dermed kan data udnyttes fleksibelt (Welbrock, 1996, s. 12).

Det er kendetegnende, at data i varehuset "frigives" til brug. Ændringer i eksisterende data – data der en gang er placeret og frigivet i data varehuset - bør i princippet ikke forekomme. Det er et krav, at data i data varehuset er gennemvaliderede, førend de annonceres som tilgængelige, og data er først tilgængelige, når de ligeledes er gennemdokumenterede med metadata.

Anvendelse af data varehuset – og ny viden

Adgang til data og dermed udbredelse af data er ikke en tilstrækkelig betingelse for en positiv anvendelse af data. Data varehuset er samlingen af data, mens udbredelsen og anvendelsen af data sker gennem diverse applikationer i virksomheden. Det er gennem applikationerne, at data analyseres, præsenteres og stilles overfor virksomhedsrelevante processer. Hermed opnås ny viden i virksomheden, typisk ved at ikke tidligere sammenførte data relateres, eller ved at data anvendes på nye områder og analyseres gennem nye metoder.

Det bør i denne sammenhæng understreges, at data varehuset i vid udstrækning også støtter den rutinemæssige vidensproduktion, der finder sted i virksomheden fx i form af standardrapporteringer. Her kan være tale om særdeles omfattende og krævende statistikker og præsentationer evt. publiceret på virksomhedens Intranet. Omfattende applikationer vil helt kunne automatisere disse virksomhedscentrale rutineprodukter.

Blandt de typer af applikationer, hvormed der produceres ny viden, findes området med data mining. Data mining udgør altså som tidligere nævnt kun en delmængde af de metoder, der udnytter data varehuset og stadig kun en delmængde indenfor virksomhedens produktionen af ny viden. Men er data mining så kun en ny etiket?

Metoder indenfor data mining

Hvad er det, der adskiller data mining fra anden analyse? I Berry og Linoffs udbredte værk fra 1997: "Data Mining Techniques" benyttes følgende definition:

> *"Data mining ... is the exploration and analysis, by automatic or semiautomatic means, of large quantities of data in order to discover meaningful patterns and rules". (Berry & Linoff, 1997, s. 5).*

Denne rummelige definition inkluderer både eksplorative og beskrivende metoder samt egentlig analyse. Metoderne anvender ikke overraskende informationsteknologi (automatisering). Der er tale om store mængder af data (men er 1.000 stort eller er 10.000 stort?), og forfatterne henleder opmærksomheden på, at en stor datasamling i sig selv ikke er tilstrækkelig for at levere viden (Berry & Linoff, 1997, s. 3). Yderligere afdækkes mønstre og regler, der er meningsfulde. Et stort antal metoder kan indeholdes i denne definition, men i denne fremstilling medtages kun nogle få af disse metoder. Kriteriet for den her foretagne udvælgelse af metoder indenfor data mining har været, at metoden er almindelig udbredt og solidt understøttet af software.

Hos Berry og Linoff stilles hypotese-tests overfor "knowledge discovery" som hhv. top-down og bottom-up. Denne bottom-up approach ses andre steder beskrevet som, at data mining snarere er data drevet end drevet af brugeren og opstilling af kriterier for falsifikation (Mena, 1999, s. 50). At metoderne i højere grad er data drevne implicerer, at der anvendes en automatiseret metode som efter nogle gennemløb af data præsenterer en model, der kun i beskedent omfang er fastlagt af en person.

Altomfattende automatiserede metoder har ofte problemer med meningsfuldheden.

En faktoranalyse kan opstilles overfor et passende problem, og der kan metodisk konstrueres tilfredsstillende faktorer – men stadigvæk er "dåben" af faktorer en vanskelige human kunst. Det er ikke automatiserbart at fortolke en vektor af ladninger og sætte begreb på den bagvedliggende faktor (Harman, 1967, s. 133).

Meningsfuldheden bør tillægges vægt når metoderne evalueres. Berry og Linoff er også opmærksomme på, at der hører aktion til meningsfuldheden:

> *".. merely finding the patterns is not enough. You must be able to respond to the patterns, to act on them, ultimately turning the data into information, the information into action, and the action into value" (Berry & Linoff, 1997, s. 18).*

Anvendelsesområder for data mining

Blandt de problemer der løses gennem anvendelse af data mining, er situationer, hvor det er vanskeligt efter den centrale begivenhed at opnå valide data. Problemet kan fx være, at kunder forlader virksomheden. En forudsætning for denne analyse er, at det er muligt at fastslå, hvorvidt kunderne er loyale, dvs. at kunderne nødvendigvis er registrerede. Eksempler kan være telefonselskaber, bankvæsen, avisabonnement etc. Dvs. tilfælde hvor der udsendes regelmæssige regninger. Men i stadigt stigende omfang registreres også andre firmatypers private kunder, oftest gennem medlemskab, og ved elektronisk handel er registreringen oftest en betingelse for, at varen vil kunne leveres, idet både betaling (via kreditkort) og modtagelse (angivelse af adresse) skal fastlægges, førend levering finder sted.

Ved "churn" forstås, at kunder afgår fra et teleselskab – og antagelig skifter til et andet. Problemet for en efterfølgende undersøgelse er, at de frafaldne kunder risikerer at have en kraftig bias. De afgåede kunder har ringe interesse i at bruge tid på at udfylde spørgeskemaer, blive interviewet, eller hvilke efterfølgende dataindsamlingsmetoder man måtte anvende. I stedet kan virksomheden så vælge at anvende data, der snarere er generelt og objektivt observeret frem for indhentet med et instrument konstrueret til formålet. Disse data beskriver kundens handlinger frem til afgangen; i dette eksempel fx længde og hyppighed af samtaler. Specielt indenfor mobiltelefoni ses rabatter ved anskaffelse af telefoner og oprettelse som abonnent, der gør det særdeles attraktivt for kunden at skifte selskab og særdeles vanskeligt for firmaet at opnå nogen profit på den enkelte kunde indenfor det første år[8]. Hvis selskabet kan "forudse" at en kunde er på vej til at "churne", kan virksomheden forsøge at fastholde kunden gennem en godt tilbud.

