

Pressmeddelande 051121

Rapport från M2005 i Las Vegas - världens största Data Mining-konferens

Över 700 analytiker från 30 länder deltog på världens största Data Mining konferens, M2005, som hölls i Las Vegas, USA, i slutet av oktober 2005. Från Sverige deltog ett tiotal av SAS Institutes kunder från branscher som bank, telekom och offentlig sektor.

Under två dagars konferens hölls inte mindre än 33 presentationer om data mining från en rad olika organisationer. Nästan hälften av föredragen handlade om olika aspekter av sälj- och marknadsföringsproblem vilket är naturligt då databaserad marknadsföring och segmentering av kundbaser är de kommersiella tillämpningar som drivit mycket av utvecklingen inom data mining de senaste tio åren.

Professor David Hand från Imperial College i London är en av världens mest namnkunniga statistiker, med ett tjugotal böcker och en mängd artiklar och uppsatser på meritlistan. Han inledde konferensen med att uppmärksamma problem som kan uppkomma vid data mining. Ofta är det sekundär data, d.v.s. sådana data som från början inte samlats in just för de aktuella analysfrågeställningarna som hämtas in från befintliga transaktionssystem och databaser. Det är vanligt att denna typ av data är av dålig kvalitet, vilket kan innebära inkonsekvent data, saknade värden, extremvärden och andra felaktigheter som kan påverka analyserna på ett negativt sätt. Ett sätt att få bukt på kvalitetsbrister i data kan vara att använda robusta modelltekniker och att vara extra omsorgsfull vid valet av förklarande variabler. David Hand gav också exempel på hur lätt man vilseleds genom att den aktuella frågeställningen inte är tillräckligt tydlig och att analysen då inte mäter det som var avsikten med analysen. Inom data mining pratas det mycket om att hitta mönster i befintligt data men det typiska är att man vill veta hur det ser ut om en månad eller ett år, dvs förklara vilket mönster som är troligt att inträffa framöver. Beroende på vilken tidsperiod som väljs och hur stora datamängder man utgår från, riskerar man att få bagateller att framstå som sensationella upptäckter. Istället för att använda standardmetoder (t.ex. Bonferroni) för att justera för andelen 'falska mönster' som är signifikanta föreslog David Hand att man kontrollerar de mönster som är signifikanta och som är 'falska'. Trots hans förmanande ord om risken för olika typer av analysfel, framhöll David Hand ändå vikten av framtida tillämpning.

Årets konferens innehöll också flera presentationer från banker, försäkringsbolag och telekombolag som beskrev tillämpningar för hantering av kreditrisker, bedrägerier, penningtvätt och andra typer av riskhantering. På det amerikanska universitetet Purdue använder man t.ex. data mining inom antiterrorism och övervakning av smittspridning. Med data från över femhundra djursjukhus har de bland annat utvecklat modeller som används för att upptäcka om husdjur uppvisar en ökad risk för olika influensaepidemier. Genom att identifiera vilka symptom som förklarar den aktuella influensan, prognostisera sjukdomsutbredningen bland djuren och beräkna förväntad tidsförskjutning för överföring mellan djur och människor, så kan de skapa larmsystem baserade på fakta istället för på medias larmrapporter som t.ex. fågelinfluensan.



Pressmeddelande 051121

Det fanns tre tydliga tekniska trender på konferensen: Hanteringen av allt större datamängder, hur prediktiv analys integreras och byggs in i olika applikationer samt analys av text. David Duling, som är utvecklingschef för SAS Enterprise Miner, visade hur några av problemen med analys av extremt stora datamängder – terabyte av data – kan hanteras med hjälp av så kallad GRID-processing. GRID-processing innebär att man fördelar analysen på datorer i ett nätverk som löser mindre delar av analysen separat. David Duling beskrev vilka strategier som visat sig vara mest effektiva vid olika typer av analyser. Slutsatsen blev att man idag kan hantera oerhört stora datamängder även med mycket komplicerade modeller, men att detta också givetvis ställer höga krav på analytiker och statistiker.

