*BENEDEK Gábor**

MESTERSÉGES INTELLIGENCIA AZ ÜZLETI VILÁGBAN

– Marketingakciók hatékonyságának elemzése statisztikai és Data Mining módszerekkel –

A cikk egy gyakorlati probléma segítségével kívánja bevezetni az olvasót a Data Mining módszertan alkalmazási lehetőségeibe. Az elemzést a dolgozat szerzői készítették és mutatták be a Clementine Users Group Conference 1999 alkalmából. Az elemzéshez használt adatbázis szintetikus, és a szemléltetés kedvéért megfelelően egyszerűsített.

A matematika és a számítástechnika területén már régóta megjelentek, és sok alkalmazásban (kép-, hangfelismerés, fordítóprogramok) szerepelnek sikeresen a mesterséges intelligencia algoritmusai. Hamar jelentkezett az üzleti világ igénye is az újszerű adatelemzési technológiára, a Data Miningra. A Data Mining sikeressége elsősorban annak tudható be, hogy a vállalatok a modern számítástechnikai lehetőségek miatt óriási, és viszonylag könnyen és gyorsan hozzáférhető adatbázisokkal rendelkeznek. Ezen adattárházak sokszor nem is adatelemzési célokat szolgáltak (hanem például a számlázást), mégis az adatok mögött rejlő információ rendkívül nagy erőt adhat a vállalat vezetőinek, stratégiai tervezőinek kezébe. Hagyományos statisztikai eszközökkel azonban reménytelennek tűnő feladat egy ilyen óriási adatbázisból az összes hasznos információ kinyerése. Ennek pedig paradox módon pont a túlságosan nagy adatbázis és a rendkívül sokféle és bonyolult (nem lineáris) összefüggések az okai. Nem beszélve arról, hogy mondjuk egy statisztikai hipotézisvizsgálathoz már az elemzés kezdetén rendelkeznünk kell valamilyen feltevéssel, melyet az adatok birtokában meg akarunk erősíteni, vagy el akarunk vetni. A Data Mining alkalmazásánál nem szükségesek a prekoncepciók, a számítógép automatikusan generálja őket. (Sokszor többet is, melyeket már tényleg a statisztikai próbák, illetve a vállalati szakértők ellenőriznek.) Az algoritmusok nem „részrehajlók", nem kerüli el a figyelmüket semmi, és képesek bonyolult összefüggéseket és kapcsolatokat is feltárni.

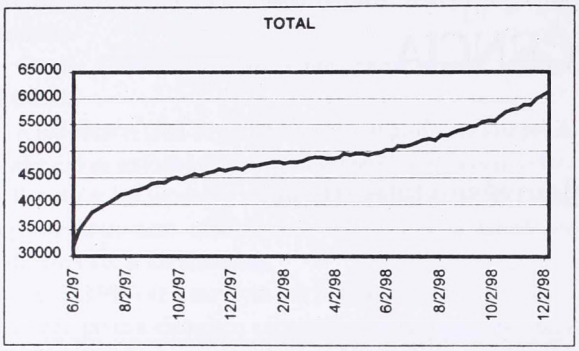
A probléma

Adott egy vállalat, amely valamilyen terméket vagy szolgáltatást árul. Az értékesítés partnereken keresztül (pl. bolt) történik. A forgalom az elmúlt időszakban eleinte erőteljes növekedést mutatott, majd megtorpant. A vállalatnak egy nagyszabású reklámakcióval sikerült kimozdulnia a stagnálásból, és újból forgalomnövekedés következett be. *(1. ábra)* A vállalat tisztában van azzal, hogy a forgalomnövekedés nem egységesen ment végbe a reklámakciót követően. Azaz összességében növekedett az eladás, de egyes boltokon belül e növekedés mértéke, tendenciája eltérő volt. A vállalat arra szeretne magyarázatot kapni, hogy vajon *milyen tényezők játszottak közre* az egyes boltoknál a reklámakció hatékonysága szempontjából. Továbbmenve arra is kíváncsi, hogy *mi jellemző azokra a boltokra,* ahol az eddig használt marketingakció hatékony, és mi azokra, ahol más reklámstratégiát érdemes alkalmazni. Sőt, pozitív válaszok esetén tovább lehetne mélyíteni az elemzést (hova

* A cikk alapjául szolgáló konferenciaanyagot a szerző Dévényi Edittel együtt készítette.

érdemes új boltokat telepíteni, milyen időpontokban, illetve időközönként érdemes új akcióval előállni stb.), ez azonban meghaladja a tanulmány kereteit.