Et andet vanskeligt undersøgelsesområde er bedrageri. Eksempler findes her typisk i forbindelse med forsikringssager. Her søger man gennem data mining at finde mønstre, der afviger fra det almindelige, og som af denne grund påkalder sig ekstra opmærksomhed ("exception reporting").

Læring: træning, validering, scoring

Den data-drevne, bottom-up "knowledge discovery" opdeles af Berry og Linoff (1997, s. 6) i en dirrigeret og en ikke-dirigeret type. Den dirigerede type forsøger at forklare et af felterne i data (fx om der fandt et salg sted), mens den ikke-dirigerede type er yderligere eksplorativ mht. at finde mønstre eller sammenhænge mellem flere felter i data. Der er dog ingen helt fastslået enighed om terminologien indenfor området, således opdeler Weiss & Indurkhya (1998, s. 7) data mining i "prediction" og "knowledge discovery".

Mens data mining og "directed knowledge discovery" er virksomhedstermer for anvendelsen af specielle metoder, benyttes indenfor datalogien begrebet "maskinel læring" (Mitchell, 1997). Man kan tale om dirigeret læring, når et udfald, et mål ("goal") eller afhængig variabel forudsiges på grundlag af andre variable indgående i en model opstået (indlært) gennem analyse

	Faktisk salg (positiv)	Faktisk ikke-salg (negativ)
Forudsagt salg (positiv)	sand positiv	falsk positiv
Forudsagt ikke-salg (negativ)	falsk negativ	sand negativ

af rækkevis af data. Weiss & Indurkhya (1998) betoner forudsigelseselementet i titlen på deres bog "Prediktive Data Mining". Ofte vil målet, der ønskes forudsagt, være binært som i eksemplet ovenfor (salg: ja/nej), men det kan også være numerisk (fx værdien af det pågældende salg). Indlæringsaspektet består i, at den opbyggede model gennem træning forbedres mht. at forudsige målet; det maskinelle består i at træningen, foretages af en computer.

Data varehuset består af flere indbyrdes relaterede tabeller i en database, men med henblik på undersøgelse gennem data mining ønskes udvalgte data omformet til en rektangulær eller flad fil (en tabel som i et regneark). Filen, der ønskes undersøgt gennem data mining, er altså ikke nødvendigvis på forhånd fastlagt i data varehuset, hvilket understreger den fleksibilitet, der ønskes af data varehuset. Til gengæld vil filen for data mining være omfattende, både hvad angår antallet af records eller poster og ofte ligeledes mht. antallet af medtagne variable. Det første sikrer at modeller vil kunne bestemmes med større sikkerhed, det sidste at der kan undersøges et meget stort antal modeller. I modsætning til hypotesetests foretages der altså ingen omfattende forudgående teoretisk begrundet udvælgelse, hvorfor data mining altid må betragtes som eksplorativ.

Når der analyseres historiske data, kendes også udfaldet. Hver record beskriver en situation eller et forløb afsluttende med beskrivelse af målet. Blev resultatet af en katalogudsendelse et salg eller ej? Målet behøver naturligvis ikke være noget ønskemål – fx kan målet være om der er foregået svindel – men information om målet må findes i data. En variabel må angive, om der var tale om svindel eller ej. Modellen foretager en klassifikation indenfor et defineret udfaldsrum (her binært).

Modellen anvender en opsplitning blandt records – oftest gennem en tilfældig udvælgelse. Der opdeles i en del der anvendes til at træne og opbygge modellen, en anden der validerer og justerer modellen, og endelig en tredje del hvor modellen testes. Behovet for store datamængder er blandt andet begrundet i, at data opdeles i flere adskilte samlinger. I samtlige datadele kendes det faktiske udfald, og dermed fås gennem testen et udtryk for modellens forudsigelsesgrad. Der er tale om en særdeles praktisk validering - "the proof of the pudding is in the eating". Modellen evalueres ved forudsigelse af udfaldene i testdatasættet. Modellens forudsigelse og de virkelige udfald sammenholdes gennem en konfusionsmatrix[9]:

Denne type opstilling anvendes også for medicinske tests, hvor gruppen af negative (som ikke har sygdommen) lykkeligvis er meget stor. Men det implicerer, at såfremt kun 100 ud af 10.000 er positive, vil der kun begås 100 fejl (falsk negative) ved at hævde at alle er negative. For at modellen kan siges at være bedre end tilfældet, må færre end 100 rubriceres fejlagtigt (eksempel fra Weiss & Indurkhya). En ukritisk udregning af fejlprocenten (1 pct.) kunne umiddelbart give indtryk af, at modellen er acceptabel. Ved sådanne skævt fordelte materialer foretages ofte indledningsvist en stratificeret sampling, således at samtlige blandt en lille gruppe få positive medtages, mens der foretages en tilfældig udvælgelse blandt den store mængde negative.