Hantering av text beskrevs av många som nästa stora tillämpningsområde för prediktiv modellering och data mining. Terry Woodfield från SAS Institute visade bland annat hur man med text mining byggt ett system för att indikera bedrägeriförsök vid anmälan av arbetsplatsolyckor. Genom att tillföra fritext från anmälningarna har försäkringsbolaget anpassat en avsevärt mycket bättre modell för att kunna förutse bedrägligt beteende. Andra tillämpningar av text mining som nämndes under konferensen var hantering av garantiärenden för att upptäcka tillverkningsfel så snabbt som möjligt, kategorisering av jobbsökningar och bidragsansökningar och användning av läkares noteringar för bättre diagnostisering.

I övrigt stod det klart att data mining idag är ett moget tillämpningsområde där analytiker och andra användare inte längre diskuterar vad prediktiv modellering ska användas till, utan snarare hur man effektiviserar analysprocessen och för ut resultaten i organisationen på bästa sätt.

Ed Gaffin från Walt Disney World var en av flera föredragshållare som beskrev hur bolagets statistiker får en allt mer central roll i organisationen, samtidigt som deras analyser och modeller flyttas ut närmare användarna – fler och fler yrkesgrupper har behov av modellerna vilket ställer nya krav på användbarhet. Flera föredrag handlade om hur företag ”byggt in” prediktiv modellering i olika applikationer. Ett stort telekombolag har exempelvis integrerat kundpreferensmodeller i sitt callcenter för att göra det mer lättillgängligt för hela företagets verksamhet och kundkontakt.

Detta var det åttonde året i rad som Data Mining-konferensen hölls. Ordförande för konferensen Michael J. A. Berry, som varit med på alla konferenser hittills, konstaterade att deltagarna blivit allt mer sofistikerade och att de analytiker och statistiker som arbetar med prediktiv modellering får en allt mer central roll i sina organisationer.

Catharina Svenningstorp
SAS Institute



Pressmeddelande 051121

Faktaruta

Data mining är en process för att nå och utforska stora mängder data och anpassa modeller för att avslöja samband och mönster i datat för att ta fram säkra beslutsunderlag. Tillämpningen av data mining sträcker sig över många branscher och områden. Inom telekom, bank, finans och försäkring används data mining för att upptäcka bedrägerier, optimera marknadskampanjer och för att identifiera de mest lönsamma prisstrategierna. Läkemedelsindustrin använder data mining för att prediktera effekten av olika typer av medicinering, läkemedelstester och kirurgiska ingrepp. Inom handel används data mining för att utvärdera effekten av rabattkuponger och specialerbjudanden för att förutse vilka erbjudanden som passar olika kunder bäst.

Data mining-verktyg strukturerar upp hela analysprocessen, från inhämtning av data, beskrivande analys och prediktiv modellering ända fram till att modellen ska utvärderas och tillämpas i verksamheten. Detta gör att man genom att arbeta med ett data mining-verktyg effektiviserar och kvalitetssäkrar analysarbetet. Vanliga analystekniker inom data mining är t.ex. klustertekniker, associations- och sekvensanalys, linjär och logistisk regression, olika typer av beslutsträd, neurala nätverk och regelbaserad induktion.

Om SAS Institute

SAS Institute är världens ledande mjukvaruföretag inom Business Intelligence och affärsanalys. SAS Institute, även världens största privatägda mjukvaruföretag, omsätter årligen runt 10 miljarder kronor i 105 länder och har 29 års erfarenhet av att utveckla verktyg och metoder som låter stora organisationer lära av sin historia, mäta och kommunicera pågående aktiviteter och inte minst skapa insikt om framtiden. Världen runt har SAS Institute totalt gjort 40.000 kundinstallationer, bland annat i 90 procent av Fortune 500-företagen. I Sverige startade SAS Institute AB år 1986 och är idag drygt 80 anställda med kontor i Stockholm och Göteborg. Bland de svenska kunderna finns landets mest betydande företag och organisationer.

Läs mer på www.sas.com/sweden