1. ábra



Az adatok

Két adatbázis állt a vállalat rendelkezésére. Az első – forgalom adatbázis – az egyes boltok (összesen 50 db) *heti forgalmát* tartalmazza akció előtt és után. A második – demográfiai adatbázis – az egyes boltok *általános jellemzőit* tartalmazza, úgymint régió; a vonzáskörzet átlagos jövedelme; boltméret; parkolási lehetőség; média ellátottság. *(1. táblázat)*

sor a demográfiai adatbázis felhasználására. Ekkor ugyanis minden bolt rendelkezik egy egyértelmű szegmenttípussal. Azt kívánjuk elérni, hogy az egyes boltok demográfiai adatai alapján *be tudjuk sorolni* őket a nekik megfelelő szegmensbe. Első ránézésre úgy tűnhet, hogy ugyanazt a feladatot végezzük el még egyszer, de gondoljuk meg, hogy egy új akció vagy egy új bolt megnyitása esetén nem rendelkezünk a jövőbeli forgalmi adatbázissal, csak a demográfiai adatokkal.

Statisztikai és Data Mining módszerek

Első feladatunk a forgalom adatbázis vizsgálata. Ehhez a statisztikából ismert idősorelemzéshez nyúltunk. Az idősorelemzés meglehetősen terjedelmes eszköztárából mi a lehető legegyszerűbb módszert használtuk; a legkisebb négyzetek elvére épülő *illesztéseket.* Minden egyes bolti idősorra négy különböző modellt illesztettünk; lineárisat, logaritmikusat, exponenciálisat és kvadratikusat. Ennek megfelelően minden bolthoz megkaptuk az adott modellhez tartozó paramétereket (*b0, b1* és a kvadratikus esetben *b2* is), az R^2-et és az *F* statisztikát.* *(2. táblázat)*

1. táblázat

Forgalom adatbázis

Time	SHOP1	SHOP2	SHOP3...
06.02.97	624.927	670.874	655...
06.09.97	692.243	724.555	723...
06.16.97	717.772	759.384	708...
06.23.97	783.636	755.236	810...
06.30.97	772.534	752.692	769...
07.07.97	824.155	752.137	759...
07.14.97	856.076	768.098	809...
07.21.97	813.699	775.112	789...
07.28.97	807.969	843.743	852...
08.04.97	863.928	876.288	874...

Demográfiai adatbázis

SHOP	Region	Attraction Zone Income	Store Scale	Parking	Media Attendance
SHOP33	Town	Average	Large	Good	Average
SHOP32	Town	Well-to-do	Small	Bad	Average
SHOP31	City	Well-to-do	Medium	Good	Average
SHOP30	Budapest	Average	Small	Bad	Average
SHOP29	Budapest	Poor	Medium	Good	Average
SHOP1	City	Average	Medium	Bad	Average
SHOP43	City	Average	Small	Good	Average
SHOP42	Village	Tight	Large	Bad	Average
SHOP41	Village	Average	Large	Good	Average
SHOP40	City	Tight	Small	Bad	Average

Ennek tükrében megfogalmazhatók azok a konkrét Data Mining célok, melyek elérését előzetesen kitűztük. Szeretnénk minden egyes bolthoz hozzárendelni a forgalmi adatai alapján néhány olyan *jellemzőt,* mely tömören jellemzi, hogy az adott boltban hogyan alakult az értékesítés. Ezek alapján megpróbáljuk *szegmentálni* a boltokat teljesítményük alapján. Nemcsak arról van szó, hogy szét szeretnénk választani egymástól a jó, illetve a kevésbé jó forgalmú boltokat. Ehhez különben is valahogyan definiálnunk kellene a „jóság" fogalmát. Ehelyett először elvégezzük a csoportosítást, majd ezután a csoportokat egyenként jellemezzük *(címkézés).* Ezután kerül

A következő vizsgálathoz egyedül ezeket a generált adatokat (modelltípusok, paraméterek és statisztikák) használtuk. Itt alkalmaztunk először Data Mining algorit-

* Röviden e jellemzők értelmezéséről. A *paraméterek* jelentése minden modell esetében más és más. Így például a lineáris modell esetén a 44-es boltnál található *b1* = 8,79 azt jelenti, hogy átlagosan minden héten 8,79-dal nőtt az adott bolt forgalma. Az R^2 az illeszkedés jóságát méri. Értéke 0 és 1 között van, és minél közelebb van az 1-hez, annál pontosabb a modell illeszkedése. Az *F* statisztika azt vizsgálja, hogy a paraméterek szignifikánsan különböznek-e 0-tól. Magas értéke megerősíti e feltételezést.