Med træningsdata opstilles modellen,

som bliver mindre og mindre fejlfyldt, jo mere kompliceret og omfattende modellen tillades at være (gennem anvendelse af længere computertid og flere iterationer). Modellen kan sammenlignes med at tilpasse en kurve til nogle observationer. Med "overfitting" menes, at modellen næsten med fuldkommenhed kan reproducere træningsdatasættet, men det er sket, fordi hvert et lille bump på kurven – enhver særhed i data - er blevet indlært. Modellen er blevet perfekt dresseret til netop denne situation. Når modellen afprøves mod et valideringsdatasæt viser det sig, at fejlprocenten for forudsigelsen igen stiger, efter at et optimalt punkt er passeret. (Berthold & Hand, 1999, s. 236).

Af det grafisk illustrerede eksempel fremgår, at den optimale model opnås ved iteration 4, hvor valideringsdata forudsiges med en fejlprocent på ca. 18. Derefter er også valideringsdata opbrugt, og modellen testes mod den tredje datadel. Efter at modellen således er fastlagt, benyttes den til at foretage "scoring" af nye data, hvor målet ikke kendes. Dermed kan der gættes på, hvorledes den pågældende gruppe eller evt. enkeltperson (fx ved beregning af kreditværdighed) vil reagere.

Data mining metoder

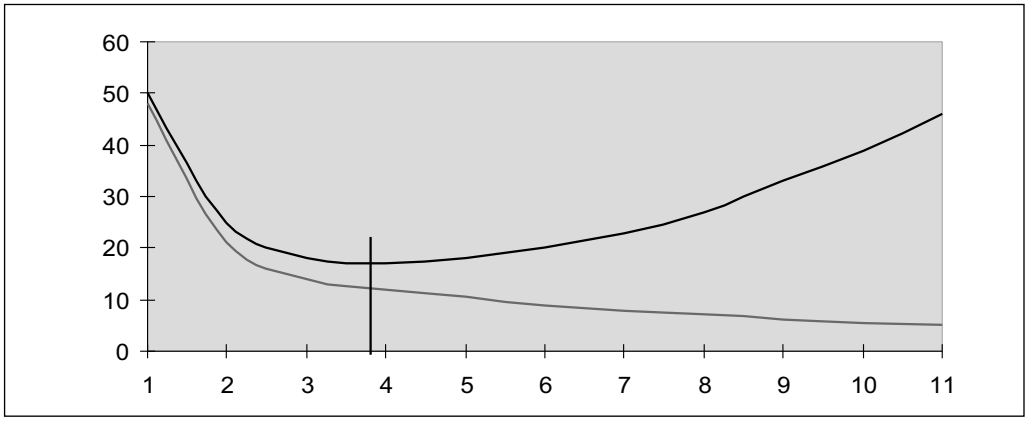
I det følgende omtales nogle af de mest udbredte data mining metoder.

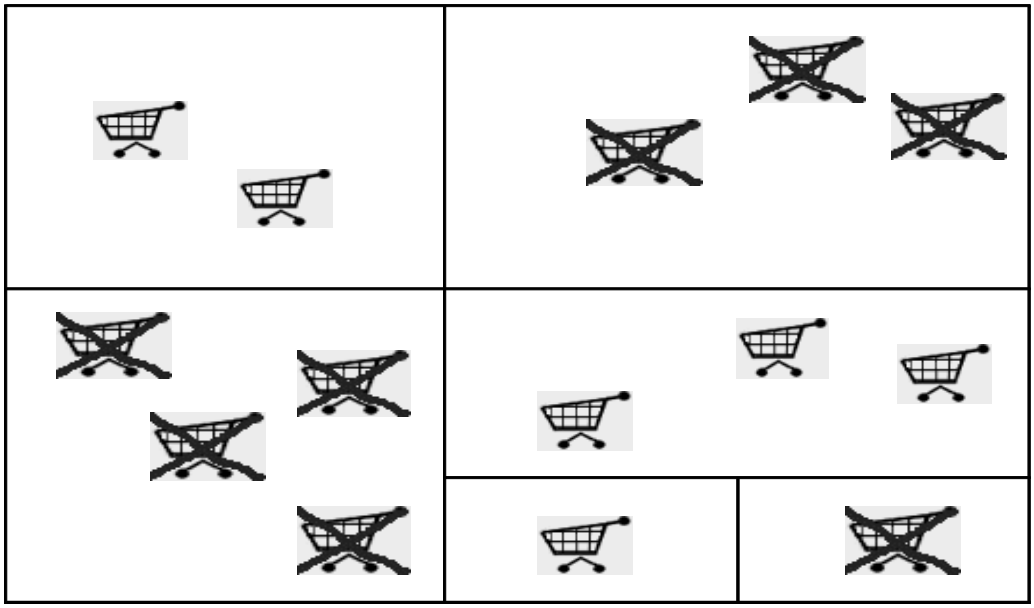
Indledningsvist bør man ved sammenligning med almindeligt anerkendte analysemetoder være opmærksom på, at data mining har andre kvaliteter. Analyse følger oftest et mønster af hypotese – undersøgelsesdesign – operationalisering - indhentning af data – hypotesetest. Heroverfor står data mining som anvender af eksisterende observationelle data uden indhentningsmæssig bias, og data i langt større målestok end der ellers ville kunne indhentes. Imod data mining bør bemærkes, at en ukritisk accept af de forhåndenværende data kan overse helt centrale og betydende variable, som blot ikke findes i data. På den anden side ses den store målestok også at have indvirkning: "hypotheses that are excellent approximations may be rejected in large samples" (Glymour et al., 1997), derfor bemærker forfatterne, at tildeling af score i data mining vil være mere attraktiv end egentlige tests.

Ved data mining er der mindre fokus på selve modellen - og dermed den bagvedliggende forklaring – og mere koncentration om en pragmatisk udvælgelse af den bedste metode til videre anvendelse.

