Data Mining af strukturnøgletal

Chr. Kaasen, forår 2002

Som begreb er *data mining* (DM) af forholdsvis ny dato. De største og mest citerede værker er alle skrevet indenfor de seneste 5 år og DM-programmer som SPSS´s Clementine og SAS´s Enterprise Miner har kun nogle få år på bagen[1]. På trods heraf - eller måske netop derfor - ses referencer til DM i mange og vidt forskellige sammenhænge - ofte nærmest som en slags "Sesam luk dig op" –tryllefor-mular, der er i stand til åbne døren til en skjult informationsskat eller for at blive i den faglige jargon finde "guldklumper" i en enorm og uoverskuelig informationsmængde. Det åbne spørgsmål er natur-ligvis: er der ægte guldklumper[2] i virksomhedens data eller er det bare glimmer? Det forhold at store, erfarne og agtværdige statistikprogramhuse har udviklet og inkluderet særskilte produkter med denne etikette i sortimentet lader formode, at de i det mindste anser det for en kommerciel salgbar artikel. Men hvordan ser det ud fra et brugersynspunkt? Er der tale om ny etikette på gamle flasker eller er der tale om en ny fagdisciplin afledt af den informationsteknologiske udvikling og med blivende værdi?

I dette notat redegøres først for nogle hovedtræk i begrebet DM og derefter vises et mindre eksempel på hvordan en manuelt udført DM-lignende analyse kan gennemføres.

[1] IBM Intelligent Miner, Oracle Darwin, SGI MineSet er andre komplette – horisontale - DM-programmer. Cognos Scenario, Business Objects Miner o.a. er mere specialiserede med et snævrere sigte

[2] KDnuggets er et fagligt forum for data mining og på deres hjemmeside http://kdnuggets.com er der faktisk noget der kan minde om "guldklumper". Mindre gyldent men mere mytologisk er Two Craws (nemlig Odins Hugin og Munin, der hver morgen flyver ud og ser hvad der sker), men med en meget interessant hjemmeside: http://www.twocraws.com

Data mining

Lidt løst udtrykt kan man sige, at filosofien bag DM er at skabe ny viden ved at "presse" upåagtet information ud af de enorme mængde rådata, som de moderne EDB-baserede registreringer har skabt. Tanken er her, at de store systematiske datamængder - eksempelvis kassebonerne i et supermarked - indeholder skjult information, som gennem en – ofte meget omfattende – DM kan gøres til brugbar viden. Afslører en DM af kasseboner – benævnt 'basket analysis' - f.eks. at nogle varer, som ikke har en brugsmæssig sammenhæng, alligevel meget ofte købes samtidig, kan denne information gøres brugbar dels gennem varernes placering i forretningen (skal de stå ved siden af hinanden eller i hver sin ende af forretningen?) og i tilrettelægningen af den ugentlige tilbudsavis, hvor man så på skift kan have de købssammenhørende varer på tilbud.

Sigte og mål

I et af hovedværkerne om DM - "Data Mining Techniques" - fra 1997 angiver forfatterne Berry og Linoff følgende definition på DM:

> "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 og Linoff, 1997, s.5)

og side 18 angives formålet med denne analyse:

> *" .....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"* (op. cit. p. 18)

SAS´s definition af DM er lidt mere kommerciel:

> "*Data mining is the process of selecting, exploring and modeling large amounts of data to uncover previously unknown patterns for business advantage*"
> (http://sas.com/technologies/data_mining)

Det centrale er altså om analysen skaber brugbar viden – det vil sige om analysens resultater skaber grundlag for mere profitable handlinger. SAS´s lidt mere dæmpede målkrav til analysen svarer mere til de forfattere, der tilkendegiver, at et bedre resultat / større forventede værdi er at foretrække, men at en reduktion af usikkerheden også er en værdifuld fordel – f.eks. ved at undgå besværlige / tidrøvende / tabsgivende kunder eller ved at accelerere beslutningsprocessen. DM er altså et beslutningsstøtteværktøj på linie med statistik, køteori og operationsanalyse. Der er dog en markant forskel: førstnævnte er i al væsentlighed induktiv medens de 3 sidstnævnte er af deduktiv karakter.

> **Eksempel:** En kontokortudsteder ønsker at bestemme hvilke faktorer der er årsag til misligholdelse. I en traditionel statistisk analyse vil første skridt være at opstille en hypotese eksempelvis at indkomsten er afgørende. Hvis dataene ikke kan verificere indkomstens betydning opstilles en ny hypotese eksempelvis at kontokortholderens gæld er afgørende for misligholdelsen. Kan dataene heller ikke understøtte denne hypotese, vil indkomst i forhold til gæld være en 3. hypotese, der kunne te-

stes. Ved en traditionel statistisk analyse genereres således en række hypoteser, som én efter én testes på dataene – altså en deduktiv fremgangsmåde. Ved store datamængder bliver denne fremgangsmåde hurtigt helt uoverskuelig. DM-teknikkerne er derimod indrettet på at kunne håndtere store datamængder. I stedet for at bekræfte hypoteser anvendes dataene til at afdække mønstre og / eller sammenhænge. I nærværende eksempel ville en DM-analyse helt ukritisk anvende alle tilgængelige informationer og kunne eksempelvis fremkomme med et resultat at alder, postnummer, indkomst og gæld – evt. i vægtet format – bedre end noget andet kan identificere potentielle misligholdere. Den deduktive fremgangsmåde vil sædvanligvis være styret af en årsag - virkning sammenhæng, medens de induktive DM-metoder ikke er begrænset af sådanne betragtninger og derfor også kan fremkomme med overraskende og uventede resultater – dvs. sammenhænge som hverken teoretisk eller logisk kan forklares, men som dog ses i eller fremgår af dataene

Som udgangspunkt er DM en "bottom up" strategi. Det vil sige at DM er *data drevet* i modsætning til sædvanlige statistiske undersøgelser - der med nævnte terminologi er en "top down"-strategi - idet sådanne er *brugerstyret* dels i form af brugerdefinerede kriterier*,* dels i form af en af brugeren formuleret hypotese. Her er dataenes funktion begrænset til 'kun' at verificere eller falsificere hypotesen iht. til de opstillede kriterier. DM er "tallenes tale" til brugeren og deres tale er mængder – nemlig overnormale hyppigheder eller koncentrationer af visse data eller data-kombinationer, der kan omsættes i ny viden og dermed forbedre beslutningsgrundlaget.

Datastyringen implicerer at der anvendes mere eller mindre automatiserede metoder, som efter nogle gennemløb af dataene kan frembringe en model. Typisk udvikles modellen på en delmængde af dataene for derefter at blive testet på en anden delmængde. Testen medfører ofte en yderligere præcisering af modellen, som derefter skal testes på en tredje delmængde af dataene o.s.f. En sådan 'træning' af modellen kan naturligvis nemt resultere i 'overfitting' hvorved forstås en næsten 100 % korrekt gengivelse af de historiske træningsdata. I sig selv er der intet negativt i at modellen trænes til høj grad af perfektionisme på et givet datasæt. Det bliver først problematisk når overfittingen skaber forventninger om at modellen vil have samme udsagnskraft i fremtiden på helt andre data. Det vil den ikke – ofte bliver dens forudsigekraft næsten halveret på fremtidige data.

Afdækningen af meningsfyldte mønstre og regelmæssigheder kræver omfattende og systematiske registreringer både mht. antal poster og mht. antal variable.[3] Lidt firkantet sagt er antallet af poster afgørende for, hvor sikkert en model kan fastlægges, medens antallet af variable er afgørende for hvor mange modeller, der kan undersøges. I DM er dataenes funktion derfor både modelkonstruktion og modelvalidering – altså en langt mere omfattende opgave end i traditionelle statistiske undersøgelser. Fokus er således flyttet fra at finde den *rigtige* model – og dermed også den bagvedliggende forklaring – til at finde de *rigtige* informationer – dvs. de mest anvendelige informationer. Der er således ingen nævneværdig forudgående teoretisk begrundet udvælgelse af variable og sammenhænge ligesom de fra hypoteseprøvningen kendte 5 og 1 % signifikansværdier er mindre væsentlige i DM. Der er dog ikke tale om en enten eller situation, men snarere både og, idet DM-analysens resultater ofte vil blive verificeret og evalueret med traditionelle statistiske metoder, der er langt bedre til at angive sammenhængenes styrke i en kendt og anerkendt form.

[3] Brugervenlig organisering af data behandles under begrebet Data Warehouses, der dog ikke vil blive omtalt her

I forhold til traditionel statistisk hypoteseprøvning er det således en meget prunkløs og pragmatisk modelevaluering i DM. Det gør naturligvis ikke DM mindre seriøs, at teoretikernes ofte noget abstrakte termer er erstattet af et meget mere håndgribeligt 'værdiskabelseskriterium'. Et sådant kriterium er i langt bedre overensstemmelse med virksomhedsledelsens dagligdag og begrebsverden, hvor lakmusprøven for beslutningernes kvalitet netop er deres positive effekt, og vil alene af den grund være mere handlingsskabende. Teoretikernes krav om 95 % eller 99 % sikkerhed vil i sig selv ofte virke handlingslammende i en beslutningssituation, idet det vil være uhyre sjældent, at der kan opnås så overbevisende sikkerhedsmarginer i erhvervslivet. En reduktion af de teoretiske sikkerhedskrav eller konfidensintervaller vil derfor også øge DM´s operationalitet.

DM tilgodeser først og fremmest brugerens behov for handleinformation frem for teoretikerens behov for overbevisende hypotese-falsifikation. Selvom den statistiske forklaringskraft kan beregnes til at være ret ringe, vil det dog ofte være således at udsagnskraften er tilstrækkelig til at være interessant i en forretningsmæssig sammenhæng. For en beslutningstager vil en statistisk signifikans på 95 ud af 100 være ganske uinteressant hvis hans normale succésrate ligger på 2 ud af 3. Alt hvad han ønsker er at den fremtidige succésrate forøges udover dette niveau. I DM bliver den logiske og metodemæssige modelevaluering derfor stort set lig med en pragmatisk modelverifikation – helst i form af en nøje overvåget og ofte klart af- og begrænset eksperiment, som kan bekræfte / afkræfte, at dens udsagnskraft giver et bedre resultat – 'the proof of the pudding is in the eating' – alternativt kan en testdatabase anvendes.

Metoder

DM gør brug af et stort antal meget forskellige metoder. Da hyppigheder, koncentrationer, samvariationer, associationer og kategoriseringer hører til blandt de hyppigste mål for analyserne, er de statistiske metoder naturligvis i overtal, men der er tilsyneladende også hentet inspiration fra køteorien og operationsanalysen. Der er dog ofte tale om en betydelig udbygning af de klassiske statistiske metoder. Udviklingen af de klassiske statistiske teknikker har sædvanligvis været baseret på anerkendte teoretiske og metodiske overvejelser og for en overskuelig og til den givne teknik tilpasset datamængde, eksempelvis mht. fordelingens art – mao. teknikkernes anvendelse hviler på en række forudsætninger af såvel teoretisk som datamæssig karakter. Dette er ikke tilfældet for DM-teknikkerne. De store metodeutilpassede datamængder har nødvendiggjort og adgangen til stadig større og mere kraftfulde computere har muliggjort en udvikling af teknikker og metoder baseret på rå og brutal regnekraft – altså en helt anderledes kraftbetonet tilgang til analysen .

En gruppering af de mange metoder kan gøres ud fra karakteren af det output de enkelte metoder frembringer. Svarende til en statisk / dynamisk synsvinkel kan DM-metoderne hensigtsmæssig opdeles i metoder, der sigter mod *at beskrive* og i metoder, der sigter mod *at forudsige*[4].

I litteraturen fremhæves at en DM-analyse altid bør starte med en meget tæt og præcis beskrivelse af de data som skal indgå i undersøgelsen herunder en bestemmelse af datatype – kontinuerte eller kategoriale variable (sidstnævnte kan yderligere splittes i ordinale (høj, mellem, lav) eller nominelle (postnumre) variable). Udover beregning af gennemsnit, standardafvigelse, standardiserede 3. og 4. momenter anbefales bred anvendelse af grafiske afbild-

[4] Opdelingen svaret stort set til det der i amerikansk literatur betegnes 'knowledge discovery' og 'predictive data mining'

ninger som f.eks. frekvensfordelinger, histogrammer samt 2- eller 3-dimensionelle afbildninger evt. vha. af pivottabeller. Sidstnævnte kan også være brugbare til homogenitetstests.