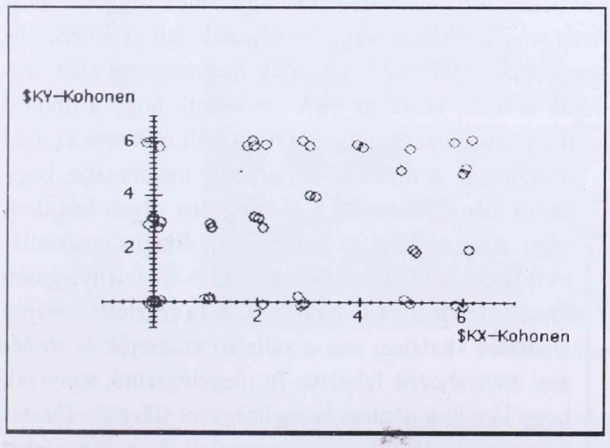
2. táblázat

Dependent	Mth	Rsqd.f.		F	b0	b1	b2
SHOP43	LOG	0,712	29	71,72	877,787	91,3291	–
SHOP43	QUA	0,948	28	254,47	975,786	5,2611	0,1425
SHOP43	EXP	0,940	29	455,41	958,726	0,0088	–
SHOP44	**LIN**	**0,805**	**29**	**119,55**	**1023,97**	**8,7948**	**–**
SHOP44	**LOG**	**0,841**	**29**	**153,89**	**923,274**	**95,8317**	**–**
SHOP44	**QUA**	**0,870**	**28**	**94,08**	**968,603**	**18,8610**	**-0,3146**
SHOP44	**EXP**	**0,795**	**29**	**112,45**	**1026,33**	**0,0077**	**–**
SHOP45	LIN	0,807	29	121,12	1037,67	7,6957	–
SHOP45	LOG	0,833	29	144,53	951,100	83,3221	–

must, mégpedig a *Kohonen Network szegmentáló eljárást.** A *2. ábra* segítségével könnyen megérthető az eljárás eredménye. Minden egyes kör a grafikonon egy-egy boltot reprezentál. Az egymáshoz közel eső boltok hasonló típusúak, azaz az illesztett modellek paraméterei és statisztikái is nagyon hasonlóak. Az egymástól messze eső boltoknál a helyzet fordított, ezek a boltok nagyon különböznek egymástól, legalábbis forgalmi adataik alapján.

Ez azonban még nem elegendő, hiszen látható, hogy egy-egy kis csoportba csupán 3–4, rossz esetben még kevesebb egyed tartozik. Ennyi különböző szegmenset

2. ábra



* A Kohonen szegmentáló algoritmus kifejlesztője dr. Teuvo KOHONEN egyetemi tanár, Neural Networks Research Centre, Helsinki University of Technology, Finland. ld. Részletesen Teuvo KOHONEN: Self-Organizing Maps (Springer Series in Information Sciences, Vol. 30. 1995; Second extended edition, 1997).

nem érdemes használni, valamilyen módon közelebb kell húzni a hasonlókat, hogy lehetőleg maximum négy-öt szegmensre csökkenjen a számuk. Továbbá jellemeznünk kell az egyes szegmenseket. Ehhez megint a statisztikához nyúltunk, és megvizsgáltuk a kohonen háló rácspontjaiban szereplő boltok *átlagos* R^2 értékét minden egyes illesztéstípusra. *(3. táblázat, ld. a 36. oldalon)*

Ezek alapján már könnyű volt meghatározni a hasonló csoportokat, ugyanis az egyes sarkokban mindig olyan csoportokat találtunk, ahol jellemzően valamelyik illesztéstípus R^2-e volt a legmagasabb. Ez azt jelenti, hogy az első táblázatban (RSQLIN), ahol a lineáris modell átlagos R^2-ét láthatjuk, a jobb alsó csoportot vontuk össze, és lineáris viselkedésűnek feltételezzük. A második táblázatban (RSQEXP) jellemzően sok a magas R^2 érték, exponenciális viselkedésűnek azonban csak a (0;3) csomópontot tartottuk. A *3. táblázat* (RSQLOG) alapján a bal alsó csoportot vontuk össze és logaritmikusnak feltételezzük. Az utolsó táblázat (RSQQUAD) alapján pedig a (0;3) kivételével kvadratikus viselkedésű csoportot vontunk össze. *(3. táblázat)* Végül a kimaradó egyedeket szintén egy csoportba vontuk, itt egyik modell sem adott jó illeszkedést. A *3. (a,b,c,d)* ábrák egy-egy bolt idősorát és a rá illesztett modelleket mutatják. *(ld. színes Melléklet)*

