HAMPEL GYÖRGY*–FABULYA ZOLTÁN**–NAGY ELEMÉRNÉ DR.:***

Adatbányászati technikák alkalmazása magyar vállalkozások adatait tartalmazó adatbázison Microsoft Excel 2007-ben

Abstract

Using a simple data mining technique, the Analyze Key Influencers, in Excel 2007 Data Mining Add-ins, we searched for relationship among the seat (county and town), the form of business, the main activity, the number of employees and the annual income of the Hungarian companies. This technique uses the Naïve Bayes algorithm. According to the used method the seat has no influencers. Most of the main activities have no influencers, but some activities (82 out of 495) have relationship with the other criteria, mainly with the form of business. The form of business (all 30 categories), the number of employees (17 of 18 categories) and the annual income (all 9 categories) are each others key influencers. Cramer's association was used to check the results of the data mining. The Cramer contingency coefficient showed similar results as the data mining, but the results also indicated that the strength of the association was less than moderate in all cases. The highest association were between the annual income and the number of employees (0.46, moderate association), the main activity and form of business (0.36, moderate association) and the annual income and the form of business (0.27, low association).

1. Bevezetés

Az adatok bősége és a hatékony adatelemzési eszközök hiánya adatokban gazdag, de információban szegény helyzetet eredményez. A tárolt hatalmas adatmennyiség hatékony eszközök használata nélkül meghaladja az emberi elme felfogó- és befogadóképességét, így az összegyűjtött adatok ritkán látogatott „adatsíremlékekké", adatarchívumokká válhatnak (Han 2006).

Az adatok mennyiségének növekedésével egyre nagyobb szerepet kapnak az adatbányászati technikák, amelyek segítségével a hatalmas mennyiségű adathalmazból komolyabb vizsgálatok nélkül fel nem ismerhető összefüggések és ezáltal többletinformáció nyerhető ki. Az adatbányászat kifinomult statisztikai módszerek segítségével keresi az adatok közötti kapcsolatot. A módszerek egyik része keresi a mintákat és trendeket, a másik része a megerősítést (Corbit 2006).

Az adatbányászat leggyakoribb felhasználási területei: a vásárlási szokások, intenzitás, termékkapcsolatok feltárása; ügyfél szokásainak feltárása, majd ügyfélmenedzsment javítása személyre szabott szolgáltatásokkal; termelési folyamatok hatékonyságának javítása, kockázati tényezők kiszűrése. Ezen kívül minden olyan terület szóba jöhet, ahol egy rendelkezésre álló, nagy adatbázis felhasználásával előre nem ismert jellemzőket, összefüggéseket keresünk (Krotos 2002).

Az intelligens adatbányászat eszközei módszerek széles skáláját integrálják magukba: matematikai programozási és statisztikai feldolgozások mellett tartalmaznak mesterséges intelligencia kutatások során kidolgozott tanuló módszereket is; továbbá a felhasználó munkáját magas szintű vizualizációval támogatják (Sántáné et al. 2008).

* Adjunktus – Szegedi Tudományegyetem Mérnöki Kar.
** Adjunktus – Szegedi Tudományegyetem Mérnöki Kar.
*** Főiskolai tanár – Szegedi Tudományegyetem Mérnöki Kar.

A korszerű irodai programcsomagok tartalmaznak olyan eszközöket, amelyek felhasználásával információ nyerhető ki az adatbázisokból. Néhány példa az Excel 2007 táblázatkezelő esetében:

- Statisztikai függvények alkalmazása, statisztikai adatelemző eszközök;
- Adatbázis-kezelő programokból átvett szolgáltatások, például rendezés és szűrés;
- Adatlisták összesítése és elemzése kimutatások készítésével.

A Microsoft SQL Server 2008 és Data Mining Add-ins for Office 2007 felhasználásával további adatbányászati lehetőségek is rendelkezésre állnak az Excel 2007-ben. Az eszközök egy része előre beállított modellel, algoritmussal dolgozik és akár komolyabb matematikai-statisztikai ismeretek nélkül is lehetőséget nyújt gyors számítások, elemzések elvégzésére *(1. ábra)*. Emellett az adatbányászat területén jártas felhasználók számára rendelkezésre állnak olyan eszközök is, ahol a vizsgálathoz szükséges algoritmusok és modellek kiválaszthatók, paraméterezhetők.

***1. ábra.* A táblázatkezelőben elérhető egyszerű adatbányászati eszközök**
(Forrás: Data Mining Add-ins for Office 2007)

2. Cél

A Központi Statisztikai Hivatal rendszeresen közöl adatokat magyarországi társas vállalkozásokról. Ez az adatbázis többek között tartalmazza a vállalkozások nevét, cég formáját, székhelyét, fő- és melléktevékenységeit, éves árbevételét, valamint a foglalkoztatottak létszámát.

Célunk volt a Microsoft Excel 2007, Microsoft SQL Server 2008 és Data Mining Add-ins for Office 2007 programok felhasználásával vállalkozásokat tartalmazó adatbázisból megállapítani, hogy az ismérvek értékei között van-e kapcsolat.

3. Módszer

A magyarországi székhellyel rendelkező társas vállalkozások táblázatban tárolt adatai: székhely megye (nominális skála, területi ismérv), székhely település (nominális skála, területi ismérv), fő tevékenység (nominális skála, minőségi ismérv), cégforma (nominális skála, minőségi ismérv), foglalkoztatottak száma kategória (ordinális skála, mennyiségi ismérv), éves árbevétel kategória (ordinális skála, mennyiségi ismérv). A kategóriák kialakításánál a KSH csoportosítását vettük figyelembe. A táblázat 476 070 rekordot, rekordonként 6 mezőt, összesen 2 856 420 adatot tartalmazott.

A vizsgált kérdés megválaszolásához, az „Analyze Key Influencers" egyszerű adatbányászati eszközt alkalmaztuk. Az eszköz a naiv Bayes algoritmus felhasználásával számítja ki a vizsgált és a többi ismérv közötti feltételes valószínűséget. A számítások eredménye egy táblázat, amely azt tartalmazza, hogy a kiválasztott ismérvek értékei milyen relatív hatással vannak más ismérvek értékeire *(2. ábra)*.

Számolva az adatbányászat veszélyeivel – hogy ott is összefüggéseket talál számunkra a számítógép, ahol valójában nincs is (Beck-Bornholdt 1999) –, ellenőrzésként a vállalko-

zások adatait tartalmazó adatbázis adatai alapján Cramer-féle asszociációs együtthatót (C) is számoltunk. Az ismérvek függetlenségét feltételezve C két ismérv közötti kapcsolat szorosságát vizsgálja, értéke 0–1 között lehet: 0, ha az ismérvek függetlenek és 1 ha az ismérvek között függvényszerű kapcsolat áll fenn (Petres 2005).

Column	Value	Favors	Relative Impact
Foglalkoztatottak kategória	0 fő	0-20 millió Ft	
Foglalkoztatottak kategória	1 fő	0-20 millió Ft	
Cégforma kategória	betéti társaság	0-20 millió Ft	
Foglalkoztatottak kategória	2 fő	0-20 millió Ft	
Cégforma kategória	részvénytársaság	4001 millió Ft felett	
Foglalkoztatottak kategória	500-999 fő	4001 millió Ft felett	
Foglalkoztatottak kategória	300-499 fő	4001 millió Ft felett	
Foglalkoztatottak kategória	1000-1999 fő	4001 millió Ft felett	
Foglalkoztatottak kategória	100-149 fő	4001 millió Ft felett	
Foglalkoztatottak kategória	50-99 fő	4001 millió Ft felett	
Foglalkoztatottak kategória	150-199 fő	4001 millió Ft felett	
Foglalkoztatottak kategória	250-299 fő	4001 millió Ft felett	
Foglalkoztatottak kategória	200-249 fő	4001 millió Ft felett	
Foglalkoztatottak kategória	2000-4999 fő	4001 millió Ft felett	
Foglalkoztatottak kategória	20-49 fő	4001 millió Ft felett	
Cégforma kategória	korlátolt felelősségű társaság	4001 millió Ft felett	
Foglalkoztatottak kategória	5000 fő felett	4001 millió Ft felett	

2. ábra. Adatbányászat Excel 2007-ben: A 20 millió forint alatti, illetve 4 milliárd forint feletti éves árbevétel ismérvei és ismérvértékei
(*Forrás:* a szerzők saját szerkesztése)

4. Eredmények

Az asszociációs vizsgálat eredményeként egyetlen mutatószámmal megállapíthattuk, hogy milyen szoros az egyes, vizsgált ismérvek közötti kapcsolat. Az adatbányászat segítségével pedig egy részletesebb táblázatot kaptunk arról, hogy az egyes ismérvek milyen hatással vannak más ismérvek értékeikre. A publikáció terjedelmi korlátaira való tekintettel itt a vizsgálatok fő eredményei olvashatók:

– Az éves árbevétel, a foglalkoztatottak száma és a fő tevékenység nem volt meghatározó abban a tekintetben, hogy egy adott vállalkozás székhelye melyik megyében, vagy melyik településen található. Ezt mind az adatbányászat, mind a C megerősítette (a Cramer-féle együttható értékeit lásd az *1. táblázatban*).

***1. táblázat.* A székhely és egyéb ismérvek közötti kapcsolat a Cramer-féle együttható alapján**

Ismérv 1	Ismérv 2	C
Megye	Fő tevékenység	0,12
Megye	Cégforma	0,04
Megye	Foglalkoztatottak száma	0,03
Megye	Éves árbevétel	0,03
Település	Cégforma	0,13
Település	Fő tevékenység	0,11
Település	Éves árbevétel	0,09
Település	Foglalkoztatottak száma	n.sz.

(n. sz.: χ^2 p = 0,05 szinten a kapcsolat nem szignifikáns)
Forrás: A szerzők saját szerkesztése

– Adatbányászat segítségével a 495-féle fő tevékenységből mindössze 82 olyan tevékenységet találtunk, amelyeknél meg lehetett határozni jelentős mértékben ható ismérveket: a cégforma (60 tevékenység esetén), a székhely település (28 esetben) az éves árbevétel, a foglalkoztatottak száma és a székhely megye (17–17 esetben). A C együttható értékeit a *2. táblázat* mutatja. Látható, hogy ebben az esetben a mutató alapján képzett sorrend kissé eltér az adatbányászat eredményeitől.

***2. táblázat.* A fő tevékenység és egyéb ismérvek közötti kapcsolat a Cramer-féle együttható alapján**

Ismérv 1	Ismérv 2	C
Fő tevékenység	Cégforma	0,36
Fő tevékenység	Foglalkoztatottak száma	0,17
Fő tevékenység	Megye	0,12
Fő tevékenység	Település	0,11
Fő tevékenység	Éves árbevétel	n.sz.

(n. sz.: χ^2 p = 0,05 szinten a kapcsolat nem szignifikáns)
Forrás: A szerzők saját szerkesztése

– Az összesen 30 cégformára leginkább ható ismérvek az adatbányászat szerint: a fő tevékenység (26 esetben), a foglalkoztatottak száma (14 esetben), az éves árbevétel (13 esetben) és a székhely település (10 esetben). A számított C értékeket a *3. táblázat* tartalmazza. Az kapcsolat szorossága alapján képzett sorrend mindkét módszer esetében megegyezik.

***3. táblázat.* A cégforma és egyéb ismérvek közötti kapcsolat a Cramer-féle együttható alapján**

Ismérv 1	Ismérv 2	C
Cégforma	Fő tevékenység	0,36
Cégforma	Foglalkoztatottak száma	0,27
Cégforma	Éves árbevétel	0,18
Cégforma	Település	0,13
Cégforma	Megye	0,04

Forrás: A szerzők saját szerkesztése

– A 18 létszám kategóriát megvizsgálva 17 kategóriában a foglalkoztatottak létszámára legtöbbször az éves árbevétel (20 eset) és a cég formája (7 eset) volt hatással. Az adatbányászati eredmények itt is összhangban álltak a C együttható értékeivel *(4. táblázat)*.

***4. táblázat.* A foglalkoztatottak száma és egyéb ismérvek közötti kapcsolat a Cramer-féle együttható alapján**

Ismérv 1	Ismérv 2	C
Foglalkoztatottak száma	Éves árbevétel	0,46
Foglalkoztatottak száma	Cégforma	0,27
Foglalkoztatottak száma	Fő tevékenység	0,17
Foglalkoztatottak száma	Megye	0,03
Foglalkoztatottak száma	Település	n.sz.

(n. sz.: χ^2 p = 0,05 szinten a kapcsolat nem szignifikáns)
Forrás: A szerzők saját szerkesztése

– Végül, a 9 árbevétel kategória ismérvértékeire a foglalkoztatottak létszáma (20 esetben), a cégforma (9 esetben) volt jelentős hatással az alkalmazott adatbányászati módszer szerint. A C együttható értékeit az *5. táblázat* tartalmazza.

5. táblázat. **Az éves árbevétel és egyéb ismérvek közötti kapcsolat a Cramer-féle együttható alapján**

Ismérv 1	Ismérv 2	C
Éves árbevétel	Foglalkoztatottak száma	0,46
Éves árbevétel	Cégforma	0,18
Éves árbevétel	Település	0,09
Éves árbevétel	Megye	0,03
Éves árbevétel	Fő tevékenység	n.sz.

(n. sz.: χ^2 p=0,05 szinten a kapcsolat nem szignifikáns)
Forrás: A szerzők saját szerkesztése

A vizsgálatok eredményeként megállapíthatjuk, hogy a vállalkozások székhelyére a többi ismérvnek nincs hatása. Az esetek többségében a végzett tevékenységek is függetlenek a többi ismérv értékétől. A cégforma, a foglalkoztatottak létszáma és az éves árbevétel – az adatbányászat szerint – kölcsönös hatást gyakorolnak egymásra, azonban a kapcsolat erőssége – az asszociációs vizsgálat szerint – legjobb esetben is közepesnek, többnyire pedig közepesnél gyengébbnek mondható.

Irodalomjegyzék

Beck-Bornholdt, Hans-Peter et al. (1997): Der Hund, der Eier legt. Rowohlt Taschenbuch Verglag GmbH. München.

Corbitt, Terry (2006): The Power of Data Mining and Warehousing. Credit Management. Apr. 2006, p. 32–33.

Han, Jiawei et al. (2006): Data Mining. Concepts and Techniques. Second edition. Elsevier Inc. San Francisco.

Harts, Doug (2008): Microsoft Office 2007 Business Intelligence. Reporting, Analysis and Measuerement from the Desktop. McGraw-Hill. New York.

Kacsukné et al. (2007): Bevezetés az üzleti informatikába. Akadémiai Kiadó. Budapest.

Krotos László (2002): Intelligens megoldások: a döntéstámogató rendszerek világa. Kód Gazdaság- és Médiakutató Intézet. Budapest.

Petres Tibor et al. (2005): Statisztika. Phare Távoktatáás 30/2005. SZTE GTK. Szeged.

Sántáné-Tóth Edit et al. (2008): Döntéstámogató rendszerek. Panem Kiadó. Budapest.