Beslutningstræ

Et beslutningstræ foretager en opdeling af data gennem en serie af logiske spørgsmål. Oftest er der tale om spørgsmål med binær svarmulighed (Ja eller Nej). "Over 25.000 kr.





indestående på kontoen?". Hvis svaret er "Ja" stilles et nyt spørgsmål, hvis svaret er "Nej" stilles et andet. Dermed breder beslutningstræet sig[10] ligesom i legen "20 spørgsmål til professoren" (Berry & Linoff, 1997, s. 244). Ved hjælp af logisk kombination af spørgsmålene (konjungerende kombination med "og") kan spørgsmålene samles til et enkelt logisk udtryk som let lader sig fremstille som en betingelsessætning fx i databasesproget SQL:

```
SELECT * from MyData where ( Indestående > 25000) AND (KundeAntalÅr < 4)
```

Dette er et eksempel på en velkendt anvendelse af beslutningstræets logiske opdeling. Det nyere ligger i at computeren anvendes til at beregne, hvorledes opdelinger bedst kan foretages i materialer. Siden 1960'erne er der udviklet og anvendt en række algoritmer[11], der alle hviler på at foretage de bedste opdelinger (Mitchell, 1997, s. 55), dvs. opdelinger hvor variationen indenfor gruppen minimeres, mens variationen mellem grupperne maksimeres.

Et simpelt konstrueret eksempel illustrerer opsplitningen med beslutningstræ. Her antages at kunderne er karakteriseret ved to dimensioner i form af kontinuerte variable (normalt vil en langt større mængde variable benyttes). Dernæst er afsat de succesfulde indkøb og indkøbsture, der ikke resulterede i salg.

Som illustreret kan der foretages opsplitning af hele udfaldsrummet i grupper med ens adfærd ved hjælp af et antal regler. Men hvis antallet af regler accepteres til at være meget stort, betyder det, at modellen i stigende grad beskriver tilfældige særheder ved læringsdatasættet. I eksemplet ovenfor er opdelingen med de enkelte indkøbsvogne i nederste højre hjørne et eksempel på en regelbygning, der antagelig blot angår disse træningsdata, idet der ikke er tale om grupper, men om regler for enkelte individer. Ved anvendelse af reglerne på valideringsdatasættet vil de komplicerede regler ikke resultere i nogen overbevisende konfusionsmatrix.

Beslutningstræets klare fordel består i, at træet og dets forgreninger er umiddelbart forståelige, og at beslutningstræet direkte kan omsættes til handlen, fx opdeling i

kundegrupper. Desuden kan beslutningstræet med smidighed behandle både kategoriale såvel som kontinuerte variable.

Man kan også betragte beslutningstræets udsagn som udtryk for regler indenfor et system af kunstig intelligens eller ekspertsystem. Hvis eksperter har udformet reglerne, kan man tale om en top-down approach. Mens hvis udsagnene er fremkommet gennem data mining, vil det være korrekt at betegne metoden som data-drevet eller bottom-up. Interessant ved en virksomhedsbetragtning er, at udsagnene vil kunne afsløre, at der i praksis følges nogle uautoriserede eller uhensigtsmæssige forretningsregler.

Metoden med beslutningstræ er en fast bestanddel af den software, der udbydes som data mining[12].