Klyngeanalyse (clustering) henregnes til databeskrivelsesteknikkerne. Sigtet hermed er at inddele datamaterialet i grupper / klynger hvor enhederne indenfor grupperne har en betydelig lighed og hvor enhederne mellem grupperne adskiller sig mest mulig. Klyngeanalysen adskiller sig fra klassifikation ved at man på forhånd ikke kan vide hvilke egenskaber eller egenskabskombinationer, der danner grundlag for klyngedannelsen medens klassifikation / segmentering altid sker ud fra kendte baggrundsvariables værdier. Klyngedannelse kan udføres ved hjælp af forskellige teknikker, hvor de underliggende algoritmer som oftest dog er baseret på varians / kovariansmatricen. Ekstrahering af en fælles faktor, der knytter alle enheder sammen i de enkelte klynger, er en overordentlig vanskelig opgave, som stort set er overladt til den menneskelige hjerne og fantasi. Den bedst kendte og mest veludviklede teknik er utvivlsomt faktor-analysen, som uanset dens eksplorative karakter normalt henregnes til den klassiske statistiks metoder og af samme grund ikke ses i de kommercielle DM-programmer. Eksempelvis er denne metode meget anvendt inden for markedsundersøgelser til at finde frem til forskellige forbrugeres livs-stilsmønstre og netop problemet med at finde en fælles dækkende etikette for de enkelte klynger har bevirket en helt intetsigende navngivning efter farver – blå, grøn, violet, rosa og grå! Nogle forfattere har dog forsøgt at identificere en fælles karakteristisk faktor for de enkelte klynger som 'konservative kvalitetsforbrugere', 'trendy forbrugere' mv.

Til beskrivelsesteknikkerne henregnes også kædeanalyse (link analysis), hvis sigte er at identificere sammenhænge – associationer – i datamaterialet. I sin grundform er det en langt enklere teknik end eksempelvis klyngeanalysen, idet samhørigheden bestemmes ved optælling af sammenfald og beregning af relative og betingede hyppigheder. Bonanalyse er en typisk kædeanalyse, hvor det undersøges hvor hyppigt forskellige varer købes samtidig – aktivitetsbetinget samhørighed – medens sekvensbetinget samhørighed beregner hyppigheden af sammenfald over tiden.

Metoder til forudsigelser underopdeles sædvanligvis efter outputtets form: 1) klassifikationsmetoder – kategorialt output, 2) regressionsmetoder – kontinuert output og 3) tidsseriemetoder – mønster output. Metoderne i denne gruppe skal trænes – dvs. de skal udvikles og testes på baggrund af et historisk materiale hvor såvel inputvariablernes som resultatvariablernes værdier er kendte.

Klassifikationsmetoderne sigter mod at identificere de karakteristika, der bedst mulig angiver hvilken klasse den enkelte observation tilhører. I den mest simple form er det en opdeling i *at være* eller *ikke at være* – at være en sandsynlig kunde eller at være en ikke sandsynlig kunde, at være misligholder af et lån eller ikke misligholder osv. Den ideologiske basis for klassifikationsmetoderne kan henføres til R. A Fishers diskriminantanalyse fra 1935, men teknisk set er der sket en betydelig udvikling af metoden. Moderne diskriminantmetoder kan inddele et observationsmateriale i mange klasser og kan håndtere mange forskellige former for afhængighed mellem de indgående variable.

Beslutningstræet henregnes sædvanligvis også til klassifikationsmetoderne. Det er en central metode i alle DM-programmer og den ses i mange varianter. Metodens popularitet skyldes dels at den logiske opdeling gør den umiddelbart forståelig og ofte også at forgreningerne direkte kan omsættes til handlinger som kundegruppering, sortimentssammensætning o. lign., dels at computeren er ideel til at sortere og lave betingede udvælgelse i et stort datamateriale. Desuden kan beslutningstræet uden problemer tage hånd om kategoriale som kontinuerte variable – hver for sig såvel som blandet. En selektering efter bopæl og købsfrekvens

(kategoriale variable) sammen købestørrelse (kontinuert variabel) kan nemt formuleres som en SQL-statement. Der findes dog også en række mere datastyrede algoritmer til udvælgelse af de bedste opdelinger, der bl.a. sikrer at variationen indenfor gruppen minimeres medens variationen mellem grupperne / grenene maksimeres som f.eks. CHAID (Chi-squared Automatic Interaction Detection) og CART (Classification And Regression Trees). Især ved kontinuerte variable bør de datastyrede algoritmer anvendes, idet en traditionel opdeling i afrundede klumper på ingen måder sikrer gode afskæringsværdier – eksempelvis kunne en sådan være indkomsten, der ofte inddeles i intervaller som 300.000 kr. til 500.000 kr., men er en indtægt på 500.001 kr. så meget bedre, at den ikke hører med til intervallet?

Kunstig neurale netværk er som regel flagskibet i de fleste DM-programmers regressionsmodeller. Neurale netværk kræver at inputvariablerne er kontinuerte. Et sådant netværk består principielt af 3 lag noder: inputnoder, skjulte indre noder (med et billede fra hjerneforskningen betegnes de ofte som neuroner, hvilket dog forekommer at være en alt for positiv allegori) og outputnode(r). Hver inputnode er forbundet med alle indre noder og alle indre er forbundet med outputnode(rne) . De skjulte noder indeholder en aktiverings / frembringefunktion, som hver især aktiverer en vægt, som indgår i beregningen af outputtet. En sådan model skal trænes rigtig meget, hvilket gøres ved at fodre den et sæt sammenhørende værdier af inputvariablerne og outputvariablerne. Ved hver fodring indstiller de indre noders funktioner sig på en værdi, som sikrer, at det ønskede output nås. De indre noder husker disse værdier og ved næste fodring anvendes de til at frembringe / beregne et output. Afviger dette beregnede output fra det aktuelle kendte output svarende til de givne inputværdier, indeholder netværket mekanismer til at korrigere vægtene, således at det beregnede resultat svarer til det realiserede resultat. Også disse nye / korrigerede vægte huskes til næste beregning. Ved hvert gennemløb bliver netværket "klogere" dvs. at de aktive noder finjusterer vægtberegningerne og/eller flere noder aktiveres for at kunne frembringe det ønskede resultat.

Efter forbillede fra den menneskelige hjerne med dens mange neuroner, der enten er aktive dvs. afgive en impuls eller ikke aktive, var de indre noders aktiveringsfunktion en diskret funktion, der kunne antage værdien 0 eller 1. Matematisk set er det en ret besværlig funktion og derfor anvendes som regel sigmoid-kurve – dvs. en S-formet kurve (altså en logistisk funktion) i intervallet 0 til 1, dvs. stort negativt input og stort positivt input vil have værdier på hhv. 0 og 1. I et snævert interval omkring 0 skifter funktionen. Når modellen trænes 'lærer' de for hvilke inputværdier de skal være aktive – altså afgive en impuls til ("tænde for") vægtberegningen af outputtet. Antallet af noder og / eller antal aktive noder og hvilke tærskelværdier der har aktiveret dem er det sædvanligvis ikke muligt at få oplyst. I store netværk kan der være flere lag skjulte indre noder, hvor første lags beregnede vægte bruges som inputværdier i andet lags noder hvorefter de beregnede vægte herfra enten anvendes i et 3. lag eller anvendes til at beregne outputtet. Men uanset om der er 1 eller flere skjulte lag noder vil outputtet være en ikke-lineær kombination af inputnoderne. Helt specielt gælder dog, at hvis outputnoden indeholder en lineær aktiveringsfunktion og at der ikke findes skjulte noder så vil det neurale netværk være en multipel lineær regression med samme antal variable, som der er inputnoder.

Et neuralt netværk fungerer altså som en "black box", der på én eller anden ukendt måde transformerer et input til et output, hvilket bevirker, at det stort set er umuligt at fortolke resultatet - herunder at give et mål for usikkerheden i resultatet, at beregne følsomheden over for variationer i inputværdierne, angive modellens robusthed o.lign. Billedligt talt er et neuralt netværk en "målet helliger midlet maskine", idet anvendelsen af et sådant netværk helt klart er udtryk for at resultatet prioriteres betydeligt højere end forklaringen. Denne totale negligering af årsag-virkning sammenhængen er udtryk for ren pragmatisme, der for en bruger / beslutningstager kan virke tiltrækkende, men den, der er afhængig af resultatet,

bruger / beslutningstager kan virke tiltrækkende, men den, der er afhængig af resultatet, kan meget vel have en helt anden opfattelse. En student, der søger optagelse på universitetet og får afslag med begrundelsen "Vort neurale netværk siger *´dur ikke´*" vil næppe finde et sådant orakelsvar passende uanset om afvisningen er baseret på hele universitets erfaringsmasse omkring studiegennemførelse. En vis forsigtighed i omgangen med neurale netværk synes påkrævet.

Alle former for regression – lineær, multipel, kvadratisk, logistisk osv. – er også standardværktøjer i DM-programmerne. Da disse modeller hviler på et kendt og teoretisk anerkendt metodegrundlag er tolkningsmulighederne langt bedre og sikrere end i et neuralt netværk. Dog kan selv disse metoder give visse forklaringsproblemer. Inkorporering af et interaktivt led i en multipel lineær regression fordi det forbedrer korrelationskoefficienten og dermed giver et mere præcist output er i sig selv ikke nogen forklaring på hvorfor og hvordan de 2 inputvariable interagerer. Kan en sådan interaktion mellem 2 inputvariable ikke logisk begrundes, er der tale om et relativt fald i forklaringsevnen og dermed tilliden til modellen selvom dens forudsigelser forbedres ved medtagelse af det interaktive led. De seneste årtiers enorme erfaringshøst med regressionsmodeller har dog givet en voksende erkendelse af deres robusthed overfor mangler og brister i de teoretiske og datamæssige forudsætninger for deres anvendelse og dermed også en vis berettiget tilbøjelighed til helt eller delvist at se bort fra en sådan logisk evaluering af modellen. For blot 25 år siden betragtedes eksempelvis dummyvariable med en betydelig skepsis – de er ikke normalfordelte; i dag indgår de uden advarende ord i enhver statistiklærebog.

Tidsseriemetoderne kan være såvel matematik som computerbaseret. Medens de matematiske modeller som hovedregel er tidsstyrede – dvs. afhængig af at begivenhederne finder sted på samme tidspunkt eller med konstante tidsmellemrum – er de computerbaserede metoder som regel begivenhedsstyrede. Computeren har en uovertruffen evne til at kunne fastholde en sekvens af begivenheder og ved ren og skær sammenligning at kunne identificere alle lignende sekvenser – også selvom tiden mellem begivenhederne varierer.

Ovennævnte hurtige og overfladiske kig i DM´s ganske omfangsrige værktøjskasse er på ingen måder udtømmende hverken mht. antal metoder, deres formåen eller deres anvendelighed. De kommercielle DM-programmer vil ofte også indeholde analysemetoder, som programhusene selv har udviklet - såkaldte proprietærmetoder – samt en procesmodel, der i en række faser angiver hvilke opgaver, der skal løses for at nå et godt resultat. En række europæiske selskaber med deltagelse af bl.a. NCR Denmark har udviklet en sådan procesmodel kaldet CRISP-DM – Cross-Industry Standard Proces for Data Mining.

Udover den omfangsrige metodeværktøjskasse er det også kendetegnende, at i DM anvendes metoderne ofte i lag. Resultaterne fra én metode anvendes ofte som input i en anden metode og resultaterne herfra kan så eventuelt anvendes i en 3. metode. Baggrunden herfor er bl.a. den, at dataene ikke er indsamlet med en speciel eller veldefineret analyse for øje i modsætning til den traditionelle statiske analyse, hvor mønstret kort kan gengives som: hypoteseformulering, undersøgelsesdesign, operationalisering, dataindsamling, hypotesetest og konklusion. I den datadrevne DM vil de tre første faser ofte blive gennemført vha. af nogle standardiserede metoder og resultaterne herfra indgår derefter som inputvariable i andre metoder. Da målet med DM er maximal handleinformation, vil en flertrinsanalyse ofte også give et mere komplet billede af virkeligheden, så også af den grund vil metoderne ofte anvendes i lag.

