Harmati Attila

Adatbányászat üzleti szemmel I. rész

A 21. századi szervezetek – legyenek azok akár profitorientáltak, akár non-profit jellegűek, magántulajdonban állók vagy állami érdekeltségbe tartozók – hatékony működésüket egy kiemelkedő eszközzel, az adatbányászattal támogathatják. A cikk az adatbányászat definiálása és átfogó bemutatása, illetve tudományok közti elhelyezése után annak alkalmazási területeit és alapvető feladatait részletezi. Ehhez kapcsolódóan betekintést nyújt az osztályozási feladatok elvégzése során általánosan alkalmazott módszerek mibenlétébe, a hangsúlyt a döntési fák, a neurális hálók és a logisztikus regressziós modellek elméleti alapjainak áttekintésére helyezve. A tanulmány folytatásában a most bemutatásra kerülő technológia gyakorlati felhasználása kerül szemléltetésre egy valós adatbázis elemzésére épülő osztályozási projekt ismertetésén keresztül.

Journal of Economic Literature (JEL) kód: *C25, C44, C45, C49*

Kulcsszavak: *adatbányászat, osztályozás, döntési fa, neurális háló, logisztikus regresszió*

A 20. század végére kifejlődő információs társadalom egyik velejáró következménye a rendelkezésre álló adatok óriási mennyisége lett. Ez egyaránt jelentkezik a tudományok szinte minden ágában, a természettudománytól kezdve az orvostudományon át a közgazdaságtanig. A már-már szinte elképzelhetetlen mennyiségben rendelkezésre álló adatok – például égitestek elhelyezkedése és amplitúdója, diagnosztizált betegségek és azok kimenete, vagy egyszerűen csak egy vásárlás tételei – paradox módon mégsem segítik az érintett tudományok képviselőit. Épp ellenkezőleg, mivel méretükből adódóan eltakarják magukból a hasznosítható információkat, ezáltal a felmerülő kérdések megválaszolását még inkább nehezítve *(Adriaans–Zantinge 2002)*.

A probléma megoldásaként fejlődött ki az adatbányászat technológiája, melynek célja az értékelhető, valamilyen módon hasznosítható információk adatbázisokból történő kinyerése. Az adatbányászat alkalmazói képessé válnak adatbázisaik lényegi vetületére koncentrálni, ezáltal segítve a döntések jobb meghozatalát, a rendelkezésre álló források hatékonyabb kihasználását, a pontosabb kutatási eredmények elérését, illetve az esetleges versenyelőny megszerzését *(Thearling 2009)*.

Ezen megfontolások zöme különösen az üzleti élet szereplői számára lényegesek, ugyanis a modern közgazdasági felfogás szerint az információ – a munkaerő, a tőke és a föld után – a negyedik fő termelési tényező. Ennek oka, hogy a materiális erőforrások mellett csakis a

Harmati Attila a Debreceni Egyetem Közgazdaságtudományi Karának végzett hallgatója, a SAS Institute Kft junior elemzője. E-mail: harmatiati@gmail.com

megfelelő információ birtokában lehet időt spórolni és költségeket csökkenteni, valamint előnyös üzleteket kötni, bevételt, így profitot növelni *(Adriaans–Zantinge 2002)*.

Az adatbányászat definiálása

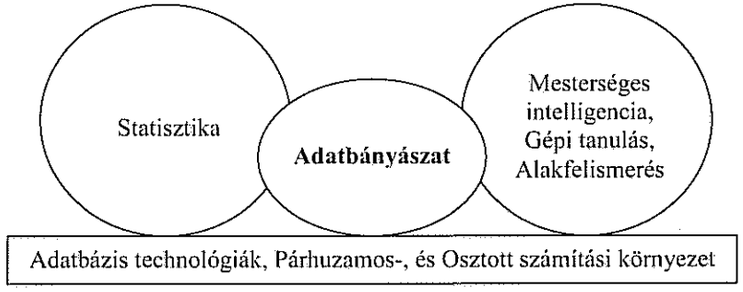
A korábbi gondolatmenetet folytatva, napjaink közgazdasági felfogása szerint a vállalatok egyik nélkülözhetetlen termelési tényezője az adatokban megtestesülő tőke, az információ. Az irányzat képviselői szerint az adatbázisokkal – a tőkéhez hasonlóan – racionálisan kell gazdálkodni, működtetni és kezelni kell azokat. Ennek oka, hogy az adatok tömegeinek kialakulásával párhuzamosan azok a 21. század legfontosabb erőforrásává váltak. Azonban az adatoknak önmagukban semmi értelmük, azokat az értékteremtő folyamatok során felhasználható információvá kell alakítani oly módon, hogy a végrehajtandó döntéseket tudás formájában támogathassák. Az adatvagyon ilyen irányú hasznosításának folyamata a világ egyik legfiatalabb tudományában, az adatbányászatban ölt testet *(Adriaans–Zantinge 2002)*.

Az adatbányászat definíció szerint „rejtett, ismeretlen, potenciálisan hasznos tudás kinyerése az adatokból, nem triviális módon" *(Adriaans–Zantinge 2002: 17.)*. A kinyerési folyamat során nagyméretű adatállományok feldolgozása történik meg magas szintű matematikai és statisztikai módszerek[1], tanuló algoritmusok, modellezési technikák és a mesterséges intelligencia automatikus, illetve félig automatikus alkalmazása által *(Fajszi–Cser 2004)*.

Az adatbányászathoz jellegzetesen kötődő tudományágak kapcsolatát az 1. ábra szemlélteti, melyen az adatállományok feldolgozására használható különböző területek, és az azokat támogató informatikai háttér integrációja látható.

1. ábra

Az adatbányászat mint tudományok keveréke



Forrás: Tan et al. 2005 alapján.

[1] *Az adatbázisokból lekérdezés útján, aggregálással, illetve alapstatisztikai vizsgálatokkal történő információnyerés nem tekinthető adatbányászatnak.*

Az adatbányászat tömör, de igen lényegre törő fogalma *Fajszi-Cser 2004* szemléltető gondolatával egészíthető ki, mely szerint az egy olyan információfeltárási folyamatot jelent, mely során az adatok között fennálló – a gyakorlati életben is jól hasznosítható –, korábban nem ismert összefüggések feltárása történik meg.

Az adatbányászat által megtestesített információfeltárást támogató informatikai eszközök alapvetően az úgynevezett szerver-kliens kapcsolatra épülnek. Ez egy olyan informatikai architektúrát takar, mely kommunikációjában a résztvevő két fél nem egyenrangú, szerepeik megosztottak. Ezt a kommunikációt jellemzően a kliens kezdeményezi valamilyen adatfeldolgozási kéréssel, például lekérdezés vagy egy művelet végrehajtása érdekében. A szerver erre reagálva végrehajtja a szükséges lépéseket, és megválaszolja a kérést. Az adatok adattárházakban[2] tárolása, illetve annak szerver-kliens kapcsolattal történő integrálása nagymértékben felgyorsítja a vállalatok rendelkezésre álló adatainak kezelését, illetve feldolgozását *(Ullman–Widom 2008)*. Ezzel összefüggésben az úgynevezett párhuzamos számítási környezet lehetővé teszi egy számítógépes program különböző részeinek párhuzamos futtatását egyszerre több processzoron, az osztott számítási környezet pedig azoknak egyidejű alkalmazását, valamint egyszerre több számítógépen és hálózaton belüli kommunikációját, ezzel még inkább lerövidítve az információfeltárás érdekében elvégzendő feladatok végrehajtási idejét *(Viczián 2002)*.

A tudásfeltárás folyamata

Az adatbányászat tágabb kontextusban a tudásfeltárás[3] 2. ábrán látható folyamatának egyik legfontosabb elemét jelenti. Ez a folyamat több, egymáshoz szervesen kapcsolódó lépés láncolata, mely során a hatékonyság elősegítése érdekében az adatbányásznak és az alkalmazási terület szakértőjének mindvégig szorosan együtt kell működnie. A folyamat főbb lépései a mintavételezés, az adatok előfeldolgozása, az adatok szükség szerinti transzformálása, az adatbányászat alkalmazása, és a kinyert tudás értékelése *(Bodon 2009)*.

A tudásfeltárás folyamatának megkezdéséhez az érintett alkalmazási terület alapos ismerete szükséges, mivel csakis ezen ismeret birtokában lehet olyan céladatbázist létrehozni, mely elemzése révén használható tudás tárható fel. Az elemezni kívánt adatok előfeldolgozása során az adatbázis zajoktól – azaz az adatokba épült véletlen hibáktól – való megtisztítása történik meg, illetve szükség esetén az osztott adatbázisok integrációja. Ezt követheti a feltárás céljából kiemelkedően fontos attribútumok – azaz változók – kiemelése a mintatér csökkentése érdekében, majd az alkalmazandó adatbányászati algoritmus típusának – azaz a végrehajtandó adatbányászati műveletnek – a definiálása. A kijelölt műveletnek megfelelő módszerek által felépített modellek alkalmazásával lehet a rejtett tudást jelentő összefüggéseket, mintázatokat feltárni, melyeket a folyamat utolsó lépéseként meg kell erősíteni az előzetes elvárásokkal és ismeretekkel történő összevetés révén. Amennyiben a feltárt tudás érvényesnek, újszerűnek és hasznosnak bizonyul,

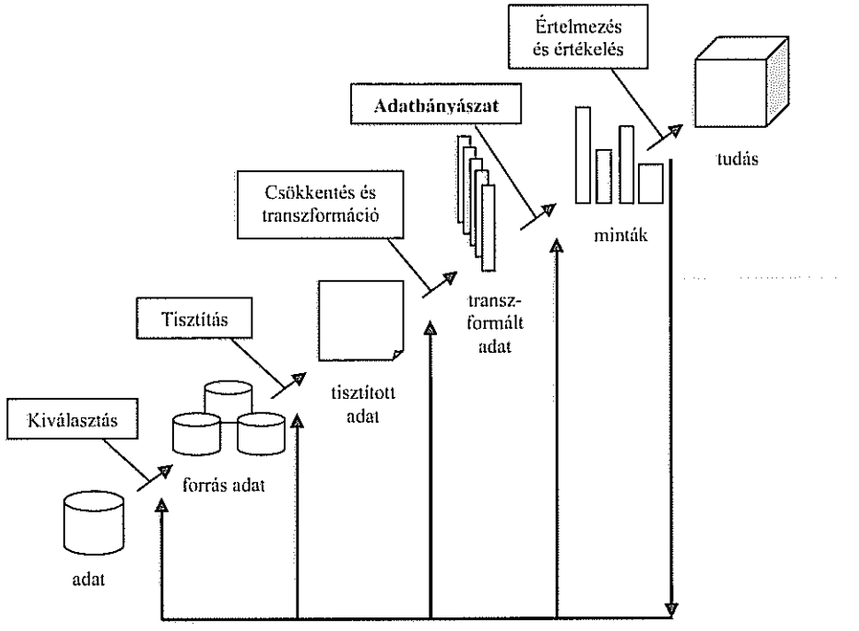
[2] *Az adattárház „egy témaorientált, integrált, nem változó időváriáns adatrendszer, melynek elsődleges célja a stratégiai döntések támogatása, a vezetői igények kielégítése" (Inmon 2005:29). Az adattárház jelenti egy szervezet adatainak legfőbb tárhelyét, melynek célja, hogy a szervezet különböző egységeinek informatikai rendszereiből átvett adatokat tárolja, és azokat szükség esetén kiszolgáltassa (Ullman–Widom 2008).*

[3] *A tudásfeltárást a „Knowledge Discovery in Databases" elnevezés alapján a szakirodalom gyakran csak KDD-ként említi (Adriaans–Zantinge 2002).*

megtörténhet annak dokumentálása, interpretálása. Ennek során kulcsfontosságú a feltárt információk megértethetősége az érintett személyekkel, így kiemelkedő szerepet játszanak a grafikus megjelenítés igen változatos eszközei. Amennyiben az értékelés során problémák merülnek fel, bizonyos elemek megkérdőjeleződnek, a folyamat visszatér egy korábbi szakaszába, és megismétlésre kerülnek az egyes lépések *(Bodon 2009)*.

2. ábra

A tudásfeltárás folyamata

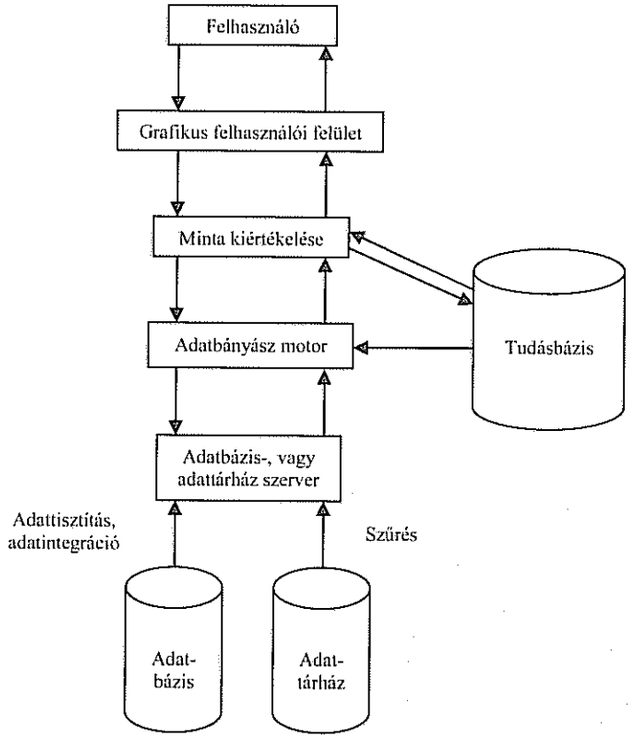


Forrás: Bodon 2009 alapján.

Az adatbányászati rendszernek kapcsolatban kell állnia mind a felhasználóval, mind az adatbázissal *(Bodon 2009)*. Ezt szemlélteti a 3. ábrán látható architektúra.

3. ábra

Tipikus adatbányászati rendszer architektúrája



Forrás: Bodon 2009: 18.

Az ábra alján látható az adatbázis, illetve az adattárház, melyek a szervezet adatainak tényleges tárhelyét jelentik. A tárhelyek és a felhasználó között a szerver teremt kapcsolatot – többek között – a szükséges adatok szolgáltatása révén. A tudásbázis az alkalmazási területre jellemző, formalizálható tudást jelenti, mely segítségével a mintatér hatékonyabban szűkíthető, egyes paraméterek, küszöbszámok pontosabban meghatározhatók. Az adatbányászmotor az adatbázist és a tudásbázist összefogva alkalmazza a felhasználó által szükségesnek tartott algoritmusokat, melyek eredményének értékelése a feltárt összefüggéseket és mintázatokat kiértékelő modulban történik meg. Ezeket az eredményeket a felhasználó a – közte és az adatbányászati rendszer között kapcsolatot teremtő – grafikus felhasználói felület által ismerheti meg, és kommunikálhatja tovább az érintettek felé *(Bodon 2009)*.

Alkalmazási területek

Az adatbányászat alkalmazhatóságáról általánosságban elmondható, hogy az szinte bárhol felhasználható, ahol nagy adathalmazok léteznek. Az üzleti életben ez különösen

olyan területeken jellemző, ahol könnyen lehet adatokat gyűjteni, például a partnerek között lezajlott tranzakciós ügyletek alapján, vagy épp a jövőbeli ügyleteik feltételéül szabott adatszolgáltatás révén. Az információfeltárás nyújtotta lehetőségek kiaknázása egyelőre igazán széles körben a felhasználás által elérhető kiemelkedően magas megtérüléssel kecsegtető, illetve a felhasználás hiányában hatalmas veszélyeket magában rejtő üzletvitel esetén terjedt el a szükséges szoftverberuházás és szakértőgárda – kiszervezés esetén pedig a szolgáltatás igénybevételének – magas költségei miatt, illetve a prudens működés fenntartásának biztosítása érdekében *(Adriaans–Zantinge 2002)*.

Mivel az adatbányászat által az ügyfelek jellemzőit tartalmazó adatbázisok elemzése révén felhasználható tudás nyerhető ki, így statisztikailag megalapozott becslések tehetők bizonyos jövőbeni változásokra, és események bekövetkezésének valószínűségére *(Szűcs 2004)*. Általánosságban megállapítható, hogy az adatbányászat két legjellemzőbb alkalmazási területe a szervezetek marketingtevékenységéhez, illetve a pénzügyi szervezetek általános működéséhez köthető. Előbbi eset két alkategóriára bontható, az ügyfélszegmentáció és a direktmarketing elkülönülő – mégis azonos, a nyereség maximalizálását szolgáló cél – tevékenységeire. A pénzügyi szervezetek ettől kissé eltérően – bár tagadhatatlanul ugyanazon végső célt követve – a törvényi előírásoknak megfelelően, illetve saját létfenntartásuk szem előtt tartásával az adatbázisukban rejlő információkat zömében hitelezési tevékenységükhöz kapcsolódóan használják fel *(Fajszi–Cser 2004)*.

A vállalatok értékesítésének kulcsfontosságú feltétele piacaik megfelelő fokú ismerete, ennek alapvető eszköze az ügyfélszegmentáció. Ennek során egy szervezet lehetséges ügyfeleinek csoportokba sorolása történik meg oly módon, hogy az egy szegmentumba tartozó egyének egymásra jobban hasonlítanak, mint egy másik csoportba tartozó társukhoz, azaz a csoportok befelé homogén, kifelé pedig heterogén sajátosságokat kell hogy mutassanak *(Kotler–Keller 2006)*. Az adatbányászat az egyének demográfiai, tranzakciós, és – az azokból képzett – dinamikus tulajdonságok elemzése révén a különböző jellegzetességeket mutató csoportok definiálásában játszik szerepet, segítségével a marketingszakemberek a hagyományos módszerekhez képest hatékonyabban választhatják ki, és célozhatják meg a legjövedelmezőbb potenciális fogyasztói csoportokat *(Fajszi–Cser 2004)*. A hatékony szegmentáció ezen túlmenően a marketingstratégiák eredményes végrehajtásának, és az ügyfélkapcsolat-menedzsment kialakításának alapját képzi, így minden túlzás nélkül állítható, hogy ez az adatbányászat egyik legfontosabb alkalmazási területe *(Adriaans–Zantinge 2002)*.

Az adatbányászat másik kiemelkedő alkalmazási módja a direktmarketinghez kapcsolódik. A direktmarketing a marketingkommunikáció egy olyan eleme, mely közvetlen csatornák – például direkt mail, katalógus, telemarketing, internetes oldalak és egyéb mobil eszközök – felhasználását jelenti a fogyasztók árukkal és szolgáltatásokkal való, közvetítők nélküli kiszolgálása érdekében. A módszer alapvető oka a költség- és időmegtakarító hatás mellett a fogyasztók részéről megnyilvánuló – az individualizálódó társadalom következtében kialakuló – egyéni figyelemre vonatkozó igény *(Kotler–Keller 2006)*. Mivel a felsorolt eszközök hatása közvetlenül mérhető, így a korábbi kampányok eredményeit tartalmazó, az ügyfelek reakcióit is felvonultató adatállományok elemzésével olyan modellek állíthatók fel, melyek segítségével az ügyfelek egy esetleges megkeresésre történő válaszadásának valószínűsége alapján értékelésre kerülnek. Az értékelés általánosításának eredményeként az ügyfelek egy széles körére vonatkozó, véletlenszerű kiválasztáson alapuló direktmarketing-kampány helyett csak egy olyan szűk csoportot

kell megcélozni, mely pozitív válaszadásának valószínűsége megfelelően magas[4]. Ennek hatására jelentősen csökkenthetők – a munkaerővel kapcsolatos, a csatornahasználati és egyéb kiadások kordában tartása által – a kampányköltségek. A költségek ésszerű szinten tartása mellett további előny a potenciális, valamint a bizonytalan vásárlókat személy szerint a leginkább érdeklő témákban megfogalmazott és célzottan eljuttatott üzenetek által elérhető hatékony értékesítés és kapcsolatépítés *(Fajszi–Cser 2004)*.

Az adatbányászat pénzügyi szervezetek világában történő alkalmazásának tárgyalása előtt szót kell ejteni arról, hogy a benyújtott hitelkérelmek elfogadásával kapcsolatban a hitelintézeteknek bonyolult intézkedések sorát kell megtennie. Kintlévőségeik kockázatának minimalizálása és a kérelmek elfogadásának optimalizálása érdekében alkalmazzák az egyedi hitelbírálat rendszerét, mely a hitelpontozási rendszerből, a verifikációból, a hitelkapacitás vizsgálatából és a megkövetelt jelzálogtárgy értékének meghatározásából áll. A hitelpontozási rendszer egy olyan kockázatértékelési rendszert jelent az ügyfelek hitelképességének megítélése céljából, mely kidolgozása során – az adatbányászat révén – a rendelkezésre álló adatok alapján könnyen kezelhető, és bárkire alkalmazható modellt alakítanak ki *(PSZÁF 2001)*. Ez a modell alkalmas arra, hogy mind a vállalati, mind a lakossági ügyfeleket kockázati kategóriákba sorolja. Az így létrehozott szofisztikált kockázatelemzések, illetve a kockázatok nagyságának egyedi, és iparági szintű pontos becslése a hitelintézetek likviditásának biztosítása[5] mellett jó eséllyel növeli azok nyereségességét is. Az adatbányászat ugyanakkor a hitelpontozás mellett az ügyfelek egy esetleges fizetésképtelensége esetén alkalmazandó, leghatékonyabb behajtási módszer meghatározására, a csalásgyanús ügyletek felismerésére, valamint az egyes jövedelmi sajátosságokat mutató ügyfélcsoportoknak kínálandó extra szolgáltatások körének meghatározására is felhasználható *(Szűcs 2004)*.

Feladatok és műveletek

Az adatbányászat feladatai az 1. táblázatban látható módon, alapvetően két kategóriába sorolhatóak annak megfelelően, hogy azok előrejelző vagy leíró funkciót töltenek be. Az előrejelző módszereket szokták prediktív módszereknek is nevezni. Ezek célja olyan modellek felállítása, melyek egy ismeretlen értékű, kitüntetett szerepű célváltozó jelenbeli vagy jövőre vonatkozó értékének becslésére alkalmasak a rendelkezésre álló adatokból. A prediktív módszerekkel szemben a leíró módszerek nem rendelkeznek kitüntetett szerepű célváltozóval, helyette az adathalmaz valamely nem triviális, mégis arra jellemző és jól interpretálható tulajdonságának felkutatását szolgálják *(Tan et al. 2005)*.

[4] *Adriaans–Zantinge (2002) könyvükben azt írják, hogy a randomizált módon történő kiválasztás alapján küldött levelek legfeljebb 3–4%-ra érkezik válasz. Ezzel szemben a tudatosan kiküldött levelek közel 20–25%-a is pozitív válaszra találhat.*

[5] *A Bázel II. egyezmény a likviditás és szolvencia fenntartása érdekében minimális tőketartalék létét követeli meg. Kedvezményként a hitelkockázatok belső minősítési rendszer alapján történő értékelése esetén alacsonyabb, a közgazdaságilag szükséges tőke mértékéhez közelítő tőkekövetelményt ír elő (PSZÁF 2005).*

Az előrejelző módszereknek számos célja lehet, melyek alapján megkülönböztethető az osztályozás, a becslés és a deviációk megtalálásának művelete. Az osztályozási műveletek során olyan modellek felállítása történik meg, melyek segítségével valamely meghatározott szempontból – a célváltozó lehetséges értékei[6] alapján kategorizálandó egyedek (azaz rekordok) hovatartozása (azaz osztályozási címkéje) azok különböző tulajdonságai (azaz attribútumérték-kombinációi) alapján meghatározhatóvá válik. A becslés művelete során felállítandó modellek feladata a vizsgált folytonos célváltozó értékének egy jövőbeli időpontra történő meghatározása, a deviációk keresésének művelete pedig a különleges üzleti partnerek megtalálására specializálódott *(Tan et al. 2005)*.

1. táblázat

Az adatbányászat feladatai és műveletei

Feladat	Műveletek
Előrejelző	Osztályozás
	Becslés és előrejelzés
	Deviációk keresése
Leíró	Szegmentálás
	Vásárlói kosár elemzése
	Szekvenciális mintázatok felismerése

Forrás: Tan et al. 2005 alapján.

A leíró módszerek az adathalmaz tulajdonságainak különböző célú feltárása alapján a szegmentálási műveleteket, a vásárlói kosár elemzésének-, valamint a szekvenciális mintázatok felismerésének műveleteit ölelik át. A szegmentálási műveletek az adatok hasonlóságon, illetve különbségen nyugvó csoportosítását jelentik, melyek révén az adatok egy meghatározott körének közös jellemzői is definiálásra kerülnek. A vásárlói kosár elemzése asszociációs kapcsolatok feltérképezésére szolgáló modellek felállítására alkalmas, a szekvenciális mintázatok felismerése pedig időben egymást követő események tényének megállapítására használható *(Tan et al. 2005)*.

A különböző műveletek leghatékonyabb végrehajtása érdekében a kívánt modellek felállítására más-más módszerek alkalmasak. Jelen tanulmány további részében a figyelem középpontjába az osztályozási műveletek módszerei kerülnek, melyek alapszintű ismerete elengedhetetlen a későbbiekben ismertetésre kerülő adatbányászati projekt értelmezéséhez.

[6] *Ilyen kimenetek lehetnek például egy direktmarketing-kampány esetén a megkeresésre reagálók, illetve nem reagálók, a hitelintézetekhez kapcsolódóan pedig a felvett hitelt visszafizetők, illetve vissza nem fizetők.*

Osztályozási módszerek

Tan et al. 2005 könyvükben részletezik az osztályozási feladatok során alkalmazható módszereket, melyek leggyakrabban a következők:

- Bayes-osztályozás: naiv Bayes-osztályozó és Bayes-hálók,
- döntési fák,
- neurális hálók,
- példány alapú módszerek,
- szabály alapú módszerek,
- SVM-módszerek[7].

A következőkben a felsorolt módszerek közül a későbbi projekt szempontjából kiemelkedő jelentőségű döntési fák, neurális hálók és logisztikus regressziós modellek elméleti alapjai kerülnek részletes ismertetésre. Azonban pár gondolat erejéig a többi módszer mibenlétével is szükséges foglalkozni.

A naiv Bayes-osztályozók olyan statisztikai osztályozók, melyek az elemek egyes osztályokhoz való tartozásának valószínűségét becslik azon feltételezés mellett, hogy az attribútumértékek egy adott osztályra gyakorolt hatása független más attribútumok értékétől. A Bayes-hálók ezzel szemben a grafikus megjelenítés segítségével alkalmasak az attribútumok részhalmazai közti összefüggések szemléltetésére is *(Han–Kamber 2004)*.

A példány alapú tanulás alapötlete az analógiák vizsgálatán nyugszik, leggyakoribb alkalmazási módja a *k* legközelebbi szomszéd módszere. Az eljárás lényege, hogy az ismert osztályozási címkével rendelkező rekordokat az attribútumok számának megfelelő többdimenziós térben egy-egy pontként kell kezelni, és az ismeretlen címkéjű elemet köztük elhelyezni. Ezután az euklideszi távolság segítségével mért, *k* darab legközelebbi szomszéd leggyakoribb osztályozási címkéjének megfelelően történik meg a vizsgált elem kategorizálása *(Fajszi–Cser 2004)*.

A szabály alapú módszereket alapvetően két csoportba lehet sorolni, a direkt és az indirekt típusba. Direkt eljárás esetén az osztályozási szabályok különböző módszerek segítségével, közvetlenül az adatokból kerülnek meghatározásra, indirekt eljárás esetén pedig korábban már elkészített modellek felhasználásával történik azok megfogalmazása *(Tan et al. 2005)*.

Az SVM-módszerek működésének lényege, hogy az eredeti megfogalmazásban komplex, nemlineáris megoldást igénylő feladatok nemlineáris transzformációk segítségével egy, a bemeneti mintatér dimenziójánál több dimenziós térbe transzformálva, lineárisan megoldhatóvá válnak *(Valyon 2007)*.

Döntési fák

Az egyik legelterjedtebb osztályozási módszer a döntési fa, mivel általa a modellépítés során felhasznált objektumok egymást kölcsönösen kizáró osztályokra történő bontásának grafikus megjelenítéséből egyszerűen leolvashatóvá és interpretálhatóvá válnak az osztályozási szabályok *(Fajszi–Cser 2004)*.

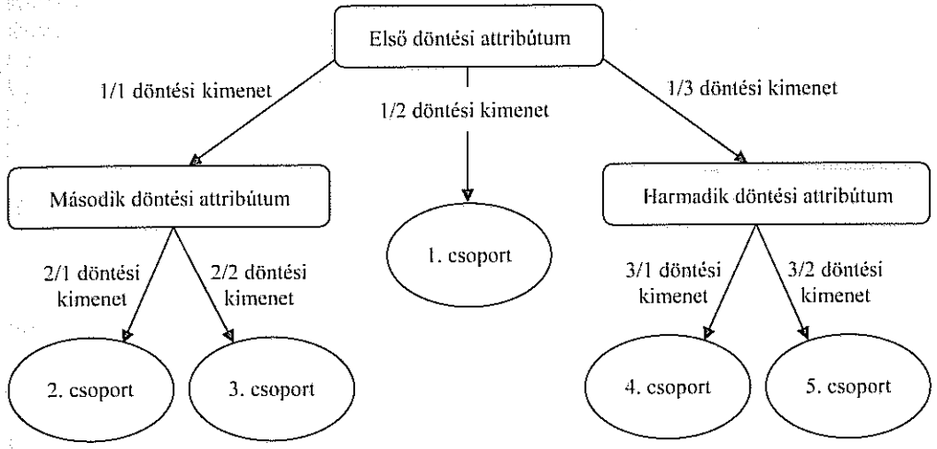
[7] *Support Vector Machines*

Definíció szerint „a döntési fa adatok alapján, logikai értéket produkáló következtetési szabályok hierarchikus sorozatának ábrázolási módja" *(Fajszi–Cser 2004:167)*.

Han–Kamber 2004 a döntési fát egy olyan gráfként határozzák meg, melyben a belső csomópontok az egyes attribútumokra – azaz az adott objektumok valamilyen tulajdonságára – vonatkozó döntési helyzeteket, az ezekből induló ágak ezen döntések lehetséges kimeneteit, az azokból származtatott levelek pedig az így képzett osztályokat jelentik. Erre látható példa a 4. ábrán. Az ábrának megfelelő egyszerű esetben az első döntési attribútumnak három lehetséges kimenete van, melyből két újabb attribútum és egy osztály adódik, majd az újabb attribútumokból két-két újabb, mind az előzővel, mind pedig egymással diszjunkt osztály keletkezik.

4. ábra

A döntési fa általános sémája



Forrás: Han–Kamber 2004 alapján

A döntési fa elmélete *Han–Kamber (2004)*, valamint *Berry–Linoff (1997)* adatbányászati technikákkal foglalkozó könyvei alapján kerülnek ismertetésre.

A fejjel lefelé ábrázolt fa felépítése annak *S*-sel jelölendő csúcsából, tehát ennek megfelelően a fa gyökeréből indul, mely *s* darab elemet tartalmaz. Első lépésben ezekre az elemekre kell megkeresni azt az attribútumot – vágóismérvet –, ami a kívánt célnak leginkább megfelelő módon választja részekre azokat. Ez a vágóismérv az entrópia, a téves osztályozási hiba[8], illetve a tanulmány második részében ismertetésre kerülő projektben definiálandó Gini-index módszereivel választható ki, melyek a vágás eredményeként létrehozandó részek homogenitásában megnyilvánuló nyereséget más-más módon értékelik, és az értékelésük alapján maximális nyereséget biztosító attribútumot választják.

Az előző módszerek közül legáltalánosabban az entrópián alapuló technika alkalmazott, melynek áttekintésén keresztül reális kép alkotható a fák kialakításának alapjairól. A

[8] *A téves osztályozási hiba az* $Error(t) = 1 - max_i P(i|t)$ *képlettel számolható ki, ahol i egy esemény kimenetei számát, t pedig a csomópontokat jelöli (Tan et al. 2005).*

módszer alapgondolata, hogy az egyes változókkal végzett esetleges particionálásból eredő információnyereségek meghatározása által definiálható a vágóismérv, ugyanis a legnagyobb nyereséget biztosító attribútum minimalizálja az osztályozáshoz szükséges információ mértékét és eredményez minimális osztályozási hibát.

Jelölje C_i az osztályozás által létrehozni kívánt csoportokat, és s_i a C_i osztályba eső elemek számát, ahol $i=1,\ldots,m$. Ezen jelölések segítségével a minta osztályozásához várhatóan szükséges információ $I(s_1,\ldots,s_m)$ nagysága:

$$I(s_1, \ldots, s_m) = -\sum_{i=1}^{m} p_i log_2(p_i), \qquad (1)$$

ahol p_i egy elem C_i osztályba való esésének valószínűsége.

Legyen egy A attribútumnak $j=1,\ldots,v$ lehetséges kimenete, mely alapján az S csúcs v darab részhalmazra bontható, jelölje ezeket S_j. Az ebbe az S_j részhalmazba eső elemek száma legyen s_j, melyekből a C_i osztályba eső elemek száma s_{ij}.

Az S_j részhalmazra vonatkoztatva a várhatóan szükséges információ $I(s_{1j},\ldots,s_{mj})$ nagysága:

$$I(s_{1j}, \ldots, s_{mj}) = -\sum_{i=1}^{m} p_{ij} log_2(p_{ij}), \qquad (2)$$

ahol p_{ij} az S_j részhalmazbeli elemek C_i osztályba esésének valószínűsége.

Ezek felhasználásával az A attribútum $E(A)$ entrópiája, tehát az A változó szerinti osztályozás információ-szükséglete:

$$E(A) = \sum_{j=1}^{v} \frac{s_{1j} + \cdots + s_{mj}}{s} I(s_{1j}, \ldots, s_{mj}). \qquad (3)$$

A kiszámított információ-szükséglet és az egyes változók entrópiái által a vágásra használandó attribútum az alábbi szabály alapján határozható meg:

$$\max_A \{Információnyereség(A)\} = I(s_1, \ldots, s_m) - E(A). \qquad (4)$$

A maximális információnyereséget biztosító attribútum alapján részekre osztható a fa csúcsa, az elemek a kimeneteknek megfelelően particionálhatók. Az így létrehozott részhalmazok a belső csomópontok, melyekre az algoritmus megismételendő. Az eljárást addig kell folytatni, míg minden részhalmaz homogén nem lesz, az így létrehozott részhalmazok lesznek a fa levelei. Emellett akkor sem folytatandó tovább a fa szerkesztése, ha az elér egy előre meghatározott mélységet, vagy egyéb, a csomópontokra vonatkozó kritérium, például információnyereségi kritérium ezt meg nem követeli.

Az elkészült fa eredményei az úgynevezett „Ha-Akkor" szabályok feltárása által válnak megismerhetővé – és a szakavatatlan szemek számára is könnyen megérthetővé –, melyek

egyszerűen alkalmazhatóak egy osztályozási címkével nem rendelkező adatállomány kategorizálására. A szabályrendszer felállításához a fa gyökerétől a levelek irányába haladva az egyes döntési attribútumokat a hozzájuk tartozó kimenetekkel kell tekinteni, és az így feltáruló vizuális útvonalakat egymást kiegészítő szabályrendszerré integrálni.

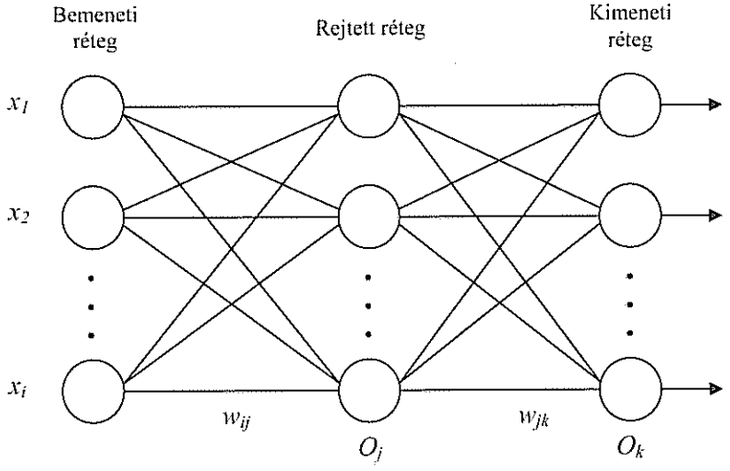
Neurális hálók

A neurális háló széles körű felhasználhatóságának köszönhetően a döntési fa mellett az osztályozási feladatok szintén igen népszerű eszköze. *Elsberry (1998)* a neurális hálót szoftver megvalósítású, párhuzamos, osztott működésre képes információfeldolgozó eszköznek definiálja. *Kristóf (2002)* egy ettől jóval szemléletesebb megfogalmazást ad, miszerint a neurális háló neuronok olyan rendszere, amely i bemenettel és k kimenettel rendelkezik, és amely az i dimenziós bemeneti vektorokat k dimenziós kimeneti vektorokká alakítja át az információfeldolgozás során.

A neurális hálók egy speciális fajtáját, az előrecsatolt többrétegű hálózatot *Han–Kamber (2004)* adatbányászati technikákkal, valamint *Bishop (1995)* neurális hálókkal foglalkozó könyve alapján kerülnek ismertetésre. Ezen hálótípus grafikus megjelenítése az 5. ábrán látható. Amennyiben egy rejtett réteget tartalmaz a háló, úgy az egyes rétegekből származó kimeneti egységek számának megfelelően kétrétegű hálónak nevezzük azt. Amennyiben az egységeket összekötő élek egyike se tér vissza egy korábbi réteg kimeneti egységébe, azt előrekapcsolt hálónak hívjuk. Ezek a fogalmak hamarosan tisztázásra kerülnek.

5. ábra

Előrecsatolt, többrétegű neurális háló



Forrás: Han–Kamber 2004: 309

Az ábrán látható a háló három rétege: a bemeneti, a rejtett és a kimeneti réteg. A bemeneti réteg a hálóba táplált adatokból áll, mely a minta attribútumainak felel meg. Ebben a rétegben az i-edik bemeneti egység I_i bemenete megegyezik annak O_i kimenetével, ahol

$i=1,\ldots,m$. Ezek a kimenetek lesznek a rejtett réteg bemenetei azon feltétel mellett, hogy a bemeneti réteg minden eleme – az úgynevezett élek által – összeköttetésben áll a rejtett réteg minden elemével. A köztük fennálló kapcsolat szorosságát egy kezdetben definiált súlyérték reprezentálja, melyet az i-edik bemeneti rétegbeli elem és a j-edik rejtett rétegbeli elem között w_{ij} jelöl. A j-edik rejtett rétegbeli elem I_j bemeneti értéke a bemeneti elemeknek a j-edik egység θ_j torzításával korrigált súlyozott összege:

$$I_j = \sum_i w_{ij} O_i + \theta_j. \tag{5}$$

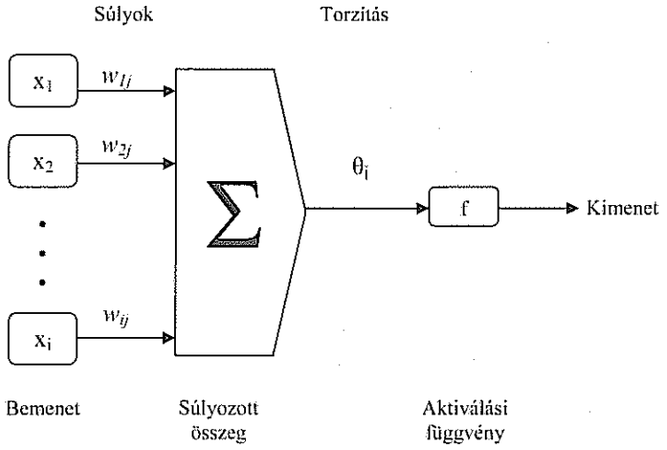
Az így megkapott rejtett rétegbeli elem I_j bemenete a széles értelmezési tartományt a 0;1 intervallum határai közé szorító aktiválási függvénnyel transzformálható annak O_j kimenetévé az alábbi módon:

$$O_j = \frac{1}{1 + e^{-I_j}}. \tag{6}$$

Ezt az O_j elemet a neuron értékének nevezzük. Előbbiek ismeretében definiálható a neuron, ami egy olyan több bemenetű, egy kimenetű eszköz, melynek kimenete a bemenetek lineáris kombinációjaként előálló érték nemlineáris függvénye *(Álmos et al. 2002)*. Az imént bemutatott folyamatot szemlélteti a 6. ábra.

6. ábra

Egy neuron sematikus képe



Forrás: Han–Kamber 2004:311

A bemeneti réteg és a rejtett réteg elemeihez hasonlóan a rejtett rétegbeli elemek is összeköttetésben állnak a kimeneti rétegbeli elemekkel, azok bemeneti elemeiként funkcionálva. A j-edik rejtett rétegbeli elem és a k-adik kimeneti rétegbeli elem közti kapcsolatot reprezentáló súlyt w_{jk} jelöli. A kimeneti rétegbeli elemek bemeneti értékeit az előző réteg kimeneti elemeinek a k-adik egység θ_k torzításával korrigált súlyozott összege adja, kimeneti értéke pedig az aktivitási függvénnyel számolható ki. Ez a kimenet az aktuális bemenetekhez tartozó osztályozási címkéket jelenti, így kategorizálva a vizsgált mintát.

A neurális hálóhoz kapcsolódó eljárás központi feladata az elemek közti súlyok optimális értékének megtalálása, melyek összességükben minimalizálják a bemenetek osztályozási címkéjének kialakításához feltárt szabályszerűség alapján meghatározott O_k aktuális, és T_k tényleges kimenetek közti négyzetes hiba nagyságát. Ez formálisan az alábbi:

$$\sum_k (T_k - O_k)^2 \rightarrow min. \tag{7}$$

A megfelelő súlyok a tanulási folyamat révén hangolhatóak be. A leggyakrabban alkalmazott neuronhálós algoritmus a hiba-visszacsatolásos algoritmus, mely során a hibák a súlyok és torzítások módosításával csatolhatók vissza, ezáltal pontosítva a hálót.

A kimeneti réteg k-adik egységénél lévő Err_k hiba az alábbi:

$$Err_k = O_k(1 - O_k)(T_k - O_k). \tag{8}$$

A rejtett rétegbeli j-edik egység *Errj* hibája a következő:

$$Err_j = O_j(1 - O_j) \sum_k Err_k w_{jk}\ , \tag{9}$$

ahol w_{jk} a j-edik rejtett rétegbeli egység és a k-adik kimeneti rétegbeli egység közti kapcsolat súlya, O_j pedig a j-edik egység aktuális kimenete.

Az így meghatározott hibák kiszámítása alapján a súlyok oly módon módosíthatóak, hogy a módosítás Δw_{ij} mértéke a w_{ij} súly esetén:

$$\Delta w_{ij} = l \times Err_j O_i, \tag{10}$$

ahol l az úgynevezett tanulási ráta. Ez leggyakrabban $\frac{1}{t}$-ként definiált, ahol a t paraméter az elvégzett iterációk, tehát az eddigi tanulási időszakok számát jelöli.

A súlyok módosítása mellett azonban a torzításokat is módosítani kell, mely módosítás $\Delta\theta_j$ mértéke a θ_j torzítás esetén:

$$\Delta\theta_j = l \times Err_j. \tag{11}$$

A szükséges módosítások után az egyes egységek értékei és a keletkezett hibák nagysága újból meghatározásra kerülnek, illetve az újabb szükséges korrekciók is. Ez a tanulási folyamat addig tart, míg a hibás osztályozások aránya elenyésző szintre nem csökken, a szükséges módosítások mértéke el nem ér egy előre meghatározott szintet, illetve az iterációk száma elér egy előre meghatározott nagyságot.

A tanulási folyamat befejezése után a döntési fához hasonlóan az osztályozásra használható szabályok feltárása következik, melyre szintén alkalmazható a „Ha-Akkor" szabályok felállításának módszere. Mivel ez sok változó esetén a nehéz interpretálhatóság következtében nem szerencsés, így helyette az eredmények megismertetésére célszerű lehet a súlyokat reprezentáló ábrák használata, osztályozásra pedig az algoritmusok számítógépes alkalmazása.

Logisztikus regressziós modellek

Az osztályozási feladatok megoldása során a gyakran alkalmazott technikák közé kell sorolni a regressziószámítást is, mely esetén a változók közti összefüggés számszerűsítése céljából sztochasztikus kapcsolatot feltételezünk egy vagy több – rendszerint X-szel jelölt – magyarázó változó és egy – Y-nal jelölt – függő változó között *(Hunyadi–Vita 2004)*.

A regressziós modellek egyik speciális típusa a logisztikus regresszió. Ez a modell olyan esetekben használható, amikor a függő változó bináris, azaz csak két értéket vehet fel. Ebben az esetben a magyarázó változók nem képesek a függő változó értékét meghatározni, csakis egy lehetséges kimenet bekövetkezésének esélyét. Ennek az esélynek a függő változóként való kezelése a függvényillesztés során – amikor is a lineáris modellekhez hasonlóan követendő cél, hogy a definiált görbe a minta n darab összetartozó X, Y adatpárja által meghatározott görbéjéhez a lehető legjobban illeszkedjen – két problémát vet fel. Az egyik, hogy az csak a 0;1 intervallumon vehet fel értéket, a másik pedig hogy az nem normális, hanem binomiális eloszlást követ. Előbbi probléma a változó transzformálása által oldható fel, utóbbi kezeléseként pedig paraméterbecsléskor a megfigyelt esemény valószínűségét maximalizáló maximum likelihood módszer alkalmazandó *(Rawlings et al. 1998)*.

Egy esemény bekövetkezési esélyének számszerűsítésére a valószínűség mellett gyakran használt mérőszám az esélyérték hányados[9], illetve a valószínűség egyéb – például a logit, a probit és a kiegészítő log-log – transzformációi által létrehozott mérőszámok[10]. Ezen transzformációk segítségével a valószínűség kifejezésére használt 0;1 intervallum oly módon képezhető le a $(-\infty;+\infty)$ intervallumra, hogy a $-\infty$ a lehetetlen, a $+\infty$ pedig a biztos esemény esélyét jelölje *(Reiczigel et al. 2007)*.

A transzformációk közül a leggyakrabban alkalmazott logit transzformáló formula a következő formában írható fel:

$$logit(Y) = \ln\big(odds(Y)\big), \tag{12}$$

[9] *Egy E esemény esélyérték-hányadosa:* $odds(Y) = \frac{Y}{1-Y}$, *ahol Y az E esemény bekövetkezésének P(E) valószínűsége.*

[10] *A kevésbé használatos probit transzformáló formula az általánosan* $\phi(x)$*-szel jelölt standard normális eloszlás eloszlásfüggvényének inverzét az alábbi módon használja:* $probit(Y) = \phi^{-1}(Y)$. *A kifejezetten ritkán alkalmazott kiegészítő log-log transzformáció alakja pedig az alábbi:* $cloglog(Y) = log\,(-log(1-Y))$.

ahol Y egy E esemény bekövetkezésének $P(E)$ valószínűségét jelöli, $odds(Y)$ pedig annak esélyérték-hányadosát *(Reiczigel et al. 2007)*.

A logit transzformáció segítségével felírt többváltozós logisztikus regressziós egyenlet az alábbi:

$$logit(Y) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m + \varepsilon, \tag{13}$$

ahol β_0 a konstans együtthatót, ε pedig a magyarázó változók értékétől független, additív, véletlen hibát jelenti. A β_j paraméter azt mutatja meg, hogy a j-edik magyarázó változó értékének egy egységgel történő megváltoztatása a többi magyarázó változó változatlanul hagyása mellett átlagosan mennyivel változtatja meg az adott esemény bekövetkezési esélyérték hányadosának logaritmusát *(Rawlings et al. 1998)*.

Rawlings et al. (1998) a logit transzformáció inverzére[11] is kifejezik az egyenletet, mellyel a bekövetkezés valószínűségének becslése az alábbi:

$$\hat{Y} = \frac{\exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m\right)}{1 + \exp\left(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m\right)}. \tag{14}$$

A paraméterek becslésére az említett „maximum likelihood"-módszer alkalmazandó. Az eljárás lényege, hogy minden megfigyelésre ki kell fejezni annak valószínűségét, hogy a függő változó éppen a megfigyelt értéket veszi fel. A likelihood függvény az egyedi sűrűségfüggvények szorzataként előálló együttes sűrűségfüggvény, mely annak a valószínűségét adja meg, hogy az egyenlet a függő változó megfigyelt értékeit veszi fel. Az optimalizálás során ennek logaritmusát, a log-likelihood függvényt kell maximalizálni *(Aldrich 1997)*.

A modell becsült paramétereinek szignifikánsságára a Wald-tesztet és a likelihood-hányados-tesztet kell alkalmazni. A Wald-teszt[12] nullhipotézise szerint egy vizsgált együttható sokasági értéke 0 – azaz a hozzá tartozó magyarázó változó nem szignifikáns a függő változó alakulásának tekintetében –, az alternatív hipotézis szerint pedig a paraméter értéke szignifikánsan különbözik nullától. A likelihood-hányados-teszt[13] nullhipotézise szerint a magyarázó változók paramétereinek mindegyike 0 értékű, így azok összességükben nem magyarázzák az eredményváltozót. Az eljárás során a likelihood-függvénynek a csak a konstans paraméterhez tartozó függvényértékét kell a paraméterek becsült értékeinek összességével számított függvényértékéhez hasonlítani *(Hosmer–Lemeshow 2000)*.

A modellillesztés jóságának értékelésére a Hosmer–Lemeshow-teszt alkalmazandó, mely próba nullhipotézise szerint a modell által előrejelzett negatív kimenetelű események

[11] *A logit transzformáció inverze egy esemény U valószínűségére az alábbi:* $invlogit(U) = \frac{exp(U)}{1+exp\,(U)}$.

[12] *A teszt próbafüggvénye az alábbi:* $(\frac{\hat{\beta}_j}{\sigma_{\hat{\beta}_j}})^2 \sim \chi^2(m-2)$.

[13] *A eljárás során felhasznált próbafüggvény a likelihood-hányados alkalmazásával az alábbi módon adható meg:* $-2ln\frac{L_0}{L_1} = -2(lnL_0 - lnL_1) \sim \chi^2(m-2)$, *ahol L0 illetve L1 a likelihood függvénynek a csak a konstans paraméterrel, illetve a paraméterek becsült értékének összességével számított értékét jelöli.*

bekövetkezésének száma nem különbözik szignifikánsan azok megfigyelt számától. Az eljárás végrehajtásához külön kell választani a pozitív és negatív kimenetelű egyedeket mind a megfigyelt, mind a modell által becsült esetben. Ezekből az adatsorokból egy kontingencia tábla készítendő oly módon, hogy elemeiket a vizsgált esemény bekövetkezésének valószínűsége alapján növekvő sorrendbe kell rendezni, majd ezeket a rangsorokat *k* darab egyenlő elemszámú csoportra kell bontani. Az így elkészített táblára meghatározhatóvá válik a Hosmer-Lemeshow-statisztika[14], mely által meghozható a modell jóságára vonatkozó döntés *(Hosmer–Lemeshow 2000).*

A paraméterek szignifikánsságának és a modell jóságának elemzése mellett a modell használhatóságát reprezentáló mutatók által a magyarázó változók hatásának mértékét is vizsgálni kell. Erre alkalmazható mutató a Cox-Snell-féle R^2együttható, illetve az ennek értékét a 0;1 intervallum határai közé szorító Nagelkerke-féle $\bar{R}^2$ együttható[15] *(Nagelkerke 1991).* Ezen mutatók a lineáris modellek esetén alkalmazott determinációs együtthatóhoz hasonló értelmezéssel bírnak. Ugyanakkor az előbbi vizsgálatok mellett a modell megkülönböztető erejére is figyelmet kell fordítani, melyet a tanulmány második részében definiálandó ROC-görbe vizsgálata által lehet megtenni *(Hosmer–Lemeshow 2000).*

Összefoglalás

Az adatbányászat cikkben ismertetett elméleti háttere és módszertani sajátosságai megfelelő alapot képeznek egy valós adatbányászati projekt folyamatának áttekintéséhez. A felvonultatott osztályozási módszerek számítógépes szoftver segítségével elvégzett empirikus alkalmazása, illetve a hatékonyságjavulást reprezentáló eredmények megerősíthetik azt a feltételezést, miszerint az adatbányászat gyakorlati alkalmazása képes releváns módon hatást gyakorolni a szervezetek eredményességére. Az elemzés folyamata és eredményei – mely az adatbányászat egy bemutatott alkalmazási területéhez, a direktmarketinghez kapcsolódik –, illetve az arra alapozott következtetések a tanulmány második részében kerülnek ismertetésre.

[14] *A próbafüggvény az alábbi alakban írható fel:* $G^2_{HL} = \sum_{j=1}^{k} \frac{(O_j - E_j)^2}{E_j(1-\frac{E_j}{n_j})} \sim \chi^2(k-2)$, *ahol nj a j-edik csoport elemeinek, Oj a j-edik csoport pozitív kimenetelű eseményeinek megfigyelt, Ej pedig a j-edik csoport pozitív kimenetelű eseményeinek becsült számát jelöli, és j=1,…,k a képzett csoportok száma.*

[15] *A Cox–Snell-féle* R^2*együttható az* $R^2 = 1 - (\frac{L_0}{L_1})^{2/n}$ *alakban írható fel, ahol n a mintaelemszám. Maximuma .* $R^2_{max} = 1 - (L_0)^{2/n}$. *Előzőek felhasználásával a Nagelkerke-féle* $\bar{R}^2$ *együttható a következő:* $\bar{R}^2 = \frac{R^2}{R^2_{max}}$.

Hivatkozások

Adriaans, P. – Zantinge, D. 2002: *Adatbányászat.* Panem, Budapest. 17. old.

Aldrich, J. 1997: *R. A. Fisher and the Making of Maximum Likelihood 1912 – 1922.* Statistical Science, vol. 12. no. 3. pp. 162-176.

Álmos Attila – Győri Sándor – Horváth Gábor – Várkonyiné Kóczy Annamária 2002: *Genetikus algoritmusok.* Typotex Kiadó, Budapest.

Berry, M. J. A. – Linoff, G. 1997: *Data Mining Techniques: For Marketing, Sales, and Customer Support.* John Wiley and Sons, Inc., New York.

Bishop, C. M. 1995: *Neural Networks for Pattern Recognition.* Oxford University Press, Oxford.

Bodon Ferenc 2009: *Adatbányászati algoritmusok.* In: http://www.cs.bme.hu/~bodon/magyar/adatbanyaszat/tanulmany/adatbanyaszat.pdf (Letöltve: 2009. február 09.) 18. old.

Elsberry, W. R. 1998: *Genetic Algorithms and Artifical Neural Networks.* Texas Chapter of the Association for Computing Machinery, Spring.

Fajszi Bulcsú – Cser László 2004: *Üzleti tudás az adatok mélyén.* Budapesti Műszaki és Gazdaságtudományi Egyetem, Gazdaság- és Társadalomtudományi Kar, Információ és Tudásmenedzsment Tanszék, Budapest. 167. old.

Han, J. – Kamber, M. 2004: *Adatbányászat. Koncepciók és technikák.* Panem, Budapest. 309., 311. old.

Hosmer, D. W. – Lemeshow, S. 2000: *Applied Logistic Regression* (Second Edition). John Wiley and Sons Inc., New York.

Hunyadi László – Vita László 2004: *Statisztika közgazdászoknak* (Harmadik átdolgozott kiadás). KSH, Budapest.

Inmon, W. H. 2005: *Building the Data Warehouse* (Fourth Edition). John Wiley and Sons, Inc., New York. 29. old.

Kotler, P. – Keller, K. L. 2006: *Marketingmenedzsment.* Akadémiai Kiadó, Budapest.

Kristóf Tamás 2002: *A mesterséges neurális hálók a jövőkutatás szolgálatában.* Budapesti Közgazdaságtudományi és Államigazgatási Egyetem, Jövőkutatási Kutatóközpont, Budapest.

Nagelkerke, N. J. D. 1991: *A Note on a General Definition of the Coefficient of Determination.* Biometrika, vol. 78. no. 3. pp. 691-692.

PSZÁF 2001: *A Pénzügyi Szervezetek Állami Felügyelete elnökének 8/2001. számú ajánlása a hitelkockázat kezeléséről.* In: http://www.pszaf.hu/bal_menu/szabalyozo_eszkozok/pszafhu_bt_ajanlirelvutmut/ajanlas_pszaf/pszafhu_ajanlirelvutmut_20050815_79.html?query=egyedi%20hitelb%C3%ADr%C3%A1lat (Letöltve: 2009. március 15.)

PSZÁF 2005: *A hitelintézetek és befektetési vállalkozások új tőkekövetelmény szabályaira (CRD) vonatkozó szakmai anyagok* (2. átdolgozott változat). In: http://www.pszaf.hu/data/cms1464363/bazel2_konzcrd_v2.pdf (Letöltve: 2009. március 15.)

Rawlings, J. O. – Pantula, S. G. – Dickey, D. A. 1998: *Applied Regression Analysis: A Research Tool* (Second Edition), Springer-Verlag New York, Inc., New York.

Reiczigel Jenő – Harnos Andrea – Solymosi Norbert (2007): *Biostatisztika nem statisztikusoknak.* Pars Kft, Nagykovácsi.

Szűcs Imre 2004: *CRM és kockázatelemzés kereszthatásainak vizsgálata adatbányászati módszerekkel.* In: http://odin.agr.unideb.hu/magisz/rendezveny/E-agrarium2004/konferencia/Publikacio/A43_doc.pdf (Letöltve: 2009. március 16.)

Tan, P. N. – Steinbach, M. – Kumar, V. 2005: *Introduction to Data Mining.* Addison-Wesley, Richmond, TX.

Thearling, K. 2009: *An Introduction to Data Mining.* In: http://www.thearling.com/index.htm#wps (Letöltve: 2009. február 02.)

Ullman, J. D. – Widom, J. 2008: *Adatbázisrendszerek* (Második, átdolgozott kiadás). Panem, Budapest.

Valyon József 2007: *Kiterjesztett LS-SVM és alkalmazása rendszermodellezési feladatokban* (PhD. értekezés tézisei). Budapesti Műszaki és Gazdaságtudományi Egyetem, Méréstechnika és Információs Rendszerek Tanszék, Budapest.

Viczián István 2002: *IBM Websphere MQ.* In: http://delfin.unideb.hu/~vicziani/pdf/ws_mq_middleware.pdf (Letöltve: 2009. február 23.)