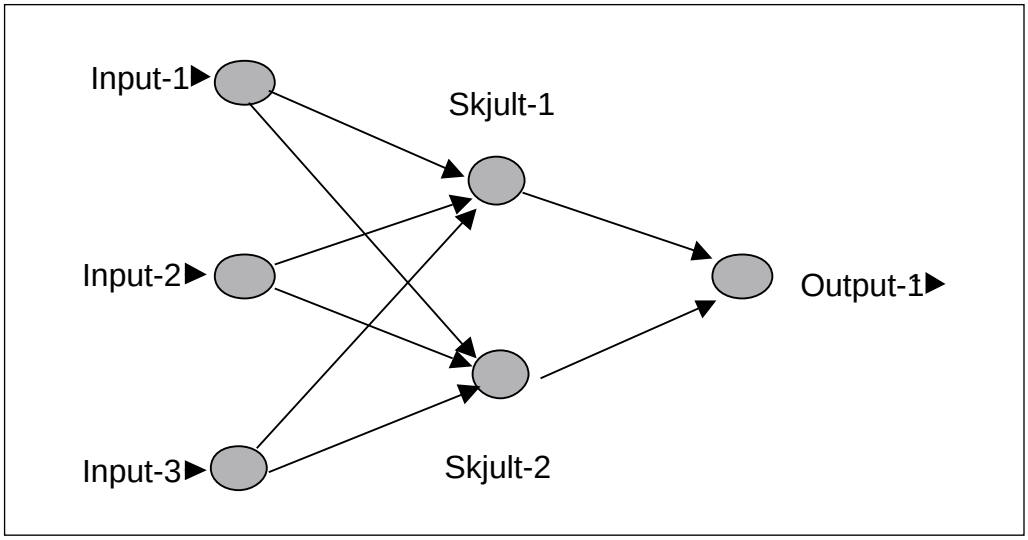
Kunstigt neuralt netværk

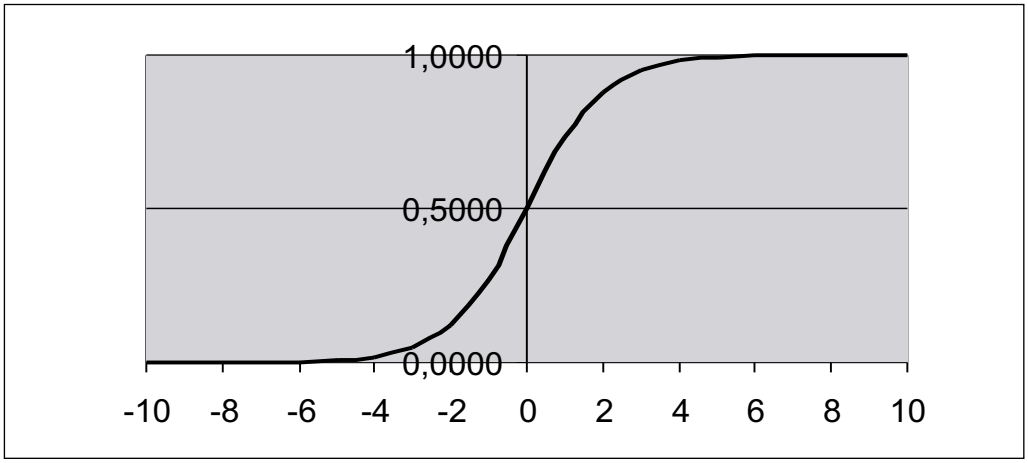
Et neuralt netværk er en model, hvor samtlige inputs er forbundne og gennem kombination transformeres til et output. Almindeligvis er både input og output skaleret til mellem 0 og 1.

Denne grafik (en lignende forekommer hos Berry & Linoff, 1997, s. 296) illustrerer hvorledes modellen bygger på et netværk af neuroner med input(s) og output (der kan forekomme flere outputs i modellen). Ligesom neuroner kan "fyre af" til andre neuroner - ved at udsende en impuls, overføre gennem en synapse og modtage gennem en dendrit - vil det kunstige netværk skulle overstige en tærskel for at påvirke efterfølgende celler. I modellen indskydes som vist yderligere neuroner i et skjult lag ("hidden layer").

Gennem oplæring fastsættes vægte for hvert enkelt input. Ved at flere inputs samles, kan der både være tale om, at en lille ændring i en input-kilde bevirker en ændring – overskridelse af en tærskel for neuronen – og at en større ændring i input ingen ændring medfører. Overskrides tærsklen for den efterfølgende neuron (fx "Skjult-1"), vil denne "fyre af" og sende videre. Transformationsfunktionen illustreres gennem den anvendte logistiske funktion (Sigmoid-kurve) i den følgende graf. Output for Sigmoid funktionen ligger mellem 0 og 1, stort negativt input og stort positiv input vil have værdier på hhv. 0 og 1. Indenfor et ret snævert område omkring 0 skifter funktionen, men funktionen er her tilnærmet liniær.

Et neuralt netværk vil ofte have et meget





stort antal input-kilder. Det store antal inputs samt anvendelsen af et skjult lag af neuroner – som ligeledes vægtes – betyder, at modellen er særdeles vanskelig at fortolke. Modellen er i praksis en black-box. Berry og Linoff udtrykker flot, at kunstigt neuralt netværk kan anvendes, når resultaterne er vigtigere end forklaringen (2000, s. 128 og 287). Den rene pragmatik kan være tiltrækkende, men der er også praktiske grunde til forsigtighed i anvendelsen af kunstige neurale netværk.

Såfremt modellen ikke kan fortolkes, er det usikkert indenfor hvilke rammer modellen er holdbar. Der kan ikke etableres advarselssystemer for, hvornår der bør skiftes model. Det bedste råd vil derfor være, at modeller med kunstigt neurale netværk bør oplæres og valideres med meget korte mellemrum. Der kan desuden være lovgivningsmæssige restriktioner overfor anvendelsen af sådanne black-box eller orakelagtige metoder. I USA er der således også begrundelsespligt for private virksomheder som fx banker i forbindelse med at yde lån. Det er ikke tilstrækkeligt at meddele: "Min computer fortæller mig, at du ikke kan få et lån. Ha' en god dag!".

Regression

Med regression vendes der tilbage til den fastere grund hvad angår modellers opbygningen og fortolkningen af modeller. Udgangspunktet er liniær regression, hvor målet m forklares med et antal kendte variable:

$$m = a + b_1X_1 + b_2X_2 + b_3X_3 + ...$$

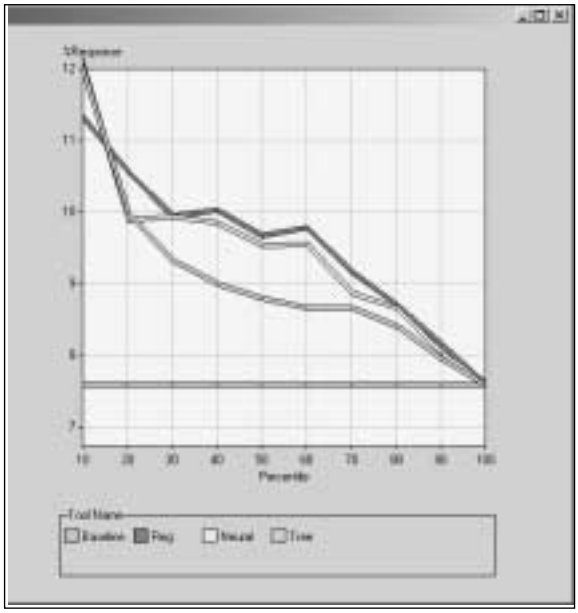
Men i data mining sammenhæng anvendes regressionsmodeller på samme pragmatiske vis som det neurale netværk. Selvom den statiske forklaringskraft af en model kan beregnes til at være temmelig ringe, kan anvendelseskraften være tilstrækkelig til, at modellen er interessant i en forretningsmæssig sammenhæng. Mens Berry og Linoff (1997) blot nævner regression og ikke betragter metoden som en egentlig data mining teknik, så anvendes metoden regression (specielt som logistisk regression) i SAS-pakkens Enterprise Miner.