DM kan anvendes i alle tilfælde hvor datamængden er tilstrækkelig omfangsrig og systematisk til at den understøtter modelkonstruktion og modeltest. De største datamængder frem-

kommer som regel i forbindelser med transaktioner og det er da også netop kunderelationer og kundeadfærd, der hyppigst fremdrages som eksempler i litteraturen. Kundeloyalitet er et af de centrale genstandsfelter for DM-undersøgelser. Ud fra en umiddelbar betragtning vil en sådan undersøgelse stille krav om at kunderne kan identificeres som individuelle kunder, hvilket bl.a. er tilfældet for banker, forsikringsselskaber, telefonselskaber, rejsebureauer, postordrefirmaer, apoteker m.fl. Ved transaktioner, der er baseret på kontantbetaling, er det derimod ikke så lige til at identificere den enkelte kunde. Dog i de tilfælde, hvor betalingen overvejende sker ved hjælp af kredit- eller betalingskort kan kortets kode identificere de enkelte salg om end kunden som sådan er ukendt. Ubemandede benzinstationer er et godt eksempel på en sådan virksomhed. Klikanalyse dvs. sporingen af internet-kunders vej frem til køb eller det modsatte er ligeledes et stort emne for DM – en analyseform Jubii bl.a. har udnyttet til at prissætte visse af portalens ydelser.

Udpegning af dubiøse debitorer, lager-, sortiments- og logistikanalyser er eksempler på en virksomhedsintern anvendelse. Et tredje område er overvågning og sporing af kriminelle handlinger. PBS bruger således datamining til at overvåge brugen af dankort bl.a. med henblik på sporing af ulovlig anvendelse af kortet, medens visse forsikringsselskaber leder efter mulig forsikringssvindel. På børserne er det insiderhandel, der er genstand for overvågning. Normalt vil selskabets egen database danne baggrund for en DM-undersøgelse, men ofte vil den blive suppleret med data fra offentlige statistikker eller speciel indsamlede data. Trafikintensiteten og konkurrenternes beliggenhed kunne eksempelvis give andre dimensioner i et benzinselskabs analyser af kundeloyaliteten.

Evaluering

Det afgørende nye i DM er først og fremmest omkostningseffektiviteten. Tidligere var omkostningerne ved statistisk vidensproduktion nærmest prohibitive: lang tids forberedelse med planlægning og design ofte med hjælp af dyre eksperter, omhyggelig og dermed dyr dataindsamling, som regel meget enstrengede og tidkrævende dataanalyse på ikke altid lige velegnede main frames og med betydelig usikkerhed om brugbarheden af de frembragte resultater var forhold, som i de fleste tilfælde gjorde en sådan vidensproduktion aldeles uinteressant for det store flertal af virksomheder.

Den omfattende og stærk stigende elektroniske registrering har sammen med et enormt prisfald på harddiske givet en datatilgængelighed, der er forbedret mange mange gange i de seneste 25 år og til omkostninger, der med datidens øjne nærmest kan betragtes som gratis. Data, der er indsamlet som et led i forretningsmæssige transaktioner, vil naturligvis ikke altid være lige velegnede til en DM, og ofte må der da også udføres en såkaldt datarensningsproces – data cleansing – der gør dataene brugbare. Da datarensningen i vid udstrækning kan gøres elektronisk vil det dog ikke medføre nævneværdigt større omkostninger ved datatilgængeligheden.

Omkostningerne ved selve databehandlingen er ligeledes decimeret mange gange. Fordoblingen af processorerne hastighed hvert andet år samtidig med at båndbredden er 4-5 doblet og prisen på 1 MB RAM er faldet til brøkdele af datidens priser har medført en enorm stigning i behandlingskapaciteten til stadig stærkt faldende omkostninger. Muligheden af at kunne parallelforbinde PC´ere i et netværk har ligeledes mangedoblet databehandlingskapaciteten.

Medens udviklingen på hardwaresiden er særdeles synlig og spektakulær er dette ikke helt tilfældet mht. softwaren. Det forhold at store stærke og velgennemtestede statistikpro-

grammer er gjort umiddelbart tilgængelige for afvikling på billige PC´ere har utvivlsomt skabt en bedre og bredere metodefortrolighed og dermed banet vejen for DM´s udbredelse. Hertil kommer meget store forbedringer i brugervenligheden både hvad angår operationalitet og resultatpræsentationen. GUI og SQL er selvfølgeligheder i den forbindelse hvorimod modelbygning med grafiske pictogramlignende objekter er af helt ny dato. Hvert pictogram dækker over en standardiseret algoritme der på forskellige måder kan føjes sammen til en komplet fuld funktionel model næsten som legoklodser, der kan anvendes til at bygge huse såvel som tog og rumstationer. En sådan visualisering skaber overblik, synliggør sammenhænge og skaber fokus. Grafisk repræsentation af inputdata og resultatvariable - såvel todimensionalt som tredimensionalt - er ligeledes udtryk for en forbedret brugervenlighed.

Historisk set er mange af metoderne udviklet indenfor de seneste par tiår af uafhængige forskere – ofte med henblik på løsning af specielle problemstillinger og ikke som et egentlig DM-værktøj. Et fælles kendetegn for disse metoder er deres eksplorative karakter og væsentligst på dette grundlag er de fagdisciplineret under et noget misvisende, men kommercielt set meget fordelagtige begreb: Data Mining. Som det normalt er tilfældet har denne afgrænsning utvivlsomt været en fordel: øget faglig og teoretisk interesse, afdækning af huller og mangler, mere konsistent begrebsdannelse, øget metodeformalisering anføres sædvanligvis som de væsentligste. Men også kommercialiseringen har sine fordele: øget tilgængelighed, større brugervenlighed, udvikling af effektive og robuste algoritmer og ikke mindst en økonomisk sikkerhed for fortsat interesse. Kommercialiseringen har dog også sine negative sider. Anvendelsen af ikke-standardiserede, fancy, men storsælgende betegnelser – tilsyneladende helst i form af akronymer – gør materien mere vanskelig tilgængelig end nødvendigt.

Implementering af DM er en omfattende og tidkrævende aktivitet, men i modsætning til tidligere tiders statistiske undersøgelser, der typisk var en enkeltstående affære, er DM beregnet til at indgå permanent i styringsprocessen. Dette sammen med muligheden for omfattende flerstrengede analyser af datamaterialet giver et langt mere facetteret og tæt nutidsbillede af virkeligheden og dermed også mange gange større sandsynlighed for at resultatet er brugbart – dvs. værdiskabende. De ofte relativt tynde sikkerhedsmarginaler stiller naturligvis skrappe krav til overvågningen af modellernes troværdighed og effektivitet og til en løbende opdatering.

Omkostningseffektiviteten er en nødvendig, men dog ikke tilstrækkelig forudsætning – den skal kombineres med et erkendt skifte i resultatevalueringen. Ændringen fra den videnskabelige forståelse og tolkning af et givet analyseresultat til en evaluering ud fra en økonomisk synsvinkel har bogstaveligt talt og i overført betydning rykket grænser. Accepten af, at en merindtægt, der overstiger meromkostningen ved vidensproduktionen, er tilstrækkelig legitimering af en analyses resultatmæssige kvalitet, bevirker at mange flere analyser vil være handlingsskabende og dermed potentielt værdiskabende. Medens videnskaben vurderer resultatet ud fra idealet 100 % korrekt er udgangspunktet for DM det aktuelle informationsniveau i virksomheden.

Ud fra traditionelle videnskabelige præmisser og kriterier kan DM i bedste fald betegnes som pragmatisme og i værste fald som krystalkuglekigning. Meget populært kan man sige at DM med den store vægt, der lægges på gennemsnit, overrepræsentation, koncentration, mønstre blot er en (videnskabelig?) metode til at danne fordomme og tommelfingerregler for beslutningstagere, som ikke har fået dem ind med mesterlæren. 'Leveregler', 'måder', ' faglige normer' o. lign., der tidligere krævede generationers erfaringer for at kunne blive formuleret, kan nu ekstraheres i brøkdele af et år.

Strukturnøgletal – et eksempel

Nærværende eksempel på en manuelt gennemført analyse baseret på et DM-lignende koncept er baseret på dagbladet Børsens Top 500 oversigt over de 500 største virksomheder i Danmark. Udover navnet på virksomheden samt en branchekode indeholder oversigten tillige nogle få centrale regnskabsdata med omsætning, balance og antal ansatte som de mest interessante. Det centrale spørgsmål i nærværende sammenhæng er: *Indeholder dataene brugbar information udover den umiddelbare der ligger i tallenes størrelse* (og som af dagbladet Børsen bl.a. er anvendt til at rangordne selskaberne efter – størst overskud, flest ansatte, største omsætningsstigning o. lign.) og i givet fald: *er det muligt at tappe den – gøre den tilgængelig og brugbar?* Dette er i al væsentlighed problemstillinger som DM forsøger at løse.

Flere steder i litteraturen diskuteres om miningen skal foregå direkte i databasen eller om der skal laves et i udtræk til en flad fil (regnearksfil). Ved virkelig store datamængder anses den direkte mining for eneste mulighed, mens andre fremhæver, at den flade fil giver mulighed for en tæt procesovervågning, som dels kan inspirere til yderligere analyser, dels målrette analysen mere effektivt end den rene datastyring kan. Analysen her er udført i Excel regneark og hovedsagelig under anvendelse af statiske metoder hentet fra "Student CD" fra Bowerman, O'Connell og Hands "Business Statistics in Practice"

I DM vil det ofte være således at man knap nok aner hvilken information man søger efter og da slet ikke hvilken form den optræder i. Selvom informationen ligger skjult i dataene vil søgningen efter den dog sædvanligvis ikke foregå helt i blinde. Tidligere erfaringer, tilfældige bemærkelsesværdige observationer, teoretiske overvejelser og lignende kan danne baggrund for en søgestrategi, der selvom den ikke er direkte målrettet, dog er metodisk og systematisk tilpasset situationen.

Udgangspunktet for nærværende analyse kan findes i stort set alle elementære økonomibøgers kapitel 1- "En virksomhed" - om end de senere års markante pensumreduktioner stort set har fjernet enhver substans i emnet. I Waarst m.fl. "Regnskabslære /Driftsøkonomi" fra 1988 er der givet en set med nutidige briller grundig og omfattende gennemgang af virksomhedsbegrebet herunder en opdeling af virksomhederne i forskellige typer og at denne opdeling bl.a. afspejler at de økonomiske forhold og at styringen er forskellig i de forskellige virksomhedstyper. Den traditionelle typologisering – industri opdelt i undergrupperne ordre-, serie- og procesproducerende, handel bestående af detail og engros samt servicevirksomheder – vil være et pejlemærke for nærværende DM.

Hvis ovenstående præmis – at virksomhedstype og økonomi er tæt forbundne – holder, må det også påvirke i det mindste strukturnøgletallene. Af de 500 datasæt beregnedes derfor følgende nøgletal: omsætning/balance, (aktivernes omsætningshastighed), omsætning/antal medarbejdere (omsætning pr. medarbejder) og balancesum/antal medarbejdere (investering pr. medarbejder) for hver virksomhed. 27 datasæt var mangelfulde og blev smidt ud af analysen, medens virksomhederne med de 40 mest ekstreme værdier blev holdt udenfor analysen. For god ordens skyld nævnes, at der ikke er foretaget kontrol af eller rettelser i Børsens data.

Børsen giver ingen forklaring på branchetilhørsforholdet, men et blik ned over de i alt 21 brancher (se tabel 1) lader formoder er der er tale om en skøn sammenblanding af juridiske, markeds-, funktions- og produktionsmæssige forhold.
Beregning af gennemsnit og spredning af de 3 nøgletal for hver af de 21 brancher viser klart, at der ikke i de brancheopdelte nøgletal kan spores et branchetilhørsforhold. Gennemsnittene varierer klart over de forskellige brancher, men spredningen gør det i de fleste tilfælde umuligt at fastslå om gennemsnittene er forskellig fra branche til branche. En økonomisk typificering af selskaberne kan derfor ikke gøres ud fra branchebetegnelsen.