A csoportok elnevezése végett azonos grafikonon ábrázoltuk az összes adott csoportba tartozó bolt értékesítési adatait. Ezek alapján négy-öt különböző viselkedésű csoportot találtunk és neveztünk el. Az elnevezések az igen szemléletes ábrák mögöttes tartalmát fejezik ki. *4. (a,b,c,d) ábra (ld. színes Melléklet)*

Az a szegmens, ahol egyetlen modell sem bizonyult használhatónak, nem reagált az akció hatására (Frigid). Az exponenciális és a kvadratikus trendet mutató szegmensek az extázisszerűen (Extasy) és a késleltetetten extázisszerűen (Delayed) reagáló szegmensneveket kapták, bár itt a különbség nem annyira szembetűnő, mint a többi csoportnál. A maradék két szegmens pedig a lineárisan növekedő (Moderate) és a gyors emelkedés után stagnáló (Jumper) szegmens volt. Összefoglalásul lássuk még egyszer, most már csoportonként szétválasztva a Kohonen grafikont! *(5. ábra) (ld. színes Melléklet)*

Most, hogy kész vagyunk az elnevezésekkel már csak egyetlen feladatunk maradt, a demográfiai adatok alapján történő besorolás. Ezt a feladatot megint egy Data Mining algoritmus, a *döntési fa* végezte el. Ez egy olyan eljárás, amikor a magyarázó változók segítségével megpróbáljuk előállítani az eredményváltozó értékét. Jelen esetben a

3. táblázat

RSQLIN Average						
	0	2	3	4	5	6
0	0.776		0.163			0.071
1	0.795		0.204			
2	0.844		0.238			**0.884**
3	0.875			0.398		**0.889**
4						**0.884**
5	0.811	0.834			**0.910**	**0.914**
6	0.793	0.838			**0.926**	**0.926**

RSQEXP Average						
	0	2	3	4	5	6
0	0.792		0.163			0.07
1	0.814		0.205			
2	0.861		0.239			**0.885**
3	**0.888**			0.402		**0.885**
4						**0.882**
5	0.801	0.820			**0.908**	**0.915**
6	0.780	0.826			**0.926**	**0.931**

RSQLOG Average						
	0	2	3	4	5	6
0	0.458		0.168			0.028
1	0.493		0.220			
2	0.565		0.215			0.697
3	0.625			0.375		0.721
4						0.747
5	**0.843**	**0.877**			0.79	0.775
6	**0.845**	**0.903**			0.772	0.727

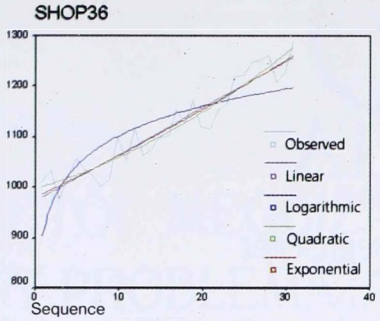
RSQQUAD Average						
	0	2	3	4	5	6
0	**0.954**		0.175			0.106
1	**0.947**		0.232			
2	**0.929**		0.255			0.889
3	**0.935**			0.406		0.889
4						0.885
5	0.866	0.897			0.911	0.916
6	0.861	0.914			**0.928**	**0.937**

magyarázó változóink a demográfiai adatbázisból rendelkezésre állnak. Eredményváltozónk pedig az öt lehetséges érték közül egyet felvevő szegmensnév. A döntési fa teljesen automatikusan keresi meg az összefüggéseket a magyarázó és az eredményváltozók között. Nézzük a végeredményt! *(4. táblázat)*

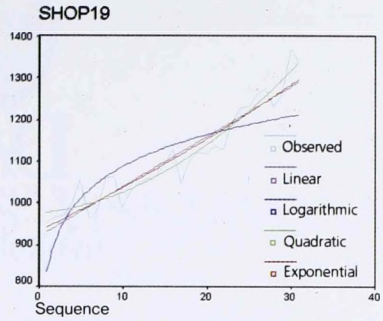
A döntési fa minden „levele" egy-egy szegmensbesorolásnak felel meg. Magát az utat a levelekig *(ág)* pedig mint döntési szabályt értelmezhetjük. Azaz például abban az esetben, ha egy bolt átlagos [Average] média ellátottsággal (Media Attendance) rendelkezett, akkor lineáris tendenciájú növekedést (moderate) mutat. Ha viszont gyenge [Poor], akkor már a vonzáskörzet átlagos jövedelmétől (Attraction Zone Income) függően lehet nem reagáló (frigid), vagy lehet késleltetett exponenciális tendenciájú (delayed). A levelek megnevezése előtt szereplő számok közül az első azt jelenti, hogy a döntési szabályt alkalmazva összesen hány bolt esik erre az ágra a mintából. A második szám pedig megmutatja, hogy ennek hány százaléka a ténylegesen olyan tulajdonságú. Azaz például az összesen 15 átlagos médiaellátottsággal rendelkező boltnak 93,3 %-a ténylegesen lineáris tendenciát mutató bolt. A fa részletes további elemzése általában már a vállalati szakértők és stratégiai döntéshozók feladata. Itt megelégszünk annyival, hogy láttuk: a döntési fa segítségével sikerült előzetes koncepciók nélkül olyan bonyolult összefüggéseket felfedezni, melyek gyakorlatilag megmagyarázzák, hogy miért lehetett sikeres egyes boltok esetében a reklámakció, és miért volt kevésbé sikeres vagy sikertelen más boltoknál.