Regressionsmetoden udmærker sig ved at modellens anvendelse af input-variablene kan fortolkes gennem vægtene. Desuden er der forbindelse til andre metoder; således vil den logistiske regression være et specialtilfælde af det neurale netværk uden noget mellemliggende lag.

Udvælgelse af metoder

De ovenfornævnte tre metoder: beslutningstræ, neuralt netværk og regression vil alle tre kunne anvendes til forudsigelse.

Udgangspunktet kunne være en postal udsendelse af et katalog, hvorfra kunder har bestilt varer. Virksomheden vedligeholder et kunderegister indeholdende information om, hvilke kunder der ved forrige udsendelse bestilte varer fra kataloget. Hvis købet kan forklares gennem de tilstedeværende variable, vil det være interessant for virksomheden at opnå adgang til et større register med tilsvarende oplysninger. Omkostningerne pr. brev/tryksag er faste, men kan der foretages en udvælgelse af de personer der mest sandsynligt vil reagere på forsendelsen, kan omkostningerne reduceres. I eksemplet her svarer gennemsnitligt godt 7 pct. positivt. Beregningerne vil endvidere kunne inkludere startomkostninger og en størrelsesangivelse for gevinsten ved hver positiv tilbagemelding. Dermed vil der gennem data mining software direkte kunne vises, hvilke grupper der bør udsendes til, og hvor stor en gruppe der i alt bør udsendes til. Udvælges en gruppe med større end gennemsnitlig sandsynlighed for at blive kunde har modellen skaffet et "lift"; hver af modellerne har da lift-kurver[13] der kan sammenlignes:



I eksemplet nedenfor ses, at beslutningstræet ("Tree") opnår det bedste lift for de første 10 pct., men skal virksomheden fx have kunder blandt 50 pct. af registeret, bør i stedet anvendes regressionsmodellen ("Reg") da denne ligger lidt over det neurale netværk ("Neural"). Bemærk at fordelen naturligvis aftager, hvis hele registeret udnyttes. Uanset model – selv uden model (illustreret ved "Baseline") – vides, at ca. 7,7 pct. af registeret vil blive kunder, såfremt de modtager kataloget.

Klassifikation og kvalitative mønstre

De ovenfor skitserede metoder har alle været klassifikationsmetoder hvor et mål (typisk med et binært udfaldsrum) fastlægges ud fra en stor mængde indgående variable. Et eksempel på en simpel model med ganske få indgående variable findes i en undersøgelse af praktiserende lægers valg af behandling (typisk ved ordinering af medicin) i forbindelse med diagnoser (Rasmussen og Falkø-Lorenzen, 1999). Udvalgte praktiserende læger blev bedt om at koble enhver medicinordinering til en diagnose. Variablene medicin og diagnose havde hver et meget stort udfaldsrum. Yderligere havde patienter ofte flere diagnoser ved samme besøg. Spørgsmålet var, om det var muligt at foretage koblingen maskinelt på baggrund af de patientbesøg, hvor der kun var en enkelt diagnose, altså patientbesøg der ikke introducerede usikkerhed om, hvilken diagnose en medicinering tilhørte? Det viste sig, at en sådan simpel model kunne forbedre forudsigelsen fra 42 pct. ved en tilfældig forudsigelse til 91 pct. ved at benytte indlæringen fra den kendte sikre sammenhæng. Undersøgelsens konklusion er, at såfremt forudsigelsesprocenten vurderes at levere en tilstrækkelig præcision for det fænomen man ønsker at undersøge, kan metoden undgå den tidsmæssigt og omkostningskrævende del, det er at bebyrde de praktiseren-

de læger med at foretage koblingen mellem medicinering og diagnose.

Klassifikation opfattes ofte som binær, og såfremt output er kontinuert (fx mellem 0 og 1), opfattes værdien som en sandsynlighed for det interessante udfald (fx et salg). I tilfældet ovenfor var der tale om et stort muligt udfaldsrum, og i andre sammenhænge kan store udfaldsrum forekomme i forbindelse med andre typer input. Således er data mining af tekst et område med kvalitativt snarere end kvantitativt input. I et projekt foretoges klassifikation af nyhedstelegrammer baseret på de nye telegrammers ordmæssige afstand til tidligere klassificerede telegrammer (Berry & Linoff, 1997, s. 165). Dette område af kvalitativt input og blandede medieformer forventes at påkalde sig stor interesse for fremtidig data mining (Mitchell, 1999).

Med "automatic clustering" foretages en gruppering af enkelttilfælde på basis af deres placering i rummet. Sammenlignet med beslutningstræet foretages der ikke nødvendigvis klare opdelinger indenfor de dimensioner tilfældene er placeret i. I stedet opereres med et afstandsmål mellem observationer. Det vanskelige i anvendelsen af cluster-metoder vil være fastlæggelsen af et validt afstandsmål. Cluster-metoderne kan med fordel anvendes som en ikke-dirigeret indgang til datamaterialet.

Association

Mens de ovenstående eksempler på data mining har anvendt datasæt med en record for hvert udfald, er man ved associationsanalyse interesseret i sammenfald af produkter. Med samme udgangspunkt indenfor salg er spørgsmålet nu ikke, om der fandt et salg sted. I stedet ønskes belyst, hvilke produkter der blev købt samtidig. Analyseformen benævnes også "basket analysis". Ved parvist[14] at kombinere samtlige varer i indkøbskurven – registreret på samme bon ved "point-of-sale" (POS) – vil man kunne analysere sig frem til, hvilke vare-kombinationer der ofte forekommer. Et af de berømte eksempler er, at øl og bleer indkøbtes samtidig om torsdagen (Berry & Linoff, 1997, s. 126; Frappaolo, 1998)[15]. Imidlertid er det ikke klart, hvorledes denne information skal anvendes. Skal disse to varer placeres tæt på hinanden? Eller tværtimod langt fra hinanden så metervis af indbydende varehylder passeres undervejs? Eller skal man vende undersøgelsen af varer med høj affinitet på hovedet og koncentrere sig om at finde antagonistiske varer, som netop ikke forekommer i det samme indkøb? Her kræves en større marketing indsigt. Endelig er praktiske spørgsmål som kurvens størrelse (antallet af varer) af betydning. Der forekommer betydeligt flere kombinationer i varerne når indkøbet foretages i Bilka, end når det foregår i Fakta!

Når informationen udstrækkes fra at være et indblik i den enkelte og isolerede indkøbskurv, og i stedet yderligere kan kobles til information om kundens øvrige indkøb, opstår der hurtigt mere intuitive anvendelser af informationen. Ved registrering gennem indkøbsforeninger - fx FDB i Danmark - kan oplysninger om medlemmets indkøb benyttes til særdeles målrettet marketing mod relevante grupper[16]. Ved indkøb over Internet vil kunden også typisk være medlemsregistreret – om ikke andet så identificeret gennem gentagen anvendelse af det samme betalingskort – derfor kan virksomheden også her sammenkoble kundens indkøb foretaget på forskellige tidspunkter.

Aktion efter data mining

Som tidligere angivet i data-mining citatet fra Berry og Linoff (1997), lægges vægten på den praktiske aktion, efter at der vha. data mining er fundet værdifuld viden. Den opnåede viden skaber værdi gennem aktion. Et eksempel herpå findes hos Amazon.com, hvor man som kunde præsenteres for, hvilke andre bøger andre kunder der købte den fremfundne bog også har købt. Hermed bliver informationen om andre kunders indkøbsmønstre til interessant viden for andre kunder og dermed

værdifuld for virksomheden. Virksomheden etablerer og udnytter effekten i et netværk af produkter og kunde-til-kunde relationer ("C2C").

Anvendelse overfor enkeltkunder eller grupper

Information om og fra flere kunder kan rettes mod en enkelt kunde. Informationen om en enkelt kunde vil også typisk kunne anvendes overfor samme kunde, men denne enkeltstående opmærksomhed er en registrering og ikke resultatet af data mining. Begge anvendelsesformer vil falde ind under CRM ("Customer Relationship Management") og personalisering.

Resultaterne af data mining vil ofte være information opnået fra og om grupper og anvendt på grupper. Ovenfor er nævnt et eksempel med masseudsendelse af postal reklame, hvor omkostningerne minimeres mens de bedst egnede grupper udvælges.

Evaluering - igen

Når data mining modellen anvendes på et nyt materiale (fx et tilkøbt kunderegister) foretager modellen en scoring af hvert kundeemne, hvorefter de ønskede udvælges. Uanset hvilken model der benyttes ved data mining, bør resultatet (målet), efter at aktion er taget, nøje registreres. For det første for at kunne foretage en evaluering af selve data mining projektet. For det andet vil den efterfølgende registrering danne grundlaget for gennem yderligere dirigeret data mining at kunne opnå sikrere viden gennem anvendelse af et nyere og større læringsdatasæt.

Processer for data varehouse og data mining er således for virksomheden kilder til en stadig udbygning af virksomhedens vidensgrundlag i data varehuset og den gennem data mining opnåede viden.

Anvendt kundeviden

Data mining baserer sig på store mængder af data. Selvom der skjuler sig enkeltpersoner og familier i materialerne, tager data mining metoder ikke højde herfor. Polemisk udtrykt kan data mining betragtes som "anvendte fordomme" ved anvendelse af gennemsnit og grupperinger. Men denne tilgang kan opblødes gennem virsomhedens intense anvendelse af personalisering. Det forventes, at virksomheder, der forstår at kombinere deres generelle viden med den personlige viden om den enkelte kunde, vil opnå store konkurrencefordele overfor virksomheder, der reagerer efter ufleksible regler. Har man været bankens kunde i 30 år, og ens kontokort inddrages i en VISA automat på Sicilien ved et overtræk på 500 kr. – ja, så skal banken ikke forvente at beholde den kunde. Og det vidste banken vel egentlig godt!

Summary

Data mining deals with new methods of analysing large quantities of data in the organisation. This survey article discusses some of the prevailing methods within data mining. Special attention is given to the objective of data mining: organisation data created in organisation processes. As still more processes produce still more data, the organisation produces an increasing flow of data. The use of the Internet in particular releases an explosion of the quantities of data. The analysis of large quantities of data demands high-level information technology. The necessity to produce suitable data for data mining also involves the establishment of a data warehouse to ensure a reliable supply of integrated and valid data into the data mining of the organisation.