Tabel 1: Virksomheder fordelt på branchekoder

BILH	19
BYGG	32
DETH	29
ENER	20
ENGH	58
ENTR	13
ITEL	62
JERN	15
KEMI	11
KONG	15
LEVN	57
MASK	27
MEDI	21
MEFO	17
MØBL	8
PAPI	12
REDE	9
SERV	19
TEKS	6
TRAN	18
UNDE	5
IALT	**473**

Kilde: Børsen top500

Infouddragning – de individuelle nøgletal

Med henblik på at afgøre om der var visse koncentrationer af nøgletallene gennemførtes først en frekvensanalyse. Selv ved forholdsvis snævre intervaller var der ikke tydelige ujævnheder i fordelingerne for de enkelte nøgletal, som kunne indikere visse koncentrationer i observationsmaterialet. Derfor sorteredes nøgletallene efter størrelse og frekvensfordelingen af førstedifferencerne viste en tydelig overvægt i de nederste og i de øverste intervaller. Overhyppigheden i de øverste intervaller kan henføres til at nøgletallene er ret udpræget højreskæve fordelt, men overvægten i de nederste intervaller må skyldes at mange observationer ligger meget tæt på hinanden. Dette bekræftedes ved en beregning af en glidende standardafvigelse over 20 og 30 observationer på de størrelses-sorterede nøgletalsværdier. Udviklingen i den glidende standardafvigelse var som forventet nærmest bølgeagtig med markante forskelle mellem bund- og topværdier – også når der sås bort fra 80 største værdier dvs. de værdier der ligger ude i fordelingens højre hale. Altså en klar indikation af atder var tydelige koncentrationer i datamaterialet og dermed også en ledetråd for den fortsatte analyse.

Sammenholdtes disse koncentrationer med Børsens branchebetegnelse kunne 4 grupper identificeres umiddelbart: én gruppe med en overrepræsentation af selskaber i JERN og MASK-branchen, en anden gruppe, hvor MEDI og KEMI-selskaberne var i overtal, en tredje gruppe domineret af ENGH-virksomheder og en fjerde gruppe bestå-ende af SERV-virksomheder. Desuden var der en indikation på at BYGG og ENTR-selskaberne kunne skilles ud. De 2 største brancher ifølge Børsen – ITEL (Information og Telekommunikation) samt LEVN - kunne ikke spores. Da denne gruppering kun er baseret på koncentrationer af nøgletallene enkeltvis plus Børsens branchekode, er der tale om en meget simpel og grov opdeling.

Infouddragning – kombination af nøgletal

Med henblik på at afdække betydningen af samspillet mellem de 3 nøgletal dannedes derfor 4 grupper med udgangspunkt i de nævnte brancher samt 4 grupper med udgangspunkt i BYGG/ENTR-, ITEL-, LEVN-, og DETH-selskaberne. Sidstnævnte 4 grupper medtoges dels af nysgerrighed, dels som kontrolgrupper. Hver gruppe bestod af 15 nærmeste naboer (k-NN, k

–nearest neighbbors) – dvs. de 15 selskaber, der lå tættest sammen målt som summen af deres kvadrede afstand fra gennemsnittene målt i standardafvigelser[5] - altså

$$Min D_k = \sum_{1}^{3} \frac{(\bar{x}_i - x_{ij})^2}{s_i^2}$$

Min D beregnedes gennem en iterativ proces, som først beregner gennemsnittene for de 3 nøgletal for alle selskaber i gruppen samt hver enkelt selskabs akkumulerede afstand hvorefter selskabet med størst afstand findes. I andet gennemløb udgår dette selskab af beregningen. Det giver 3 nye gennemsnit og nye afstande for alle selskaber incl. det udeladte selskab og selskaberne med de 2 største afstande udpeges. I 3. gennemløb udelades disse 2 selskaber, nye gennemsnit og afstande beregnes, 3 fjerneste selskaber udpeges o.s.fr. Efter N-15 gennemløb vil de resterende 15 selskaber være de 15 selskaber, der ligger tættest på hinanden og det antages at de giver et tilfredsstillende estimat på gruppens tyngdepunkt..

Med udgangspunkt i disse 8 fixpunkter beregnedes vha. ovenstående udtryk alle øvrige selskabers afstande til hver enkelt gruppe. I første omgang bestemtes gruppetilhørsforholdet blot som den korteste afstand, men det resulterede i nogle ret diffuse og konturløse grupper. Eksempelvis sugede JERN-MASK-gruppen næsten 25 % af alle selskaber til sig uden at det styrkede forventningen om, at der kunne identificeres en jernindustriel klynge i datamaterialet. Med henblik på at få en noget skarpere profil frem skærpedes kravet til de øvrige selskabers gruppetilhørsforhold til kun at omfatte de selskaber, som faldt indenfor det rum som gruppens 15 selskaber udspændte og det gav bonus. Selv med denne skærpede betingelse for et gruppetilhørsforhold næsten 4-dobledes JERN-MASK-gruppen, men der var nu et fælles kendetegn for stort set alle selskaber – nemlig at de var serieproducerende virksomheder. Gruppen omdøbtes derfor til PROD-gruppen.

En tilsvarende udvikling sås omkring KEMI-MEDI-gruppen – den opsugede bryggerierne, sukker- og oliefabrikker m.fl. Tydeligvis var KEMI-MEDI-gruppen samlingspunkt for Procesindustrien, PROC. ENGH-gruppen, der var den næststørste, omfattede selskaber fra alle brancher, som beskæftigede sig med import og engroshandel. Den mindste gruppe var SERV-virksomheder med ca. 5 % tilslutning fra forskellige brancher. Måske bortset fra BYGG/ENTR-gruppen, der lidt klarere manifesterede sig som en entreprenørgruppe (ordreproducerende industri eller Business to Business), var den ingen entydig samling omkring de sidste 4 grupper.

Infouddragning – sammenvejning af nøgletal

Indtil nu har hele grupperingsarbejdet været baseret på en ligevægtning af de 3 nøgletal og at denne ensvægtning ville give et tilfredsstillende udgangspunkt for en beregning, der kunne klassificere de enkelte virksomheder i forskellige typer svarende til lærebøgernes opdeling. Som nævnt ovenfor var dette ikke umiddelbart tilfældet. Den anvendte grupperingsteknik, k-NN, bevirker, at det rum, som de 3 variable udspænder stort set bliver helt symmetrisk målt i standardafvigelser, men er det nu også tilfældet i virkelighedens verden? Kunne man ikke forestille sig, at en høj omsætning pr. medarbejder skyldtes en stor investering pr. medarbejder eller at en lav omsætningshastighed for aktiverne kompenseredes af en høj omsætning

[5] Da standardafvigelsen for de 3 nøgletal var noget forskellige anvendtes de relative afstande fra gennemsnittet, idet der herved undgås at beregningsalgoritmen først eliminerer selskaber med store absolutte afstande

pr. medarbejder? Hvis det er tilfældet vil den ovenfor anvendte symmetriske gruppedannelse uheldigvis lige netop skære sådanne virksomheder væk.

Et andet problematisk forhold er, om det centrum, som de 15 oprindelige selskaber i grupperne konstituerer, nu også er det sande centrum for de pågældende grupper? Hvis det beregnede centrum afviger systematisk fra det sande centrum, vil det medføre en fejlagtig afskæring af selskaber i koncentreringsberegningerne. Det ville derfor være ønskeligt om det rum, som de forskellige grupper falder i, kunne afgrænses mere frit i forhold til det oprindelige centrum og i forhold til de mere individuelle forskelle i nøgletallene såvel indenfor gruppen som i forhold til de øvrige grupper.

En lille allegori: Da der – tilfældigvis - er tale om 3 nøgletal kan hver gruppe opfattes som 3-dimensionale legemer med ukendte men sandsynligvis forskellige former. Én gruppe kunne være appelsinformet, en anden pæreformet, en tredje bananformet, en fjerde på størrelse med en vindrue og en femte på størrelse med en melon. Da alle 'frugterne' ligger i den samme frugtskål er det ønskeligt at der kunne lægge nogle 'snit' ind, der maksimerer sandsynligheden for at man får 'appelsin' for sig og 'banan' for sig o.s.fr. Med nedennævnte model skæres nogle fuldstændig lige snit ind i 'frugtkurven' og med de forskellige former in mente vil det helt sikkert medføre at der 'hugges en hæl og klippes en tå', men at kernen forbliver intakt. Det er muligt at en 'krumkniv' ville være et bedre redskab, men hvilken krumning skulle så vælges?

Diskriminantanalysen er en velegnet teknik til klassifikation når inputvariablerne er kontinuerte. Metoden beregner med hvilken vægt de enkelte variable skal indgå for at opnå den bedst mulige grænsedragning mellem 2 eller flere grupper. Antallet af variable og observationer er i nærværende tilfælde dog for lille til simultan opdeling i mere end 2 grupper og derfor anvendtes en additiv model til at bestemme den bedste afskæringsværdi mellem grupperne 2 og 2, altså

$$d = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + e$$

hvor x_1, x_2 og x_3 er de 3 nøgletal, der indgår i analysen og hvor d er en diskret variabel, der antager værdien 0 når selskabet tilhører gruppe 1 og værdien 1 når selskabet tilhører gruppe 2. Ved hjælp af mindste kvadraters metode – multipel lineær regression – kan vægtene b bestemmes. Da d kun antager værdierne 0 eller 1 vil algoritmen beregne vægte, b_1, b_2 og b_3 , som resulterer i at selskaber i gruppe 1 får værdier, der spreder sig rundt omkring 0 og selskaber i gruppe 2 får værdier, der spreder sig rundt omkring 1 med en risiko for en vis overlapning. Der må derfor også fastlægges en afskæringsværdi som deler intervallet fra 0 til 1 i to, således at når d < afskæringsværdien tilhører selskabet gruppe 1 og d > afskæringsværdien er det et selskab tilhørende gruppe 2 Hvis residualleddet, e, er N(0,1) fordelt kan den bedste afskæringsværdi bestemmes som den gennemsnitlige $\overline{d}_1$ for selskaberne i gruppe 1 og den gennemsnitlige $\overline{d}_2$ for selskaberne i gruppe 2 vægtet med antallet af observationer i de 2 grupper, dvs.

$$d_{opt} = \frac{n_1\overline{d}_1 + n_2\overline{d}_2}{n_1 + n_2}$$

hvor n_1 og n_2 er antal selskaber i hver af de 2 grupper. Hvis standardafvigelsen på residualleddet for de 2 grupper er forskellige opnås dog en mere præcis afskæring ved også at vægte gennemsnittene d_1 og d_2 med standardafvigelserne:

$$d_{opt} = \frac{s_{e1} n_1 \bar{d}_1 + s_{e2} n_2 \bar{d}_2}{s_{e1} n_1 + s_{e2} n_2}$$

hvor s_{e1} og s_{e2} er standardafvigelsen på restleddet. Sidstnævnte afskæring gav i almindelighed det bedste resultat – dvs. færrest fejlklassificerede selskaber.

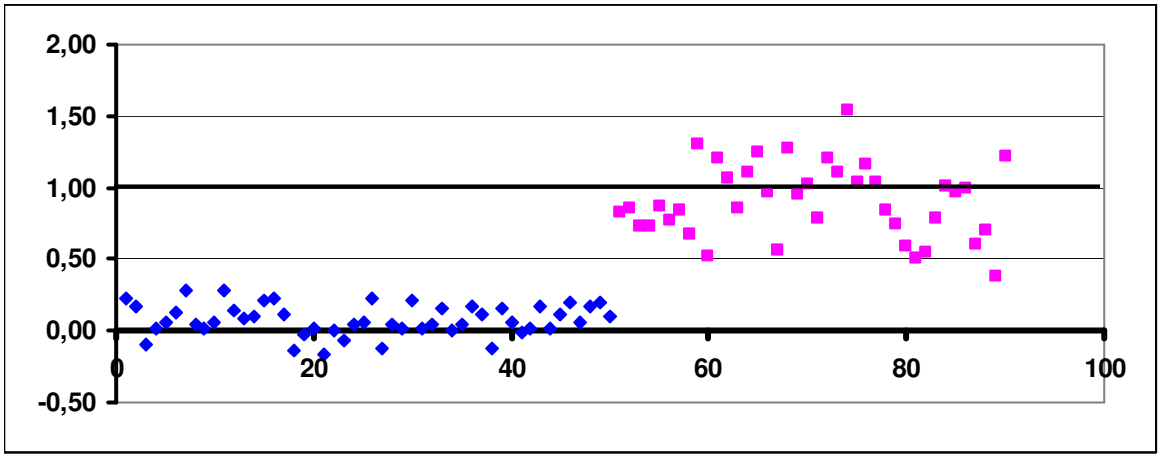
Som nævnt var det kun muligt at bestemme vægtene og afskæringsværdier for 2 grupper ad gangen og med 8 forskellige grupper vil det resultere i 28 beregninger. Da det var noget uklart hvordan de enkelte grupper lå i forhold til hinanden og da hver gruppe skulle afgrænses mod 7 andre og helst så klart som muligt og da den samme stikprøve skulle anvendes i alle 7 beregninger valgtes nogle ret 'snævre' stikprøver ud af de ovenfor nævnte grupper – nemlig de 50 nærmeste naboer i PROD, 40 nærmeste naboer i ENGH og de 20 nærmeste naboer i de øvrige grupper bortset fra SERV som kun omfattede 18 selskaber. Valget af nærmeste naboer sikrer at afstanden mellem de enkelte grupper er størst mulig og dermed også størst sandsynlighed for at grupperne kan adskilles, men omvendt betyder det også at stikprøvens variabilitet er ret begrænset og at dens repræsentativitet for gruppen som helhed er mindre end sædvanlig. Konsekvensen vil være at klassifikationsevnen i stikprøven vil overestimere den faktiske opdelingsformåen i hele populationen.