4. táblázat

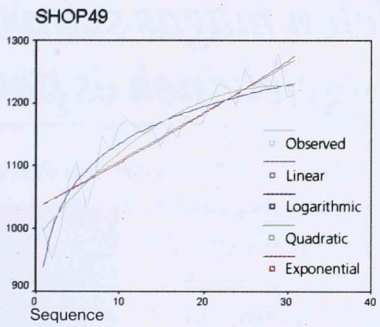
Media Attendence[Average Poor]
  Media Attendence Average (15; 0.933) → moderate
  Media Attendence Poor
    Attraction Zone Income [Average Poor Tight] (15; 0.933) → frigid
    Attraction Zone Income Well-to-do (4; 1.0) →delayed
Media Attendence Good
  Store Scale Large (7; 1.0) →jumper
  Store Scale [Medium Small]
    Region [Budapest City] (5; 0.8) →extasy
Region [Town Village] (4; 1.0) →delayed



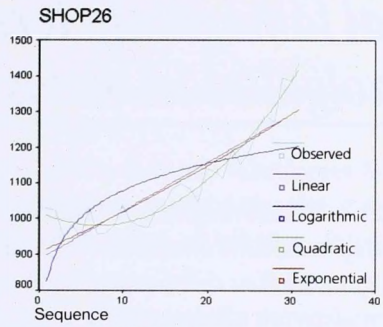
3. a) ábra: Lineáris trendet mutató forgalomnövekedés



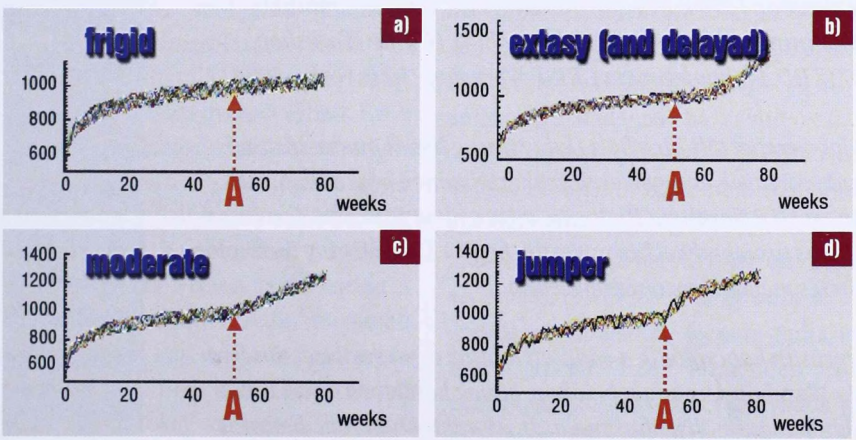
3. b) ábra: Exponenciális trendet mutató forgalomnövekedés



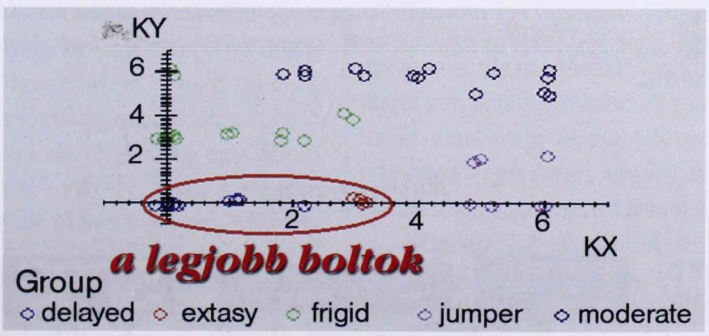
3. c) ábra: Logaritmikus trendet mutató forgalomnövekedés



3. d) ábra: Kvadratikus trendet mutató forgalomnövekedés



4. ábra: Szegmensek idősorai



5. ábra: Kohonen network