Noter

1. Pulterrum og vidanspunpe forekommer som begreberne "knowledge attic" og "knowledge pump" hos Hejst, Spek, Kruizinga (1998, s. 26).
2. Se www.spss.com/clementine.
3. Hvor www.kdnuggets.com reklamerer med at være " Your Guide to Data Mining, Web Mining, Knowledge Discovery, and e-CRM".
4. Både "Very Large Data Bases" og "Knowledge Discovery in Data" (senere udvidet med "and Data Mining") har som interessegrupper under den faglige organisation ACM (Association for Computing Machinery – www.acm.org) egne konferencer og publikationer.
5. Et dansk varehus er ikke helt det samme som et US "warehouse", som snarere dækker over et dansk lager eller pakhus. Alligevel anvendes her fordanskningen data varehus, fordi udtrykket er mundret og har vundet frem i brug. Omvendt benyttes termen "data mining" på engelsk, men ofte med dansk bøjning som i den bestemte form "data miningen".
6. Ligesom en IP-adresse af internetudbyderen kan benyttes til flere abonnenter, kan en PC (hvor cookie-filen ligger) også benyttes af flere brugere. Individidentificeringen er altså stadig ikke er helt sikker.
7. I softwaren SAS Enterprise Miner findes der således et specielt led hvorigennem manglende data kan erstattes efter flere metoder ("imputation").
8. Ifølge Thomas Zizzo, "Churn, baby, churn" i Electronic Business, July 2000. Nyligt etablerede telefonselskaber i Danmark viser også regnskaber som kun er attraktive set i fremtidens håbefulde skær.
9. Berry & Linoff ville gerne give matricen et andet navn (2000, s. 55), men navnet "confusion matrix" hænger ved selvom matricen gerne skulle skabe oversigt snarere end forvirring.
10. Almindeligvis tegnes beslutningstræet oppefra og nedefter; således passer billedet egentlig bedre på træets rødder. Men da der indenfor beslutningstræer tales om blade, grene, kviste og forgrening og ikke mindst om at foretage beskæring ("pruning"), så er metaforen træ ganske veletableret.
11. I SAS Enterprise Miner skrives således blot at der er valgt en hybrid af det bedste fra CHAID, CART og C4.5 algoritmerne (SAS, 1999a, s. 9).
12. Faktisk er beslutningstræet så udbredt indenfor data mining at nogle produkter (fx Analysis fra det danske firma Targit A/S) har beslutningstræet som deres "single click data mining".
13. Disse lift-kurver er frembragt med SAS Enterprise Miner udlånt af SAS Institute, København.
14. Analyser kan foretages på højere niveauer end de parvise kombinationer.
15. Historien bliver nogle gange udlagt som ren fiktion, og alene en historie der skulle promovere kædeforretningen WalMart.
16. FDB har potentielt et enestående materiale pga. de mange medlemmer, mange kæder og mange butikker.

Litteratur

Berry, Michael J.A.; Linoff, Gordon S.: Mastering Data Mining: The Art and Science of Customer Relationship Management, John Wiley & Sons, New York, NY, 2000.

Berry, Michael J.A.; Linoff, Gordon: Data Mining Techniques: For Marketing, Sales, and Customer Support, John Wiley & Sons, 1997.

Berthold, Michael & Hand, David J. (eds.): Intelligent Data Analysis. An introduction, Springer, Berlin, 1999.

Borghoff, Uwe & Pareschi, Remo (eds.): Information Technology for Knowledge Management, Springer, 1998.

Brandi, Søren & Hildebrandt, Steen (eds.): Kompetenceguldet, Børsens Bøger, København, 2000.

Frappaolo, Carl: Defining knowledge management: four basic functions, 1998, Computerworld (US), February 23, 1998.

Glymour, Clark; Madigan, David; Pregibon, Daryl; Smyth, Padhraaic: Statistical Themes and Lessons for Data Mining, Data Mining and Knowledge Discovery 1, 11-28 (1997), Kluwer Academic Publisher, 1997.

Harman, Harry H.: Modern Factor Analysis (2. ed), University of Chicago Press, 1967.

van Heijst, Gertjan; van der Spek, Rob; Krunzinga, Eelco: The Lessons Learned Cycle, i Borghoff & Pareschi, 1998.

Inmon, W.H.: Building the Data Warehouse (2.ed.), John Wiley & Sons, 1996.

Kimball, Ralph: The Data Warehouse Toolkit: Practical Techniques for Building Dimensional Data Warehouses, John Wiley & Sons, New York, NY, 1996.

Kimball, Ralph; Merz, Richard: The Data Webhouse Toolkit, John Wiley & Sons, New York, NY, 2000.

Laudon, Kenneth C. and Laudon, Jane Price: Management Information Systems: Organization and Technology in the Networked Enterprise (6.ed.), Prentice-Hall, Upper Saddle River, NJ, 2000.

Mena, Jesus: Data Mining Your Website, Digital Press, Boston, MA, 1999.

Mitchell, Thomas M.: Machine Learning, McGraw-Hill, 1997.

Mitchell, Thomas M.: Machine Learning and Data Mining, Communications of the ACM 42 (11; 31-36), 1999.

Rasmussen, Karsten Boye: Datadokumentation. Metadata for samfundsvidenskabelige undersøgelser, Odense Universitets Forlag, Odense, 2000.

Rasmussen, Karsten Boye; Falkø-Lorentzen, Erik: Statistical linkage of treatment and diagnosis. Report for Fyns Amt, 1999.

Rothenberg, Jeff: Ensuring the Longevity of Digital Documents, 1995, Scientific American January 1995.

SAS: Enterprise Miner. Applying Data Mining Techniques. Course Notes, 1999a, SAS 56606, Cary, NC, 1999.

SAS: Getting started with Enterprise Miner Software, Version 3.0, SAS 56869, Cary, NC, 1999.

Weiss, Sholom M.; Indurkhya, Nitin: Predictive Data Mining: A Practical Guide, Morgan Kaufman, 1998.

Welbrock, Peter R.: Strategic Data Warehousing Principles Using SAS Software, SAS Institute Inc., Cary, NC, 1998.

Zizzo, Thomas: Churn, baby, churn, Electronic Business, July 2000.