For PROD – ENGH kan diskriminantfunktionen estimeres til:

d = -1,44 + 0,72*AOH – 0,27*oms/medarb. + 1,11*inv/medarb.

med en korrelationskoefficient på 0,89 og med vægte, der alle er signifikante med p-værdier på under 0,05. Rent statistisk er det udtryk for at der er en meget klar forskel mellem de 2 grupper PROD – ENGH. Disse statistiske mål er dog af begrænset interesse, idet de ikke umiddelbart kan relateres til opgaven: at afgrænse grupperne fra hinanden.

Figur 1 Diskriminantværdier for PROD - ENGH



Scatterdiagrammet figur 1 viser de beregnede d-værdier for de i alt 90 selskaber i de 2 grupper. Det ses at produktionsvirksomhedernes d-værdier fordeler sig rundt 0 og engroshandelens omkring 1. Gennemsnittene kan beregnes til $\bar{d}_{PROD}$ = 0,08 og $\bar{d}_{ENGH}$ = 0,90, altså en afstand på 0,82. Da standardfejlen på residualleddet e, S(e), kan beregnes til 0,208 kan den relative afstand beregnes til 3,98. Afstanden målt i standardafvigelser er det bedste generelle udtryk for funktionens diskriminations-evne, idet det er sammenligneligt over alle beregnin-

ger. Den optimale afskæringsværdi kan beregnes til 0,367 og da min(d_{ENGH}) = 0,386 og max(d_{PROD}) = 0,284 ses det at den giver en perfekt klassifikation af de 90 selskaber. Denne fuldstændige tvedeling er dog et resultat af stikprøveudvælgelsen og er opnået på bekostning af modellens generalitet, idet en afstand på 3,98 standardafvigelser i en normalfordelt population ville resultere i at ca. 4 % af observationerne – svarende til ca. 4 selskaber - ville blive fejlklassificeret.

I tabel 2 er vist regressionskoefficienter og afstande målt i standardafvigelser for alle 28 diskriminantfunktioner. I en normalfordelt population vil en afstand på 2 standardafvigelser resultere i at ca. 1 ud af 3 observationer vil blive fejlklassificeret, men den snævre stikprøveudvælgelse resulterer i en væsentligt bedre diskrimination mellem stikprøverne.

Tabel 2: Regressionskoefficienter og afstande for de 8 grupper

		PROD	ENGH	PROC	SERV	BYGG	DETH	ITEL	LEVN
PROD	b_0 =		-1,44	-1,04	0,37	-1,80	-1,10	-1,51	-1,60
	b_1 =		0,72	0,32	0,65	1,23	0,41	0,80	0,81
	b_2 =		-0,27	-0,06	-1,28	-0,23	0,18	-0,11	-0,08
	b_3 =		1,11	0,93	0,24	0,89	0,64	0,92	0,99
	Afst		3,98	4,09	2,13	1,24	2,62	3,07	2,66
ENGH	b_0 =			-0,14	2,04	2,42	0,31	0,85	0,81
	b_1 =			0,20	-0,57	-0,62	0,30	0,05	0,16
	b_2 =			-0,44	0,32	0,24	-0,52	-0,37	-0,48
	b_3 =			0,81	-1,19	-1,19	0,57	0,29	0,35
	Afst			3,10	4,09	2,48	1,22	1,20	1,22
PROC	b_0 =				0,99	1,21	0,98	-0,22	1,06
	b_1 =				0,15	0,08	-0,06	0,58	-0,11
	b_2 =				-0,20	0,09	0,34	-0,14	0,47
	b_3 =				-0,41	-0,72	-0,67	0,18	-0,81
	Afst				5,44	3,75	1,59	0,59	1,86
SERV	b_0 =					-1,12	-0,63	-0,47	-0,74
	b_1 =					0,52	0,28	0,22	0,37
	b_2 =					0,02	0,14	0,23	0,19
	b_3 =					1,43	0,63	0,43	0,60
	Afst					1,76	2,95	2,62	2,69
BYGG	b_0 =						-1,18	-0,82	-1,49
	b_1 =						0,48	0,30	0,71
	b_2 =						-0,08	0,11	-0,21
	b_3 =						0,98	0,63	1,17
	Afst						1,32	1,18	0,98
DETH	b_0 =							0,72	1,33
	b_1 =							-0,18	-0,32
	b_2 =							0,08	0,24
	b_3 =							0,02	-0,52
	Afst							0,31	0,18
ITEL	b_0 =								0,74
	b_1 =								0,03
	b_2 =								0,00
	b_3 =								-0,25
	Afst								0,23

Hvor b_1 er koefficienten til AOH, b_2 er koefficienten til oms/medarb., b_3 er koefficienten til inv/medatb.

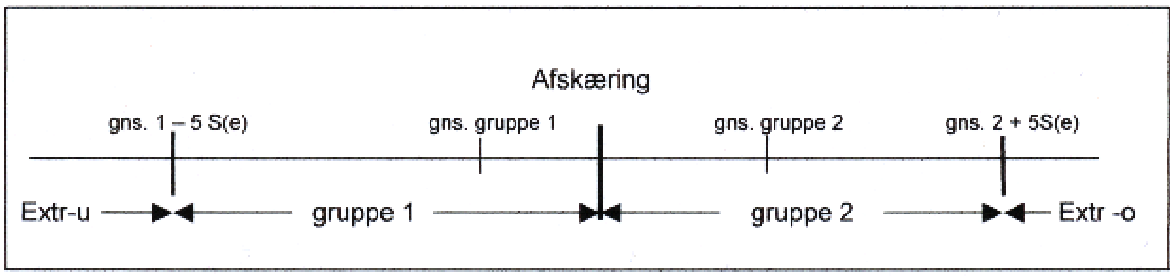
Note: farverne angiver signifikansniveau: skarp gul p < 0,01 og lys gul p< 0.05

Bortset fra BYGG-gruppen er produktionsvirksomhederne med de 7 diskriminantfunktioner relativt vel defineret vis-a-vis de øvrige 6 grupper, idet afstandene til de øvrige grupper ligger på 2,13 til 4,09 standardafvigelser. Den store lighed mellem nøgletallene for BYGG og PROD skal antagelig søges i det forhold at flere BYGG-virksomheder i højere grad er produktionsvirksomheder end de er entreprenørvirksomheder – dvs. ordreproducerende. Som forventet er der en meget klar forskel mellem nøgletallene for grupperne PROD, ENGH, PROC og SERV, medens afgrænsningen de 4 øvrige grupper er ret begrænset og mellem de sidste 4 grupper indbyrdes ikke er nogen konstaterbar forskel.

Gruppering af selskaberne

Diskriminantfunktionernes effektivitet er testet på alle 433 selskaber. Principielt kan de opstillede diskriminantfunktioner kun tvedele observationsmaterialet i 2 klasser – enten gruppe 1 eller gruppe 2 afhængig af om d er mindre end eller større end den beregnede afskæringsværdi. Resultatet var imidlertid at en række selskabers d-værdier var meget store eller meget små – dvs. det måtte antages at disse selskaber faldt helt udenfor det rum de 7 diskriminantfunktioner afgrænser rundt omkring hver gruppe. Derfor opdeltes udfaldsrummet for d i 4 klasser, nemlig extreme-u, gruppe 1, gruppe 2 og extreme-o. jfr. fig. 2.

Figur 2: d´s udfaldsrum



Under og overgrænsen for ekstremværdierne blev fastsat ret arbitrært til gennemsnit gruppe 1 minus 5 gange standardfejlen på estimatet og tilsvarende for overgrænsen. I en normalfordelt population ville 3 S(e) have været tilstrækkelig, men på baggrund af at de snævre stikprøver, der dannede grundlag for estimeringen af diskriminantfunktionerne, må det antages at den beregnede S(e) underestimerer den sande fejl på e.

For at afgøre om et selskab tilhørte eksempelvis produktionsgruppen må der gennemføres 7 tests vis-a-vis alle de øvrige grupper. Som følge heraf kunne et selskab maximalt 7 gange blive klassificeret i én og samme gruppe 1. På baggrund af testene opstilledes en såkaldt konfusionsmatrice, der viser hvor mange gange det enkelte selskab blev klassificeret som gruppe 1 selskab, som gruppe 2 selskab o.s.fr. Resultatet af denne klassifikation er vist i bilag A. Her angiver en gruppebetegnelse med store bogstaver at selskabet i 7 ud af 7 tilfælde er klassificeret i den pågældende gruppe, stort begyndelsesbogstav: at selskabet i 6 ud af 7 gange er klassificeret i gruppen og 4 små bogstaver: at selskabet i 5 ud af 7 tilfælde er klassificeret i gruppen.

Af de i alt 433 selskaber blev de 133 selskaber klassificeret som produktionsselskaber – i 92 tilfælde entydigt, medens de resterende 41 selskaber havde 1 eller 2 gruppe 2 klassifikationer. 78 selskaber blev i fem eller flere tilfælde klassificeret som extreme, medens 30 virk-

somheder blev klassificeret i mindst 4 forskellige grupper, hvorfor deres gruppetilhørsforhold ikke kunne bestemmes. Som forventet er der ingen økonomisk fællesnævner for Informations og Telekommunikationsgruppen (ITEL) – kun 8 ud af 62 selskaber kan med en vis velvilje siges at udvise nogle meget beskedne økonomiske fællestræk. Overraskende nok fandtes en økonomisk set ret homogen gruppe i levnedsmiddelgruppen centreret omkring slagterierne / landbrugsforarbejdende virksomheder. En umiddelbar hypotese ville være, at det var en 'andelsselskabsfaktor', der her var trukket ud af dataene, men da flere selskaber aldrig har været andelsselskaber, må homogeniteten antagelig være forårsaget af det ensartede virkefelt og konkurrenceforholdene. DETH-selskaberne – i alt 13 er klassificeret som sådan – er på trods af klassifikationen økonomisk set meget forskellige.

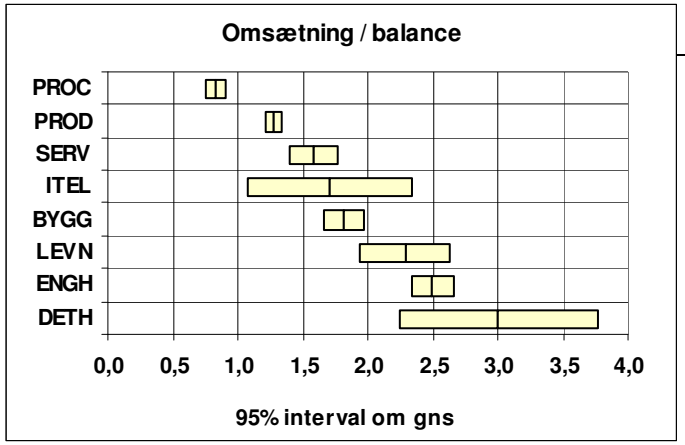
I alt 49 virksomheder er blevet klassificeret som procesvirksomheder. En nøjere gennemgang af PROC-selskaberne viser imidlertid, at nok afgrænser de 7 diskriminantfunktioner procesvirksomhederne, men det fælles kendetegn for alle virksomheder i denne gruppe er kapitalintensive virksomheder. Til servicegruppen er henregnet i alt 19 selskaber, men heraf hører kun syv til den SERV-branche, som Børsen har udarbejdet – 12 selskaber er altså henregnet til andre brancher af Børsen.

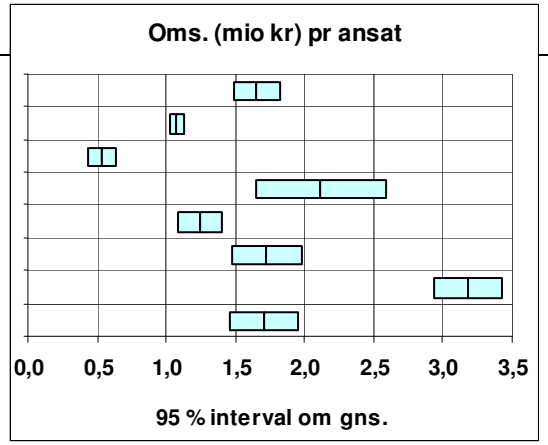
Ved en gennemgang af bilag A vil det ses at den statistiske typifisering af de 433 virksomheder ikke er fejlfri, men det er egentlig ikke så væsentlig i denne forbindelse. Det væsentlige er at dataene – den kategoriale branchekode og de kontinuerte: omsætning, balancesum og antal ansatte - i sig selv indeholder en ikke-synlig information, som gør det muligt, at foretage en sådan ret entydig klassifikation samt naturligvis implikationen af denne klassifikation.

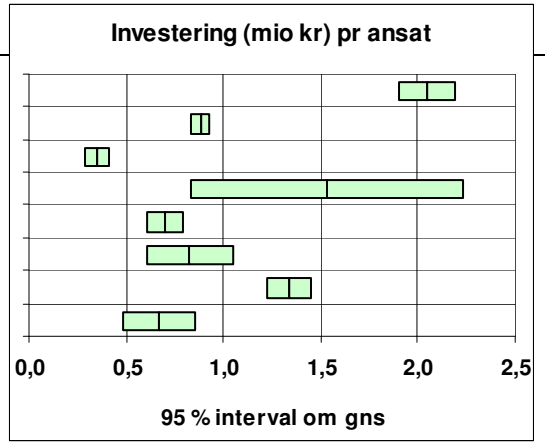
Resultatet

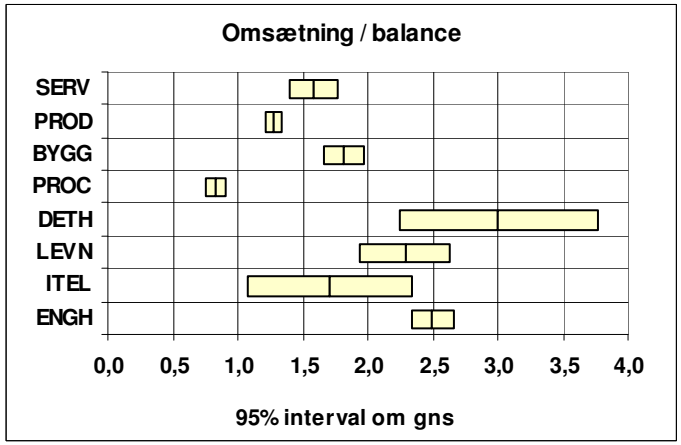
I figur 3 næste side er vist det egentlige mål med denne data mining – nemlig hvordan kan den typiske produktionsvirksomhed, den typiske servicevirksomhed osv. karakteriseres ud fra nogle centrale økonomiske data. For hver virksomhedstype er beregnet gennemsnittet af aktivernes omsætningshastighed, omsætning pr. ansat og investering pr ansat samt et 95 % konfidensinterval. I øverste tredjedel er de forskellige virksomhedstyper sorteret efter aktivernes omsætningshastighed, i midterste tredjedel efter omsætning pr. ansat og i nederst efter investering pr. medarbejder.

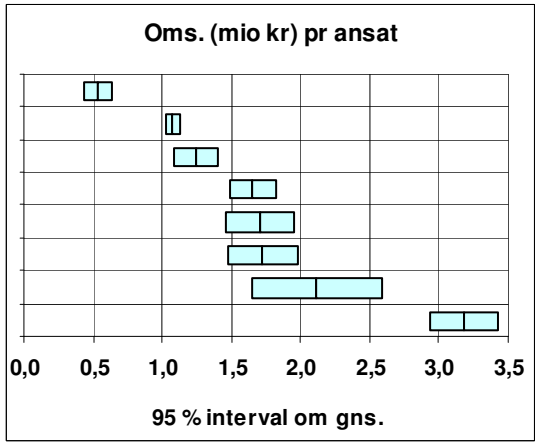
Det ses, at i procesindustrien omsættes aktiverne uhyre langsomt – ca. 0,8 gange pr år. Det meget snævre konfidensinterval indikerer, at det er et gennemgående og meget karakteristisk vilkår for hele industrien. Aktivernes omsætningshastighed er også særdeles entydig i den serieproducerende industri, men hastigheden er 50 % højere end i procesindustrien – mao. procesindustrien skal investere 1,25 kr. for at få 1 krones omsætning medens den mere traditionelle industri klarer det for ca. 0,80 kr.! Det har naturligvis en meget klar konsekvens for styringen af den anden betydelig ressource – nemlig medarbejderne. Medarbejdereffektiviteten – målt som omsætning pr medarbejder – er da stort set også omvendt proportional med aktivernes omsætningshastighed – i procesindustrien er omsætningen ca. 1,7 mio. kr. pr medarbejder medens serieindustrien ligger så lavt som 1,1 mio. kr. Den meget store forskel i investeringsomfanget i de to industrier – i gennemsnit 2,1 mio. kr. pr medarbejder i procesindustrien mod beskedne 0,8 mio. kr. i serieindustrien – kompenseres derimod i udpræget grad gennem den opnåede bruttoavance på de solgte produkter.

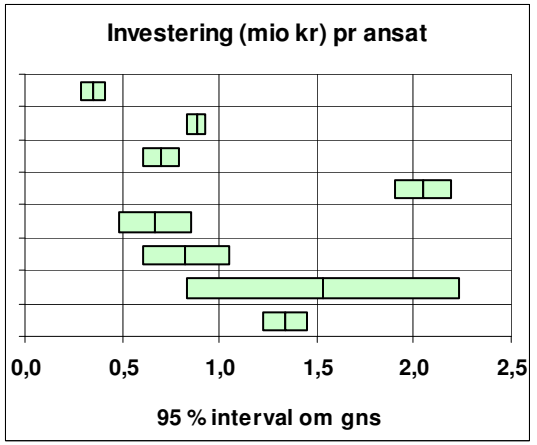

Omsætning / balance
PROC
PROD
SERV
ITEL
BYGG
LEVN
ENGH
DETH
0,0 0,5 1,0 1,5 2,0 2,5 3,0 3,5 4,0
95% interval om gns

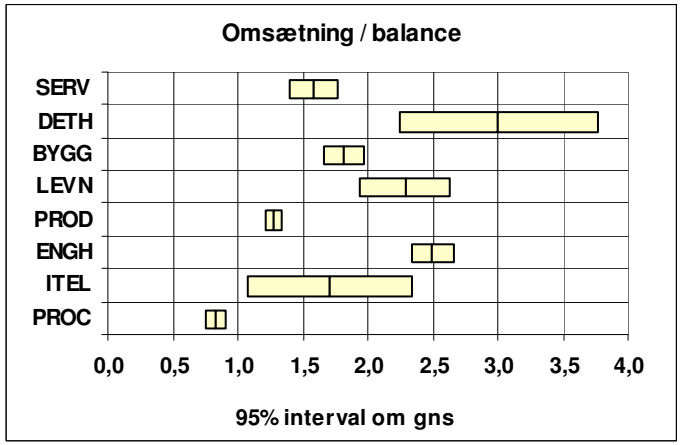

Oms. (mio kr) pr ansat
0,0 0,5 1,0 1,5 2,0 2,5 3,0 3,5
95 % interval om gns.

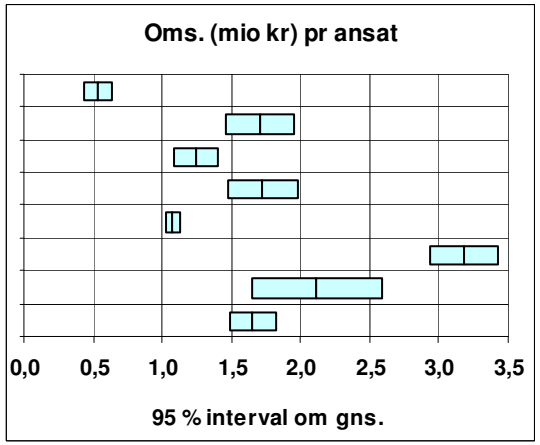

Investering (mio kr) pr ansat
0,0 0,5 1,0 1,5 2,0 2,5
95 % interval om gns

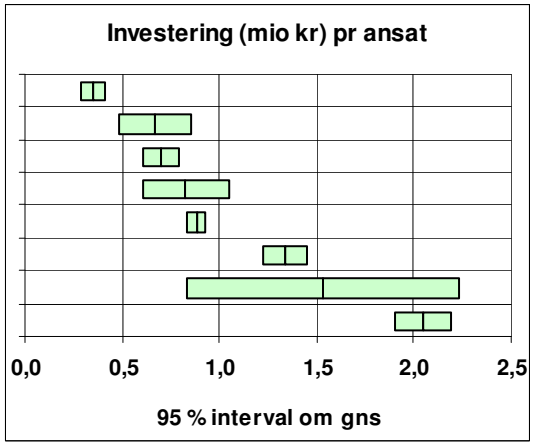

Omsætning / balance
SERV
PROD
BYGG
PROC
DETH
LEVN
ITEL
ENGH
0,0 0,5 1,0 1,5 2,0 2,5 3,0 3,5 4,0
95% interval om gns


Oms. (mio kr) pr ansat
0,0 0,5 1,0 1,5 2,0 2,5 3,0 3,5
95 % interval om gns.


Investering (mio kr) pr ansat
0,0 0,5 1,0 1,5 2,0 2,5
95 % interval om gns


Omsætning / balance
SERV
DETH
BYGG
LEVN
PROD
ENGH
ITEL
PROC
0,0 0,5 1,0 1,5 2,0 2,5 3,0 3,5 4,0
95% interval om gns


Oms. (mio kr) pr ansat
0,0 0,5 1,0 1,5 2,0 2,5 3,0 3,5
95 % interval om gns.


Investering (mio kr) pr ansat
0,0 0,5 1,0 1,5 2,0 2,5
95 % interval om gns

Opgjort på denne måde kommer servicevirksomhederne noget uventet på 3. pladsen som den mest kapitalintensive virksomhed med ca. 0,60 kr. investering pr. omsætningskrone. Der er en noget større variabilitet i serviceselskaberne mht. aktivernes omsætningshastighed, idet konfidensintervallet er flere gange større end i den serieproducerende industri – 0,38 mod 0.12. Da der ingen overlapning er mellem de 2 konfidensintervaller, er det ensbetydende med, at den gennemsnitlige AOH statistisk set er klart større i serviceindustrien end i serieproduktionen, jfr. at b_1 = 0,65 i tabel 2 er markeret let skraveret, hvilket indikerer, at der på 5 %-niveauet er en signifikant forskel mellem omsætningshastighederne i de to virksomhedstyper. Den kunstige gruppe ITEL kan overhovedet ikke karakteriseres ud fra de givne nøgletal. Derimod er BYGG / ordreproducerende industri ret veldefineret ved de givne nøgletal, idet konfidensintervallet størrelsesmæssigt svarer til service og engroshandel. Det ses at der er en vis overlapning af konfidensintervallerne for SERV og BYGG, hvorfor dette nøgletal ikke er særlig pålidelig ved en skelnen mellem de 2 grupper.

Som tidligere nævnt er der et vist statistisk belæg for en levnedsmiddelgruppe primært identificeret ved en høj omsætning af aktiverne - 2,3 gange pr. år – hvilket relativt klart adskiller den fra PROC, PROD, SERV og til dels BYGG medens den stort set er sammenfaldende med ENGH – jfr. den næsten fuldstændige overlapning af konfidensintervallerne. Dog er der en markant forskel mellem de to virksomhedstyper når der ses på omsætning pr. ansat – 1,7 mio. kr. mod 3,2 mio. kr. – og på investering pr. ansat – 0,9 mio. kr. mod 1,4 mio. kr. Rent statistisk kan levnedsmiddelindustrien og detailhandelen ikke adskilles - der er næsten 100 % overlapning af de 2 gruppers konfidensintervaller – men man er næppe i tvivl når man står foran forretning eller et slagteri. Konfidensintervallet er dog ret stort for alle 3 nøgletal, hvilket implicerer en noget mindre homogenitet i gruppen.

Ikke uventet har handelsvirksomhederne den hurtigste omsætning af aktiverne – 2,5 gange pr. år i engros og 3 gange i detailhandelen. Men spredningen på detailhandelens nøgletal gør det helt uanvendelig som determinant for denne gruppe. Et (godt?) gæt på årsagen til denne næsten absurde store spredning for iøvrigt ret ensartede virksomheder kunne være, at nogle detailhandelsvirksomheder udfører visse engrosfunktioner som f. eks. import og finansiering.

Omsætningen pr. medarbejder – i gennemsnit 1,7 mio. kr. for de i alt 355 virksomheder, der er fordelt på de 8 grupper – varierer fra godt en halv million pr. servicemedarbejder til 3,2 mio. kr. i engroshandelsvirksomheder. Nogenlunde samme spredning er der på investeringerne pr. ansat – fra 0,35 mio. kr. i service til 2,1 mio. kr. i procesindustrien. Flertallet af virksomheder ligger på omkring trekvart mio. kr. pr. næse. At engroshandelen ligger på næsten det dobbelte må antagelig henføres til den betydelig finansiering af varehandelen, der ofte anføres som én af de væsentligste opgaver for denne type virksomheder.

Anvendelsen af resultatet

Og udbyttet af denne data mining? Tjaaaaa.....og dog. Først og fremmest at dataene som sådan indeholder mere information end deres absolutte værdier, og at det under anvendelsen af forskellige værktøjer er muligt at uddrage en sådan ikke-synlig information, nemlig en gruppering / typificering af erhvervsvirksomheder på en måde, der stemmer overens med den traditionelle lærebogs systematisering af virksomhedsbegrebet. Den praktiske anvendelighed - der er det andet krav til en vellykket DM - er i modsætning til informations ekstraheringen specifik og konkret. Tallene kan anskues som en form for samhørende type-karakteristiske normtal, der i den konkrete situation - analyse / vurdering - giver en ret-

tesnor for en klassificering af en virksomhed ud fra en begrænset, men lettilgængelige datamængde.

Tabel 3: Virksomhedstypernes karakteristiske nøgletalsværdier

Nøgletal / Virksomhedstype	Omsætning pr. balancekr.	Omsætning pr. medarb, mio. kr.	Investering pr medarb. mio. kr.
Serieprod. industri	1,28	1,07	0,88
Procesprod. industri	0,83	1,65	2,05
Byggeri afh. industri	1,81	1,24	0,70
Levnedsmiddel industri	2,29	1,73	0,83
Information & telekom.	1,71	2,12	1,53
Engroshandel	2,49	3,18	1,34
Detailhandel	3,01	1,70	0,67
Service	1,58	0,54	0,35

Betegnelsen 'rettesnor' skal tages helt bogstaveligt. Tallene prætenderer ikke at være absolutte eller entydige, men, som nævnt under diskussionen af DM-begrebet, blot et forbedret beslutningsgrundlag. Udover den usikkerhed, der kan henføres til procedure- og beregningsmæssige forhold vil prisudviklingen naturligvis relativt hurtigt gøre tallene invalide. Omsætning pr. heltidsansat medarbejder vil være mest følsom over for prisudviklingen på kort sigt, men på blot mellemlangt sigt vil alle nøgletallene ændre sig som følge af prisændringer. Ændringer i produktiviteten virksomhedstyperne imellem – der sædvanligvis er noget langsommere – vil tillige medføre en forskydning af nøgletallene indbyrdes.

Analysens resultater kan muligvis også anvendes i et helt andet perspektiv. De sammenhørende værdier af nøgletallene kan betragtes som et kvantitativt udtryk for den bærende idé i de enkelte virksomhedstypers økonomistyring og de vilkår, der er gældende for udøvelsen af styringen, og dermed også en norm / et mål for styringens effektivitet givet de forskellige styringsgrundlag, der er gældende for de forskellige virksomhedstyper. Det er måske mere forståeligt nu, at aktiemarkedet har vendt tommelen nedad for ISS´s interesse for Sophus Berendsen. Sophus er i bund og grund en produktionsvirksomhed og skal styres og ledes som sådan, medens ISS´s ledelseserfaring stort set kun rækker til servicevirksomheder - alle de produktionsvirksomheder som ISS har ejet, har man afhændet som regel som underskudsforretninger. Det er måske også mere forståeligt nu hvorfor Magasin har så vanskeligt ved at få enderne til at hænge sammen. Med en økonomisk struktur, der svarer til procesindustriens, men med en indtjening, der svarer til detailhandelens, må der være en betydelig ubalance i systemet.

oxo OXO oxo

Bilag

Bilag A – klassificering af virksomhederne

Plac 01	Selskab	Branche	Type
382	Scaniadam Holding	BILH	
320	Hessel, Ejner Holding	BILH	ENGH
374	Nellemann Holding	BILH	ENGH
376	MAN Last og Bus	BILH	ENGH
467	Reinhard Nielsen	BILH	ENGH
115	Ford Motor Company	BILH	EU
40	Daimler Chrysler Skand.	BILH	extr
80	Semler Holding (11)	BILH	extr
83	Interdan (12)	BILH	extr
149	Toyota Danmark	BILH	extr
198	Nic Christiansen Holding (2	BILH	extr
229	Stena Metall	BILH	extr
244	Volvo Personvogne Danma	BILH	extr
249	Volvo Lastv. og Busser	BILH	extr
259	Citroen Danmark	BILH	extr
295	Fiat Automobiler DK	BILH	extr
298	Andersen Motors	BILH	extr
364	Scania Danmark	BILH	extr
429	Lastas	BILH	extr
19	Danske Trælast	BYGG	
48	Icopal	BYGG	BYGG
174	H+H Holding	BYGG	BYGG
200	Consenta Holding (30)	BYGG	BYGG
266	Flügger	BYGG	BYGG
426	Danogips	BYGG	BYGG
434	Betonelement	BYGG	BYGG
436	Hansengroup	BYGG	BYGG
450	H+H Fiboment	BYGG	BYGG
462	Expedit	BYGG	BYGG
64	DLH	BYGG	ENGH
111	Brødrene Dahl	BYGG	ENGH
142	Bygma	BYGG	ENGH
358	Anco Træ	BYGG	ENGH
390	M. J. Eriksson	BYGG	Engh
427	Color line Danmark	BYGG	ENGH
334	Sjælsø Gruppen	BYGG	EU
253	KPC Byg	BYGG	extr
350	Phønix Contractors	BYGG	PROC
31	Velux Industri	BYGG	PROD
37	Rockwool International	BYGG	PROD
72	C.W. Obel	BYGG	PROD
211	Vest-Wood	BYGG	PROD
228	Junckers Industrier	BYGG	PROD
274	Saint-Gobain Glass Nor.	BYGG	PROD
370	SP Group (41)	BYGG	PROD
393	Spæncom	BYGG	PROD
428	Arvid Nilsson	BYGG	PROD
488	Dansk Industri Invest	BYGG	PROD
497	Inter Primo	BYGG	PROD
293	Kemp & Lauritzen	BYGG	SERV
496	Rationel Vinduer	BYGG	SERV
231	Fredgaard Radio	DETH	
280	Synoptik	DETH	
6	FDB Koncernen	DETH	DETH
30	Føtex	DETH	DETH
32	Bilka Lavprisvarehus	DETH	DETH
90	Jysk Sengetøjslager	DETH	DETH
123	Aldi Holding	DETH	DETH
129	OBS Danmark	DETH	DETH
138	KNI	DETH	DETH
145	Hennes & Mauritz	DETH	DETH
181	Fona Gruppen Holding	DETH	DETH
307	Bauhaus Danmark	DETH	DETH
330	Dreisler Storkøb	DETH	DETH
466	Løvbjerg Supermarked	DETH	DETH
492	Audionord International	DETH	DETH
12	Dansk Supermarked	DETH	ENGH
223	Harald Nyborg	DETH	Engh
490	BSH Hvidevarer	DETH	ENGH
45	Fleggaard Holding	DETH	extr
233	Inbodan	DETH	extr
297	Jaco Gruppen	DETH	extr
359	F.Salling	DETH	extr
363	Nuance Global Trad.	DETH	extr
372	Rema 1000 Danmark	DETH	extr
401	Imerco	DETH	extr
477	Expert	DETH	extr
122	Wessel & Vett	DETH	PROC
421	Hanssen & Co. Holding	DETH	PROC
213	Jensen Group	DETH	PROD
78	Q8 Danmark (10)	ENER	ENGH
22	DONG	ENER	EU
51	Mærsk Olie og Gas	ENER	EU
56	Eltra	ENER	EU
60	Energi E2	ENER	EU
106	NESA	ENER	EU
151	HNG	ENER	EU
160	Naturgas Midt-Nord	ENER	EU
175	D.F.N. Olie (27)	ENER	EU
252	KE Energiforsyning	ENER	EU
415	Disam	ENER	EU
437	Haahr Benzin	ENER	EU
474	Denerco Oil	ENER	EU
11	Statoil Danmark	ENER	extr
25	Dansk Shell	ENER	extr
41	Hydro Texaco Holdings	ENER	extr
57	Elsam (7)	ENER	extr
197	DK-Benzin	ENER	extr
230	Elbodan	ENER	extr
395	Star Tour	ENER	extr
279	Carl F. Petersen	ENGH	
299	Østsjællands Andel	ENGH	
338	Mesco Denmark (39)	ENGH	
410	P. Brøste (47)	ENGH	
24	DLG	ENGH	ENGH
26	KFK	ENGH	ENGH
47	Solar Holding	ENGH	ENGH
75	Lemvigh-Müller Holding	ENGH	ENGH
113	Sanistål	ENGH	Engh
196	Gasa Århus (28)	ENGH	ENGH
208	A. & O. Johansen	ENGH	ENGH
217	Inco Fællesindkøb	ENGH	ENGH
260	Hedegaard	ENGH	ENGH
262	Consiva Holding	ENGH	ENGH
272	IKEA	ENGH	Engh
406	Triumph Int. Textil	ENGH	ENGH
433	Lyreco Danmark	ENGH	Engh
442	DBK-Bogdistribution	ENGH	ENGH
451	Bøg Madsen	ENGH	ENGH
452	3M	ENGH	ENGH
469	CC&CO Holding	ENGH	Engh
479	L'Oreal	ENGH	ENGH

Plac 01	Selskab	Branche	Type
332	Birns Jernstøberi	JERN	PROD
472	Alstom Power Flowsystems	JERN	Prod
485	Ferrosan	JERN	PROD
491	Ib Andresen Industri	JERN	PROD
236	Colgate-Palmolive	KEMI	ENGH
493	Brenntag Nordic	KEMI	ENGH
82	Hempel Gruppen	KEMI	PROC
84	Auriga Industries	KEMI	PROC
91	Nycomed Gruppen (14)	KEMI	PROC
161	Haldor Topsøe	KEMI	PROC
225	Kemira Danmark	KEMI	Proc
336	Sun Chemical	KEMI	PROC
367	BASF Health & Nutrition	KEMI	Proc
38	Akzo Nobel	KEMI	prod
424	Henkel-Ecolab	KEMI	prod
13	J. Lauritzen Holding	KONG	
379	Jern Holding	KONG	
17	Skandinavisk Holding	KONG	Engh
141	TK Development	KONG	EU
14	FLS Industries	KONG	PROC
36	ØK	KONG	PROD
50	Monberg & Thorsen	KONG	Prod
94	VT Holding	KONG	PROD
118	Incentive	KONG	prod
120	Superfos	KONG	PROD
146	Micro Matic Holding	KONG	PROD
147	Schouw & Co.	KONG	Prod
386	Therp Holding	KONG	PROD
411	NTR Holding	KONG	PROD
440	Schou-Fondet	KONG	PROD
245	Nestlé Danmark Holding	LEVN	
119	Grey Scandinavia	LEVN	ENGH
130	BHJ	LEVN	ENGH
178	Jysk Cater	LEVN	ENGH
267	DLF AmbA	LEVN	ENGH
277	Kraft Foods Danmark	LEVN	ENGH
318	Fiskernes Fiskeindustri	LEVN	ENGH
387	Esbjerg Fiskeindustri	LEVN	ENGH
408	Kurt Skare Holding	LEVN	ENGH
448	Beauvais (49)	LEVN	ENGH
270	Rose Poultry	LEVN	EU
419	Saga Lax	LEVN	EU
21	Dagrofa	LEVN	extr
125	CP Kelco (19)	LEVN	extr
172	F. Uhrenholt Holding	LEVN	extr
222	Kangamiut Fish Holding	LEVN	extr
255	Cerestar Scandinavia	LEVN	extr
257	DSB Resturanter	LEVN	extr
457	Central Soya European Pro	LEVN	extr
4	Danish Crown	LEVN	LEVN
53	Dat-Schaub	LEVN	LEVN
59	Steff-Houlberg	LEVN	LEVN
157	Arovit Petfood (23)	LEVN	LEVN
171	Nowaco	LEVN	LEVN
209	TICAN	LEVN	LEVN
263	A. Espersen	LEVN	LEVN
273	Lactosan-Sanovo	LEVN	LEVN
309	SFK (37)	LEVN	LEVN
310	Norway Seafoods Denmark	LEVN	LEVN
340	Rynkeby Foods	LEVN	LEVN
362	Cerealia Danmark	LEVN	LEVN
378	Hjem-Is-Europa	LEVN	LEVN
470	Rahbekfisk (53)	LEVN	LEVN
5	Carlsberg (2)	LEVN	PROC
9	Danisco	LEVN	PROC
65	Aarhus Olie	LEVN	Proc
70	Chr. Hansen Holding	LEVN	PROC
89	Royal Greenland	LEVN	PROC
137	Bryggerigruppen	LEVN	proc
256	Polar Seafood Greenland	LEVN	Proc
294	Unilever Danmark	LEVN	PROC
366	Alfa Laval LKM	LEVN	PROC
468	Daka Amba	LEVN	PROC
500	Albani Bryggerierne	LEVN	PROC
156	Schulstad	LEVN	PROD
166	Dandy Holding	LEVN	PROD
190	Toms Fabrikker	LEVN	PROD
261	KelsenBisca (35)	LEVN	PROD
264	Hatting Bageri	LEVN	PROD
271	Danpo	LEVN	Prod
319	Tholstrup Cheese Holding	LEVN	PROD
324	CO-RO Holding	LEVN	PROD
326	Danske Spritfabrikker	LEVN	Prod
343	Harboes Bryggeri	LEVN	prod
371	Kongskilde Industries	LEVN	PROD
432	DDG Holding	LEVN	Prod
447	Gøl Holding	LEVN	PROD
43	Vestas Wind Systems	MASK	
104	Aalborg Industries	MASK	
232	Danewco	MASK	
251	Electrolux Home Products	MASK	ENGH
348	Thermo King Co	MASK	ENGH
71	Viborg Gruppen Holding	MASK	EU
8	Borealis (3)	MASK	extr
460	New Holland Danmark	MASK	extr
42	NKT Holding	MASK	PROC
303	Crisplant	MASK	Proc
384	Niro (44)	MASK	Proc
18	Danfoss	MASK	PROD
74	NEG Micon	MASK	prod
98	Grundfos	MASK	PROD
132	DISA (20)	MASK	PROD
167	APV Pasilac	MASK	Prod
202	Vestfrost	MASK	Prod
291	Aage V. Kjærs Maskinfabri	MASK	PROD
361	Svend Møller Hansen Hld	MASK	PROD
394	GPV Industri	MASK	PROD
455	Skiold Holding	MASK	Prod
459	Thrige-Titan	MASK	PROD
461	Scanvaegt International	MASK	PROD
463	Denka Holding (52)	MASK	PROD
464	Nilpeter Holding	MASK	prod
484	Disa Industries	MASK	Prod
169	Stubkjær	MASK	
349	Chista Consult (40)	MEDI	ENGH
397	Astra Danmark	MEDI	ENGH
179	K. V. Tjellesen	MEDI	EU
185	Bayer	MEDI	EU
44	Nomeco	MEDI	extr
193	Brdr. Lembcke	MEDI	extr
304	Max Jenne	MEDI	extr
445	Bukkehave	MEDI	extr

Plac 01	Selskab	Branche	Type
486	Bonnier Publications	MEDI	extr
10	Novo Nordisk	MEDI	PROC
52	H. Lundbeck	MEDI	PROC
77	Løvens Kemiske Fabrik	MEDI	PROC
100	Alpharma (16)	MEDI	PROC
311	Tellabs Denmark	MEDI	Proc
88	Coloplast	MEDI	PROD
110	William Demant Holding	MEDI	Prod
188	Radiometer	MEDI	PROD
284	Løgstør Rør Holding	MEDI	Prod
285	Cook Denmark Internationa	MEDI	PROD
331	Dako	MEDI	PROD
385	F.E. Bording	MEDI	PROD
327	Glunz & Jensen	MEFO	
409	Color Print	MEFO	
61	Int. Masters Publishers	MEFO	ENGH
29	Egmont Fonden	MEFO	extr
288	CIA Denmark	MEFO	extr
341	Initiative Universal	MEFO	extr
101	Carl Allers Etablissement	MEFO	PROC
187	TV 2	MEFO	Proc
107	Det Berlingske Officin	MEFO	Prod
117	DR	MEFO	PROD
180	Dagbladet Politiken	MEFO	Prod
242	Søndagsavisen	MEFO	Prod
248	Jyllands-Posten	MEFO	Prod
265	Aarhuus Stiftsbogtryk	MEFO	PROD
407	Aalborg Stiftstidendes	MEFO	PROD
422	Gyldendal	MEFO	PROD
425	Fyens Stiftstidende	MEFO	PROD
346	HTH Køkkener	MØBL	
444	Ilva Holding	MØBL	
124	Royal Scandinavia	MØBL	PROD
150	Tvilum-Scanbirk	MØBL	prod
268	Bodilsen Holding	MØBL	PROD
289	ITH	MØBL	PROD
329	Licentia Group	MØBL	PROD
398	Egetæpper	MØBL	PROD
403	Plus Pack	PAPI	
465	Esselte	PAPI	
287	Papyrus	PAPI	ENGH
438	Inventing Denmark	PAPI	ENGH
494	Færch Holding	PAPI	PROC
140	Rosti	PAPI	PROD
176	SCA Packaging Denmark	PAPI	PROD
205	Hartmann, Brødrene	PAPI	PROD
212	Schur International	PAPI	PROD
214	Glud & Marstrand	PAPI	PROD
275	Danapak	PAPI	PROD
305	Bantex	PAPI	SERV
144	D/S Torm	REDE	EU
170	D/S Norden	REDE	extr
220	Tschudi & Eitzen Bulkers	REDE	extr
69	DFDS	REDE	PROC
423	Mols-Linien	REDE	PROC
33	Odense Staalskibsværft	REDE	PROD
103	MAN B&W Diesel	REDE	PROD
269	Weco-Rederi	REDE	Prod

Plac 01	Selskab	Branche	Type
317	Ø. S. Invest	REDE	PROD
183	Kilroy Travels	SERV	ENGH
396	Novasol	SERV	ENGH
81	Statoil Detail	SERV	extr
168	Københavns Lufthavne	SERV	extr
314	Tumlare Corporation	SERV	extr
441	Eurest	SERV	extr
453	Bladkompagniet	SERV	extr
482	SOS International	SERV	extr
487	Grønlandsfly	SERV	proc
92	Sophus Berendsen	SERV	PROD
355	Statens Serum Institut	SERV	prod
480	KOFF	SERV	PROD
7	ISS	SERV	SERV
15	Group 4 Falck (4)	SERV	SERV
23	Post Danmark	SERV	SERV
58	Gate Gourmet North. Eur	SERV	SERV
235	Deloitte & Touche	SERV	serv
456	Scandic Hotel	SERV	SERV
489	Novia Holding	SERV	SERV
344	Møller & Co Gruppen	TEKS	
99	IC Company (15)	TEKS	BYGG
478	Fibertex	TEKS	PROC
114	Kansas Wenaas	TEKS	PROD
131	Brandtex	TEKS	Prod
135	Ecco Holding	TEKS	SERV
20	SAS Danmark (5)	TRAN	
68	DSV	TRAN	ENGH
300	Leman (36)	TRAN	ENGH
328	Sterling European	TRAN	Engh
431	ASG Holding	TRAN	engh
458	Iveco Danmark (51)	TRAN	ENGH
475	Bestfoods Nordic	TRAN	ENGH
116	Sund og Bælt	TRAN	EU
164	World Tourist Rejseb. (25)	TRAN	extr
226	United Shipping Agencies	TRAN	extr
347	Scandlines Danmark	TRAN	extr
93	Maersk Air	TRAN	PROC
377	Cimber Air-Holding	TRAN	PROC
380	Em. Z. Svitzer (42)	TRAN	PROC
16	Maersk	TRAN	prod
420	Royal Artic Line	TRAN	PROD
240	Arriva Danmark	TRAN	SERV
254	Combus	TRAN	SERV
54	Dansk Tipstjeneste	UNDE	extr
27	Lego Gruppen (6)	UNDE	PROC
414	Scanbox Entertainment	UNDE	Proc
204	Top-Toy	UNDE	PROD
413	Kompan	UNDE	Prod

EU = ej undersøgt

Bilag B – gennemsnitsværdier alle observationer

				Eksklusiv 78 extreme værdier:		
	Oms / bal	***0ms / Ans***	***Bal / Ans***	***Oms / bal***	***0ms / Ans***	***Bal / Ans***
count	433	433	433	355	355	355
mean	2,046	3,130	1,605	1,656	1,714	1,146
standard error of the mean	0,080	0,209	0,074	0,042	0,055	0,034
confidence interval, 95% lower	1,889	2,719	1,460	1,573	1,605	1,078
confidence interval, 95% upper	2,203	3,541	1,749	1,740	1,823	1,214
sample standard deviation	1,660	4,352	1,531	0,798	1,045	0,650
tandardiseret 3. moment skewness	3,663	3,654	2,661	1,167	1,663	1,415
Standardiseret 4. moment: kurtosis	21,226	16,844	8,586	1,600	3,109	2,791

Bilag C – rangordnede gennemsnit pr. virksomhedstype

Rangordnede gennemsnit pr. virksomhedstype					
Oms / bal		0ms / Ans		Bal / Ans	
PROC	0,83	SERV	0,54	SERV	0,35
PROD	1,28	PROD	1,07	DETH	0,67
SERV	1,58	BYGG	1,24	BYGG	0,70
GNS	**1,66**	PROC	1,65	LEVN	0,83
ITEL	1,71	DETH	1,70	PROD	0,88
BYGG	1,81	**GNS**	**1,71**	**GNS**	**1,15**
LEVN	2,29	LEVN	1,73	ENGH	1,34
ENGH	2,49	ITEL	2,12	ITEL	1,53
DETH	3,01	ENGH	3,18	PROC	2,05
95 % konfidensinterval omkring GNS					
Nedre	1,57	Nedre	1,60	Nedre	1,08
Øvre	1,74	Øvre	1,82	Øvre	1,21

Bilag D – normaliserede afstande mellem virksomhedstyperne

Afstande i standardafvigelser S(e)								
	PROD	ENGH	PROC	SERV	BYGG	DETH	ITEL	LEVN
PROD		3,98	4,09	2,13	1,24	2,62	3,07	2,66
ENGH	3,98		3,10	4,09	2,48	1,22	1,20	1,22
PROC	4,09	3,10		5,44	3,75	1,59	0,59	1,86
SERV	2,13	4,09	5,44		1,76	2,95	2,62	2,69
BYGG	1,24	2,48	3,75	1,76		1,32	1,18	0,98
DETH	2,62	1,22	1,59	2,95	1,32		0,31	0,18
ITEL	3,07	1,20	0,59	2,62	1,18	0,31		0,23
LEVN	2,66	1,22	1,86	2,69	0,98	0,18	0,23	

I en todimensional afbildning kan de 8 grupper placeres således i forhold til hinanden: