Statisztikai Szemle

Közzététel: 2018. december 4.

A tanulmány címe:

A case-based reasoning alkalmazása a hazai mikrovállalkozások csődelőrejelzésére

Szerzők:

Kristóf Tamás, a Budapesti Corvinus Egyetem egyetemi docense, az MTA Statisztikai és Jövőkutatási Tudományos Bizottság tagja, e-mail: tamas.kristof@uni-corvinus.hu

DOI: https://doi.org/10.20311/stat2018.11-12.hu1109

A case-based reasoning alkalmazása a hazai mikrovállalkozások csődelőrejelzésére

Kristóf Tamás,
a Budapesti Corvinus Egyetem egyetemi docense, az MTA Statisztikai és Jövőkutatási Tudományos Bizottság tagja
E-mail: tamas.kristof@uni-corvinus.hu

A tudomány fejlődése számos adatvezérelt és mesterségesintelligencia-eljárással tette lehetővé a csődelőrejelző modellek teljesítményének javítását. Ezek egyike az utóbbi időben növekvő népszerűséget elérő CBR (case-based reasoning – esetalapú következtetés). A tanulmány célja a CBR alkalmazási lehetőségeinek vizsgálata magyarországi mikrovállalkozások mintáján klasszikus csődelőrejelzés keretében, a klasszifikációs erő szempontjából összevetve a CBR-t a szakirodalomban és a gyakorlatban leggyakrabban alkalmazott három csődelőrejelzési eljárással (a döntési fával, a logisztikus regresszióval és a neurális hálóval). Az empirikus vizsgálat 1 828 hazai mikrovállalkozás 2017-ben megfigyelt csődeseményének bekövetkeztével (csődeljárás, felszámolási eljárás vagy kényszertörlés megindításával) foglalkozott 2015. és 2016. évi beszámolóadatokból számított pénzügyi mutatók alapján. A szerző eredményei szerint érdemes figyelmet szentelni a CBR-nek a csődelőrejelzéshez hasonló klasszifikációs problémák megoldása során. Az empirikus vizsgálat tesztelő mintájában szereplő megfigyelések esetén azonban a CBR-modell előrejelző ereje elmaradt a neurális háló és a logisztikus regresszió teljesítményétől.

TÁRGYSZÓ:
Case-based reasoning.
Csődelőrejelzés.
Mikrovállalkozások.

DOI: 10.20311/stat2018.11-12.hu1109

A KSH (Központi Statisztikai Hivatal) nyilvántartása szerint a 2016 végén, Magyarországon működő 693 662 cég közül 682 475 a mikro-, kis- és középvállalkozások sorába tartozott; utóbbiakból 649 773 volt kevesebb mint 10 fővel működő mikrovállalkozás, amelyek 1 070 153 főt foglalkoztattak (*KSH* [2017]). A mikrovállalkozások nemzetgazdasági fontosságuk mellett számosságuk miatt is alkalmasak arra, hogy különböző empirikus vizsgálatok – így a csődelőrejelzés – tárgyát képezzék. A csődelőrejelzés a statisztikában az esetek többségében bináris klasszifikációs probléma, amelyben a csődesemény a vállalkozással szemben megindított csődeljárással, felszámolási eljárással vagy kényszertörléssel specifikálható. A mikrovállalkozások csődelőrejelzése az alkalmazott módszertan tekintetében nem tér el a közép- és a nagyvállalatok csődelőrejelzésétől, adatgyűjtéskor ugyanakkor lényegesen könnyebb helyzettel szembesül a kutató a csődbe jutott mikrovállalkozásokról rendelkezésre álló, megfelelő mennyiségű adat következtében.[1] A különböző méretű vállalkozások csődmodellépítése ezzel egyidejűleg a modellváltozókat illetően jelentősen eltérhet egymástól, mivel a kisebb vállalkozások rövid távú túlélése szempontjából az árbevétel realizálása, illetve növekedése általában kritikusan fontos ismérv, szemben a nagyobb vállalkozásoknál jellemző cash flow, eladósodottsági, tőkeszerkezeti és likviditási mutatókkal.

A vállalati pénzügyek területén napjaink változatlanul népszerű kutatási témája a hatékony csődelőrejelzési modellek kifejlesztése, hiszen ezek a megbízható csődvalószínűség-becslésen keresztül alapvető elemei a hitelintézetek kockázatkezelési tevékenységének. A többváltozós statisztikai klasszifikációs módszertan alkalmazása a csődelőrejelzés területén *Altman* [1968] diszkriminanciaanalízissel készített, úttörő és méltán világhírű csődmodelljével kezdődött. Magyarországon az első diszkriminanciaanalízis-alapú csődmodell az 1990-es évek elején készült (*Virág–Hajdu* [1996]). Az első csődmodellek megjelenését követő évtizedekben a tudomány fejlődése számos adatvezérelt statisztikai és mesterségesintelligencia-módszertannal tette lehetővé a csődelőrejelzési modellek előrejelző képességének javítását.[2]

[1] A KSH adatai szerint felszámolási eljárás 2017-ben 74 középvállalkozással és 13 nagyvállalattal szemben, csődeljárás pedig, minden vállalkozási kategóriát egybevéve, összesen 39 esetben indult Magyarországon (*KSH* [2018]). Gyakorlati hüvelykujjszabály a többváltozós statisztikai bináris klasszifikációs módszerek alkalmazásakor, hogy legalább 50 megfigyelésnek mindkét osztályban lennie kell a modellfejlesztés alapjául szolgáló tanulási mintában, különben az eredmények összehasonlíthatósága korlátozott (*Kristóf* [2008]). A kismintás csődelőrejelzési probléma logisztikus regresszióval történő kezeléséről részletes áttekintést ad *Hajdu* [2004] publikációja.

[2] Erről Magyarországon is több publikáció született (lásd például *Hajdu–Virág* [2001], *Hajdu* [2003], *Kristóf–Virág* [2012], *Virág–Nyitrai* [2014], illetve a *Statisztikai Szemle*ben *Kristóf* [2005], *Nyitrai* [2014], *Nyitrai–Virág* [2017]). Jelen tanulmánynak terjedelmi okokból nem célja a csődelőrejelzés fejlődéstörténetének és módszertanának a Bayes-klasszifikációt, a diszkrimanciaanalízist, a logisztikus regressziót, a döntési fákat, a neurális hálókat, a gépi tanulási eljárásokat és a metamódszereket egyaránt felölelő részletes bemutatása.

A tanulmány célja a csődelőrejelzés „főáramlatában" és gyakorlatában viszonylag kevéssé elterjedt, ugyanakkor a komplex, változó üzleti környezetben jól alkalmazható, ígéretes problémamegoldó és döntés-előkészítési eljárás, a CBR alkalmazási lehetőségeinek megvizsgálása hazai mikrovállalkozások csődelőrejelzésében, komparatív értékelést adva a CBR és a leggyakrabban alkalmazott módszertanok teljesítményében tapasztalt különbségekről.

A CBR[3] a mesterséges intelligencia módszertani családba tartozó eljárás, amely múltbeli tapasztalatok alapján igyekszik új problémákra megfelelő megoldásokat találni. Az utóbbi években a CBR növekvő népszerűségre tett szert a gyakorlatban, amelynek oka, hogy alkalmazásakor nem merül fel túltanulási probléma[4], megfelelő magyarázó erőt képes felmutatni a célváltozó előrejelzésekor, valamint karbantartása lényegesen egyszerűbb, mint a mesterséges intelligencián alapuló rendszerek többségéé. A módszer szélesebb körű publikációjának hiánya arra vezethető vissza, hogy a kezdeti empirikus vizsgálatok a CBR alacsonyabb szintű előrejelző képességét igazolták az iparági legjobb gyakorlatként alkalmazott neurális hálókhoz és logisztikus regresszióhoz viszonyítva. Napjainkra számos technika áll azonban már rendelkezésre, amellyel a CBR előrejelző képessége javítható. Jelen tanulmány is erre igyekszik empirikus vizsgálattal alátámasztott módszertani megoldást nyújtani.

A tanulmány először áttekintést nyújt a CBR fejlődéstörténetéről, módszertanáról, alkalmazási feltételeiről, a csődelőrejelzés területén rendelkezésre álló nemzetközi empirikus vizsgálatok tapasztalatairól, valamint gyakorlati kihívásairól. Ezt követően összehasonlító empirikus vizsgálat keretében azt vizsgálja, hogy a klasszifikációs erő szempontjából mennyire lehet létjogosultsága a CBR-nek a magyarországi mikrovállalkozások csődelőrejelzésében, összevetve a CBR-t a szakirodalomban és a gyakorlatban leggyakrabban alkalmazott három csődelőrejelzési eljárással (a döntési fával, a logisztikus regresszióval és a neurális hálóval).

1. A CBR fejlődéstörténete

A CBR elméleti és módszertani alapjait *Schank* [1982] fogalmazta meg. A szerző az ún. dinamikus memória elméletében első ízben írta le a gondolkodás memóriaalapú megközelítését, és fogalmazott meg architektúrát a gondolkodási rendszer számítógépes megvalósítására. A CBR alapelve, hogy az emberi szakértelem analógiai és

[3] Az eljárás MBR (memory-based reasoning – memórialapú következtetés) néven is ismert.

[4] A túltanulási probléma az a jelenség, amikor a tanulási mintán épített modell túlságosan az ismert adatbázis sajátosságaira specializálódik, amelynek révén az ismert adatokon nagyon jó klasszifikációs teljesítményre képes, ugyanakkor új adatokon csupán romló teljesítményt nyújt, így korlátozottan alkalmazható.

kísérleti alapokon old meg komplex problémákat, és képes tanulni korábbi problémamegoldási tapasztalatokból. A memóriában történő visszakeresésre azonban az is igaz, hogy az ember egyrészt hajlamosabb élénkebben emlékezni az első élményeire, másrészt a frissebb élmények is élénkebb hatást válthatnak ki benne (*Brown–Gupta* [1994]).

A CBR szisztematikus keresést tesz lehetővé a memóriában/esettárban, amellyel az aktuális vizsgálat tárgyát képező esethez leginkább hasonlóakat igyekszik kinyerni (*Kolodner* [1993]). A gépi tanulással ezáltal számottevően nagyobb adathalmazból képes a leginkább releváns eseteket megtalálni, mint az ember. A gyakorlati alkalmazás során a CBR rendszerkarbantartása lényegesen egyszerűbb és hatékonyabb lehet, mint a statisztikai modelleké, hiszen csupán új eseteket kell hozzáadni a rendszerhez, és nem szükséges új modelleket építeni (*Bryant* [1997]).

A csődelőrejelzés szempontjából releváns első CBR-publikációt ismereteink szerint *Buta* [1994] készítette el. A szerző 1 039 egyesült államokbeli vállalat 1991–1992. évi pénzügyi mutatóinak felhasználásával épített CBR-modellt azzal a céllal, hogy a vállalati kötvények jövőbeni besorolását előre jelezze. A minta 10 százalékát tesztelési céllal elkülönítette, amelyen a modellt alkalmazva 90,4 százalékos besorolási pontosságot ért el.

Bryant [1997] 1975 és 1994 között vizsgálta 2 ezer normál és 85 csődbe jutott egyesült államokbeli feldolgozóipari vállalat pénzügyi mutatóit, és épített több időszak dinamikus mutatóit figyelembe vevő CBR-modelleket, amelyek teljesítményét benchmark modellként *Ohlson* [1980] világhírű logisztikus regressziós modelljével hasonlította össze. A 10 százalékos tesztelő mintán valamennyi CBR-modell gyengébb teljesítményt ért el, mint a logisztikus regresszió, annak ellenére, hogy másodfajú hiba tekintetében nem voltak rosszabbak a CBR-modellek.

Jo–Han–Lee [1997] dél-koreai vállalatok csődelőrejelzési adatbázisára építettek diszkriminanciaanalízis-, neurálisháló- és CBR-alapú modelleket. A CBR-modell teljesítménye 81,8 százalékos besorolási pontosságú volt, ami elmaradt a diszkriminanciaanalízis 82,4, illetve a neurális háló 86,4 százalékától.

Zurada–Lonial [2005] az Egyesült Államok egészségügyi szektorában modellezték nemteljesítő követelések megtérülési valószínűségét neurális hálókkal, döntési fákkal, logisztikus regresszióval, CBR-rel és metamódszerekkel. Az egyes modellek teljesítményét a klasszifikációs modelleknél gyakran alkalmazott ROC- (receiver operating characteristic – kumulált besorolási pontosság) és lift-görbével hasonlították össze. Összességében legjobb eredményt a logisztikus regresszióval, legrosszabbat a CBR-rel értek el, a legalacsonyabb elsőfajú hibát ugyanakkor a neurálisháló-modell követte el.

A CBR-modell teljesítményét vizsgáló összehasonlító empirikus vizsgálatok területén *Ahn–Kim* [2009] tanulmánya hozta meg az áttörést. A szerzők 1 335 normál és 1 335 csődbe jutott dél-koreai nehézipari vállalat mintáján genetikus algoritmusokkal

optimalizálták a CBR példakiválasztása során alkalmazott változósúlyozást, jelentősen javítva a hagyományos CBR-modell előrejelző képességét. Benchmark modellként 8, 16, 24 és 32 neuronból felépült, köztes rétegű neurális hálókat alkalmaztak. A tesztelő mintán a CBR 86,7, míg a legjobb neurális háló 85,4 százalékos besorolási pontosságot ért el, amivel első ízben sikerült igazolni a CBR versenyképességét a megbízható csődelőrejelzésben.

A csődbe jutott vállalkozások valós populációban megfigyelhető relatíve alacsonyabb arányát reprezentálta *Li et al.* [2014] kínai vállalkozásokból álló csődelőrejelzési mintája, melyen a szerzők a CBR-t alkalmazták hierarchikus klaszterelemzéssel kombinálva, hogy javítsák az esetkinyerés hatékonyságát. A modell teljesítményét összevetették a hagyományos CBR-rel, a logisztikus regresszióval, a támogató vektorok módszerével (support vector machine-nel) és a diszkriminancia-analízissel. A klaszterelemzéssel kombinált CBR valamennyi eljárás teljesítményét felülmúlta a besorolási pontosság és a mintában kis arányt képviselő, csődbe jutott esetek felismerése tekintetében.

A nemzetközi empirikus vizsgálatok eredményei arra engednek következtetni, hogy a CBR napjainkra túljutott a „gyermekbetegségeken", és megfelelő technikákkal kombinálva, ígéretes alternatívát jelenthet a csődelőrejelzésben az iparági legjobb gyakorlatként alkalmazott módszerek mellett.

2. A CBR módszertani leírása és alkalmazásának feltételei

A CBR folyamatát az 1. ábra mutatja be. A CBR úgy old meg új problémákat (csődelőrejelzés esetén klasszifikál új megfigyeléseket), hogy először az esettárból kinyer egy vagy több korábban tapasztalt esetet, és megfelelő módon felhasználja az azokban levő információkat, majd felülvizsgálja a megoldást a korábbi esetek alapján, végül pedig megőrzi az új tapasztalatot az esettárba való belefoglalással.

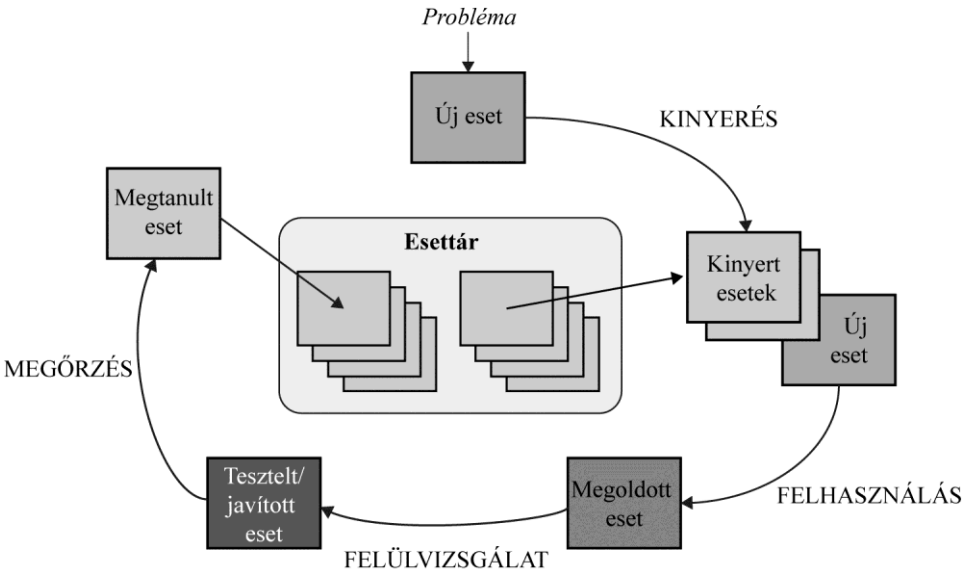
A CBR a nemparaméteres *k* legközelebbi szomszéd eljárással[5] rokon módszer, amely hasonlósági függvény[6] alkalmazásával generál besorolásokat a tárolt esetek tulajdonságainak felhasználásával (*Park–Han* [2002]). Alkalmazása nem feltételez előzetes függvénytípust, valamint nem igényli a vizsgált sokaságra vonatkozó paraméterek (átlag, szórás) becslését; valószínűségelméleti háttere miatt ugyanakkor

[5] A *k* legközelebbi szomszéd eljárás csődelőrejelzésben való alkalmazásáról *Nyitrai* [2015] tanulmánya ad részletes áttekintést hazai vállalkozások mintáján.

[6] Számos korábbi tanulmány arra világított rá, hogy a *k* legközelebbi szomszéd eljárás teljesítménye érzékeny lehet a hasonlósági függvényre, valamint az irreleváns változókra (*Yip* [2006]), amelyek megoldása többféle távolságfüggvény súlyozása, illetve a változók relevanciájának megfelelő figyelembevétele lehet.

alkalmazásának feltétele legalább 50 megfigyelés megléte. A módszer érzékeny az adatok lokális struktúrájára, ami felvetheti az anekdotikus evidencia problémáját (*Goodwin* [2009]). Az eljárás az input megfigyelések és a tárolt esetek közötti célváltozó értéke szempontjából is releváns hasonlóságokat távolságmértékkel számítja ki.

1. ábra. A CBR folyamata



Forrás: *Aamodt–Plaza* [1994] 47. old.

Valamely új eset és a tárolt esetek közötti hasonlóságot számos módon lehetséges meghatározni. Amikor az eseteket változóvektorok reprezentálják, gyakran alkalmazott megközelítés a változóértékek közötti távolságok súlyozott összege (*Jarmulak–Craw–Rowe* [2000]), amelynek tipikus numerikus függvénye:

$$\frac{\sum_{i=1}^{n} W_i \times sim\left(f_i^1, f_i^R\right)}{\sum_{i=1}^{n} W_i}, \qquad /1/$$

ahol W_i az *i*-edik változó súlya, f_i^1 az input eset *i*-edik változójának értéke, f_i^R a kinyert eset *i*-edik változójának értéke, és $sim\left(f_i^1, f_i^R\right)$ az f_i^1 és f_i^R közötti hasonló-

sági függvény, ami gyakran az euklideszi távolság. A CBR-modellezés lényege a megfelelő f_i változók, W_i változósúlyok és R kinyert példaesetek megtalálása, amire a szakirodalomban és a gyakorlatban több optimalizáló módszer is rendelkezésre áll.

A CBR és a k legközelebbi szomszéd módszer között fontos különbség, hogy míg az utóbbi a tanulási mintában szereplő eseteket az euklideszi térben levő pontokként tárolja, addig a CBR komplex szimbolikus leírásokként (*Zurada–Lonial* [2005]). Új megfigyelések klasszifikálásakor a CBR először megvizsgálja, hogy létezik-e ugyanolyan tárolt eset. Ha van, akkor annak a megoldását adja eredményül. Ha nincs, akkor a tanulási mintában szereplő megfigyelések komponensei között igyekszik hasonlóságot keresni. Az eljárás ezeket a hasonló eseteket tekinti legközelebbi szomszédoknak.

A CBR feltételezi, hogy a változók numerikusak, egymásra ortogonálisak és standardizáltak (*Matignon* [2007]). A folytonos pénzügyi mutatók teljesítik a numerikus változóra vonatkozó alkalmazási feltételt; a második két követelmény teljesítésére pedig megfelelő technikát kínál a főkomponens-elemzés.[7] Ezáltal a legközelebbi szomszédok megtalálására szolgáló nemparaméteres modell input változói az eredeti változók helyett maguk a főkomponensek. A változók standardizálása azért fontos, mert a CBR algoritmusa valamennyi euklideszi távolságpárt vizsgál, amelyeket a változók értékkészlete befolyásol.

A CBR alkalmazásakor kulcsfontosságú kérdés a mintában szereplő megfigyelések tárolásának, valamint a legközelebbi szomszédok meghatározásának módja. Erre igazoltan hatékony megoldásnak bizonyul az RDT- (reduced dimensionality tree – dimenziócsökkentő fa) módszer (*Liu–Zhang* [2012]). Az RDT bináris döntési fákat épít, az adathalmazt a megfigyelések közötti legnagyobb varianciát biztosító dimenziók alapján folyamatosan részmintákra választva szét, és közben egyre kevesebb megfigyelést hagyva az egyes részmintákban. A szétválasztás általában az egyes csomópontok mediánértékei szerint történik. A döntési fa túltanulásának elkerülésére számos leállítási kritérium specifikálható.

A CBR-modellezés másik fontos kihívása – a k legközelebbi szomszéd eljáráshoz hasonlóan – magának a k paraméternek a meghatározása, ami gyakran, az adatok eloszlását és a változók számát is figyelembe véve, próbálgatással történik.[8] A k legközelebbi szomszéd becslések meghatározott régión belül a leggyakoribb célváltozó-kategóriába való tartozás és a régiót körülvevő adatpontok átlagai alapján egyaránt végezhetők. Minél nagyobb a tanulási minta, annál jobbak lehetnek a legközelebbi szomszéd becslések. Ezzel egyidejűleg k értéke az RDT-ben a bináris elválasz-

[7] A CBR fejlődéstörténetének áttekintésekor idézett publikációk egyike sem alkalmazta a főkomponens-elemzést a modellváltozók egymásra ortogonálissá tételére, ami – ugyan empirikusan nem bizonyítható, de – oka lehetett a tapasztalt alacsonyabb teljesítménynek.

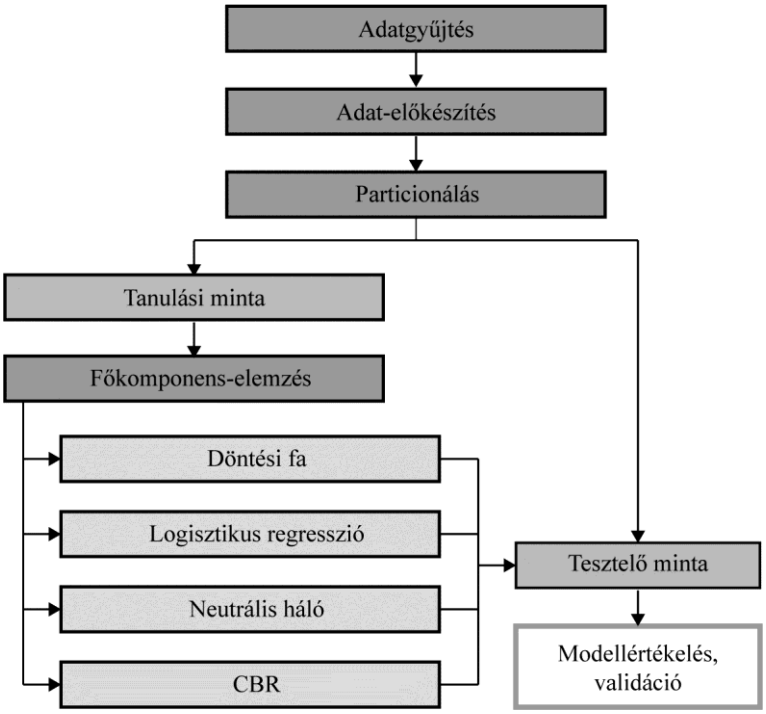
[8] Emiatt célszerű a modellezés során folyamatosan keresztvalidációt végezni, hiszen a modellteljesítmény akár jelentősen függhet a k értékétől.

tás leállítását is meghatározza, hiszen nem lehet a döntési szabályok végeredményeként kevesebb megfigyelés adott ágon, mint a legközelebbi szomszédok száma.

3. CBR-csődmodellezés hazai mikrovállalkozások mintáján

A CBR módszertanának és a releváns nemzetközi empirikus vizsgálatok eredményeinek megismerését követően jogosan merül fel kutatási kérdésként, hogy mennyire lehet létjogosultsága a CBR-nek a magyarországi vállalkozások csődelőrejelzésében. Az összehasonlító empirikus vizsgálat a hagyományos bináris klasszifikációs probléma megoldása keretében azt vizsgálja, hogy a CBR-modell teljesítménye a modellfejlesztés során nem használt megfigyeléseket tartalmazó tesztelő mintán jobb vagy rosszabb-e, mint a leggyakrabban alkalmazott három eljárásé (a döntési fáé, a logisztikus regresszióé vagy a neurális hálóé).

2. ábra. Az empirikus vizsgálat folyamata



Megjegyzés. Itt és a további ábráknál, táblázatoknál CBR (case-based reasoning): esetalapú következtetés.

Az adatgyűjtés 1 828 magyarországi társas mikrovállalkozásra terjedt ki. Ez a mintanagyság ugyan általánosított következtetések levonását nem teszi lehetővé, de a fejlődéstörténetnél taglalt nemzetközi empirikus vizsgálatok során felhasznált elemszámok tükrében, megfelelő nagyságot képvisel a módszer alkalmazási lehetőségeinek vizsgálatára. A mikrovállalkozások köre a hatályos európai uniós és hazai jogszabályok mikro-, kis- és középvállalkozásokra megalkotott definíciója alapján 2 millió euró árbevétel és mérlegfőösszeg, valamint 10 fő foglalkoztatotti létszám alatti vállalkozásoknak feleltethető meg. A megkérdőjelezhető adatminőség és a hiányzó értékkel rendelkező beszámolók nagy száma következtében az adatgyűjtésből kizártuk a 10 millió forint árbevétel alatti vállalkozások. Tekintettel arra, hogy jelen empirikus vizsgálatnak nem célja a banki portfolióra, nemzetgazdasági ágazatra vagy más speciális, rétegzett mintavételt igénylő célra történő modellfejlesztés, a modellezési adatbázist egyszerű véletlen mintavétellel állítottuk össze. A minta legfontosabb jellemzőit, nemzetgazdasági ágak szerinti megoszlását, csődkarakterisztikáit és pénzügyi tulajdonságait az 1. táblázat foglalja össze.

1. táblázat

A minta összetétele és legfontosabb pénzügyi jellemzői

Nemzetgazdasági ág	Összes vállalkozás száma (db)	Működő vállalkozások száma (db)	Csődbe jutott vállalkozások száma (db)	Átlagos árbevétel (ezer Ft)	Átlagos mérlegfőösszeg (ezer Ft)	Átlagos bruttó cash flow (ezer Ft)
Adminisztratív és szolgáltatást támogató tevékenység	100	59	41	94 894	109 149	189
Egyéb szolgáltatás	23	16	7	75 969	103 587	1 683
Építőipar	201	92	109	70 729	65 910	–8 178
Feldolgozóipar	220	101	119	82 272	112 839	–6 035
Humánegészségügyi, szociális ellátás	69	37	32	28 698	62 961	4 432
Információ, kommunikáció	98	47	51	53 689	67 117	8 645
Ingatlanügyletek	110	50	60	82 601	128 370	48 934
Kereskedelem, gépjárműjavítás	532	269	263	83 722	88 050	–2 968
Közigazgatás, védelem, kötelező társadalombiztosítás	2	1	1	19 539	42 164	–5 381
Művészet, szórakoztatás, szabadidő	20	11	9	58 869	54 877	–10 554
Oktatás	14	6	8	40 810	37 785	5 100
Pénzügyi, biztosítási tevékenység	31	11	20	43 792	64 208	–4 368
Szakmai, tudományos, műszaki tevékenység	179	90	89	51 607	101 533	2 438

(A táblázat folytatása a következő oldalon.)

(Folytatás.)

Nemzetgazdasági ág	Összes vállalkozás száma (db)	Működő vállalkozások száma (db)	Csődbe jutott vállalkozások száma (db)	Átlagos árbevétel (ezer Ft)	Átlagos mérlegfőösszeg (ezer Ft)	Átlagos bruttó cash flow (ezer Ft)
Szálláshely-szolgáltatás, vendéglátás	111	63	48	52 050	38 002	–209
Szállítás, raktározás	105	54	51	98 710	422 609	2 981
Villamosenergia-, gáz-, gőzellátás, légkondicionálás	3	–	3	34 712	17 707	–21 061
Vízellátás, szennyvíz gyűjtése, kezelése, hulladékgazdálkodás, szennyeződésmentesítés	10	7	3	52 700	109 823	8 482
Összesen	*1 828*	*914*	*914*	73 073	106 665	1 370

A klasszikus csődelőrejelzési definíciót alkalmazva, a modellezési célváltozó meghatározása egyéves előretekintéssel: a 2016. évi beszámoló fordulópontját követően, adott vállalkozással szemben 2017. január 1-je és 2017. december 31-e között csődeljárás, felszámolási eljárás vagy kényszertörlés indult-e. Jelölése bináris (1/0) értékkel történt.

Annak érdekében, hogy az alkalmazott klasszifikációs módszerek minél több jól teljesítő és fizetésképtelen vállalkozás tulajdonságaiból tudjanak összefüggéseket felállítani, a célváltozó szerinti besorolás szempontjából az adatbázisban szereplő megfigyelések fele (914 vállalkozás) 2017. december 31-én működő vállalkozás volt, míg a másik fele (914) olyan vállalkozás, amely 2017 folyamán csődeljárás, felszámolási eljárás vagy kényszertörlés alá került.

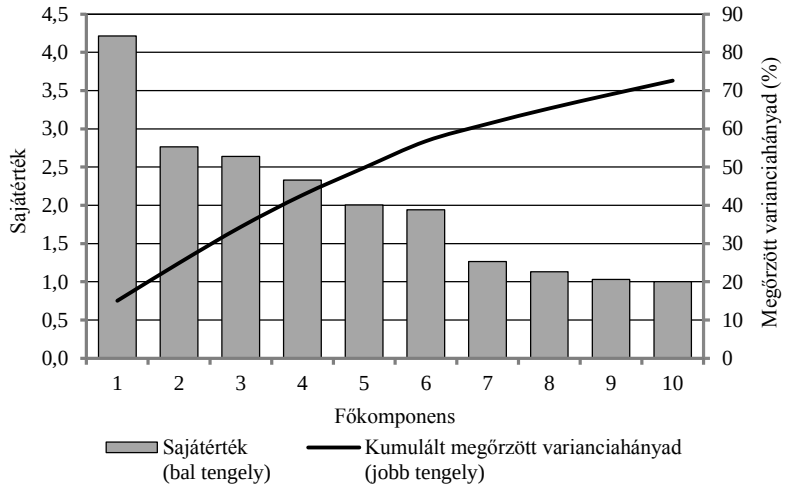
A magyarázó változók 2016. évi beszámolóadatokból számított pénzügyi mutatók voltak. Tekintettel arra, hogy számos pénzügyi mutató időszakra vonatkozó eredménytételt és időpontra vonatkozó mérlegtételt viszonyít egymáshoz, a mérlegtételek előző és tárgyévre vonatkozó értékeit átlagolni kellett; ezért az adatgyűjtés során követelmény volt, hogy a 2015. évi éves beszámolók adatai is rendelkezésre álljanak. Mindkét év figyelembevételére a növekedési mutatók miatt is szükség volt.

A magyarázó változók a pénzügyi elemzés, illetve a csődelőrejelzés szakirodalmában és gyakorlatában alkalmazott pénzügyi mutatók közül kerültek ki (*Virág–Fiáth* [2010]). Az éves beszámolóadatokból a következő 28 pénzügyi mutatót képeztük: *1*. árbevétel arányos EBITDA (earnings before interest, taxes, depreciation, and amortization – kamatok, adózás és értékcsökkenési leírás előtti eredmény), *2*. Árbevétel arányos nyereség, *3*. Árbevétel növekedési üteme, *4*. Befektetett eszközök aránya, *5*. Befektetett eszközök saját finanszírozása, *6*. Cash flow/árbevétel, *7*. Cash flow/összes tartozás, *8*. Dinamikus jövedelmezőségi ráta, *9*. Dinamikus likvi-

ditási ráta (EBITDA-alapon), *10*. Dinamikus likviditási ráta (üzemieredmény-alapon), *11*. EBITDA-jövedelmezőség, *12*. Eszközarányos árbevétel, *13*. Eszközarányos nyereség, *14*. Forgóeszközök aránya, *15*. Készletek forgási sebessége, *16*. Készpénz-likviditási ráta, *17*. Likvid pénzeszközök aránya, *18*. Likviditási gyorsráta, *19*. Likviditási ráta, *20*. Mérlegfőösszeg növekedési üteme, *21*. Nettó forgótőkearány, *22*. Rövid lejáratú kötelezettségek forgása, *23*. Sajáttőkeerő-dinamika, *24*. Saját vagyon aránya, *25*. Sajáttőke-arányos nyereség, *26*. Tőkeellátottsági arány, *27*. Vevők forgási sebessége, *28*. Vevők/szállítók.

Az adat-előkészítés a hiányzó értékek, a nullával való osztások, a kettős negatív osztások és az outlier (kiugró) értékek kezelését foglalta magába. A hiányzó értékek pótlását és a nullával való osztásokat az egyes pénzügyi mutatók értelmétől függően medián, csonkolt minimum vagy csonkolt maximum imputációval oldottuk meg. A csonkolt minimum és a csonkolt maximum értékeket az egyes változók mutatószámítási anomáliáiban nem érintett értékek 1. és 99. percentilisében állapítottuk meg. A kettős negatív osztás problémája a sajáttőkeerő-dinamika és a sajáttőke-arányos nyereség mutatóknál merült fel, amelynek megoldása a csonkolt minimum alkalmazása volt.[9] Az outlier értékek csonkolása úgyszintén az egyes mutatószámértékek 1. és a 99. percentiliseinek alapján történt.

3. ábra. A főkomponensek sajátértéke és megőrzött varianciahányada



A kidolgozott csődelőrejelzési modellek megfelelő validációjának biztosítása érdekében a modellezési adatbázist 75 : 25 százalék arányban particionáltuk tanu-

[9] Veszteséges és negatív saját tőkéjű vállalkozások sajáttőke-arányos nyeresége alapesetben pozitív számot ad eredményül, akárcsak a két időszakban negatív saját tőkéjű cégek sajáttőkeerő-dinamikája, ami különösen a csődbe jutott vállalkozások pénzügyi mutatóinak csődelőrejelzési célú alkalmazását torzíthatja súlyosan.

lási (1 371 megfigyelés) és tesztelő mintákra (457 megfigyelés). A további modellezési lépéseket a tanulási mintán, míg a visszaméréseket a tesztelő mintán hajtottuk végre.

A CBR alkalmazási feltételei közül a modellváltozók egymásra való ortogonalitásának és standardizáltságának biztosítása érdekében a 28 változóból 10 főkomponenst képeztünk. Az eljárás során követelményként támasztottuk a főkomponensek szignifikanciáját a Barlett-féle χ^2-próba alapján; számuk meghatározásakor pedig *Kaiser* [1960] gyakorlatban rendkívül elterjedt szabályát vettük alapul, aki úttörő publikációjában az 1 feletti sajátértékű főkomponenseket tekintette megbízható faktoroknak. Az 1 sajátértékküszöb alkalmazását az is indokolja, hogy 1 az az átlagos variancia, amikor *p* számú, 1 varianciájú indikátort elemzünk. Figyelemmel voltunk továbbá a sajátértékek csökkenésének ütemét jelző hüvelykujjszabályra is, amelyet a 3. ábrán követhetünk nyomon. A 10 főkomponens kumulált megőrzött varianciahányada 72,6 százalék volt.

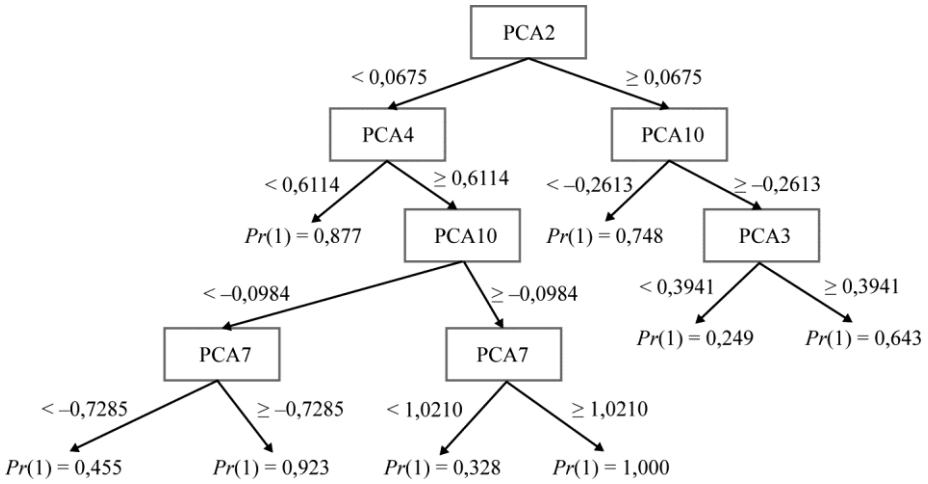
4. A négy csődmodell értékelése

A négy módszerrel kidolgozott csődmodellek mindegyike a 10 főkomponensből és a modellezési célváltozóból épült fel a tanulási mintán. Jelen tanulmánynak terjedelmi okokból nem célja a döntési fa, a logisztikus regresszió és a neurális háló módszerének részletes ismertetése, mivel azokról korábban Magyarországon is már számos publikáció megjelent.[10]

A döntési fára épülő modell kidolgozása a CHAID- (chi-squared automatic interaction detection – khínégyzetalapú automatikus interakció-detektálás) algoritmussal történt. A döntési fa mélysége a tanulási mintából kiindulva négy szintet ért el. Legerősebb particionáló változónak a 2. főkomponens bizonyult. A fa második szintjén a 4. és a 10. főkomponens, a harmadik szinten a 10. és a 3. főkomponens, míg a negyedik szinten, két ágon is, a 7. főkomponens biztosította a leginkább homogén szétválasztást. A döntési fa túltanulásának elkerülése érdekében az elágazásképzést kontrolláló leállítási kritériumok a tökéletesen homogén osztályképzés, a szülő- és gyermekágakra paraméterezett minimális elemszám, valamint a homogenitásimutató-változás minimális mértéke voltak. A döntési fa felépítésének alapjául szolgáló főkomponensintervallum-határokat és a döntési szabályok eredményeképpen adódó csődvalószínűségeket a 4. ábra tartalmazza.

[10] Lásd például *Kristóf* [2008], *Kristóf–Virág* [2012].

4. ábra. A döntési fa felépítése és a szabályokból eredő csődvalószínűségek



Megjegyzés. PCA (principal component analysis): főkomponens-analízis; *Pr*: főkomponensintervallumhatár. A nyilak mellett az elválasztást jelentő intervallumhatárok szerepelnek.

A logisztikus regresszió modelljét a változók erősségének hatássorrendjében történő beléptetését megvalósító forward stepwise eljárással építettük fel, 5 százalékos beléptetési kritérium alkalmazásával. Első lépésben a 6., második lépésben a 4., harmadik lépésben a 2., negyedik lépésben a 3., ötödik lépésben az 1., hatodik lépésben az 5. főkomponenst léptettük be a modellbe. A paraméterek optimalizálása a Newton–Raphson-módszerrel történt. A végső modell a konstanson kívül hat változót tartalmaz. A modell a likelihood-arány próba alapján szignifikáns (szabadságfok: 6, $\chi^2 = 461{,}24$, p-érték: 0,000). A becsült paraméterek együtthatóit, standard hibáit, valamint a Wald-tesztek eredményeit és a p-értékeket a 2. táblázat tartalmazza.

2. táblázat

A logisztikus regressziós modell paraméterei

Változó	Együttható	Standard hiba	Wald-teszt	*p*-érték
Konstans	0,4899	0,0863	32,19	0,000
PCA4	–1,3047	0,1015	165,37	0,000
PCA6	–1,3130	0,1136	133,66	0,000
PCA2	–1,0919	0,1231	78,67	0,000
PCA3	0,6966	0,0972	51,33	0,000
PCA1	–0,2550	0,0482	27,97	0,000
PCA5	–0,2420	0,0545	19,73	0,000

A neurális háló modelljét három köztes réteget tartalmazó, többrétegű perceptronháló-formában[11] fejlesztettük ki. A neurálisháló-modellezés fontos kérdései a tanulási folyamat során kialakuló hálóstruktúra és a neuronsúlyok optimalizálása. A neuronsúlyok optimalizálása felfogható olyan nemlineáris függvényoptimalizálásnak, amely a háló által elkövetett hiba minimalizálását tűzi ki célul. Jelen empirikus vizsgálatban a függvényoptimalizálás a szakirodalomban és a gyakorlatban számtalan neurálisháló-modellezés során – különösen a 100 neuronsúly alatti modellek esetén – már igazoltan jól teljesítő Levenberg–Marquard-algoritmussal (*Mammadli* [2017]) történt. A Levenberg–Marquard-eljárás az exponenciális függvénycsaládból származó függvények négyzetes hibáit igyekszik minimalizálni, amellyel elhanyagolhatóan alacsony konvergenciahibájú paraméterbecslést valósít meg. A túltanulás elkerülése érdekében a tanulási ciklusok addig folytatódtak (azaz a neuronsúlyokat addig mentettük el), míg a tesztelő mintán mért hiba a legalacsonyabb nem lett, és utána növekedésnek nem indult. Az iterációk során végül a három köztes réteggel, összességében 37 neuronnal rendelkező háló bizonyult leginkább optimálisnak.

A CBR-modell a tanulási mintán a *k* legközelebbi szomszédokat euklideszi távolság alapján, a korábbiakban említett RDT-módszerrel határozta meg. A *k* legközelebbi szomszéd megfigyelés célváltozójának értéke alapján valamennyi legközelebbi szomszéd egy bináris (1/0) szavazatot adott a vizsgált megfigyelések hovatartozására vonatkozóan, ahol 1 jelölte a csődbe jutott vállalkozások célváltozójának értékét. A CBR-modell alapján becsült csődvalószínűség a szavazatok számtani átlaga.[12] A modellben figyelembe vett *k* értékét próbálgatással, valamint a modellteljesítmény folyamatos visszamérésével választottuk ki. Tapasztalatok alapján az alacsony (1 és 5 közötti) *k* érték a modell túltanulásának növekedését idézte elő a kevesebb ismert esetre történő specializálódás miatt. Ez a tanulási mintán extrém esetben a neurális hálónál is jobb besorolási pontosságot tett lehetővé, a tesztelő mintán azonban rontotta az eredményt, ami miatt nem bizonyult célszerűnek túl kevés legközelebbi szomszédot specifikálni. A kísérletezésekből ugyanakkor az is kiderült, hogy 15-nél több szomszédot sem indokolt választani, mert dimenzionalitási problémák merülhetnek fel a nem nagy mintaelemszám és a 10 modellváltozó következtében. Emiatt a végső modellben *k* értéke 15 lett.

A négy csődmodell teljesítményét a klasszifikációs modellek értékelésében és validációjában gyakran alkalmazott ROC-görbe[13] segítségével hasonlítottuk össze.

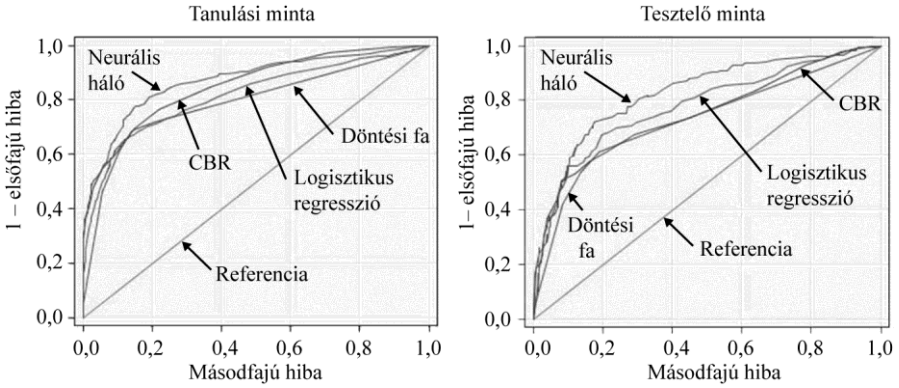
[11] Perceptron: mesterséges neuron.

[12] Például négy legközelebbi szomszéd alkalmazása esetén, ha 3 szomszéd csődbe jutott vállalkozás, míg 1 működő vállalkozás, akkor a becsült csődvalószínűség 3/4 (75%).

[13] A ROC-görbe azt vizsgálja, hogy a modell által becsült valószínűségi értékek mennyire jelzik megbízhatóan az outputkategóriába való tartozást, amennyiben az eredeti besorolás ismert (*Stein* [2005]). A ROC-görbe referenciája a 45°-os egyenes, ami a véletlen találgatásnak felel meg. Annál jobb az értékelése a csődmodellnek, minél jobban elválik a ROC-görbe a 45°-os egyenestől. A ROC-görbéből számított objektív statisztikai mutató a görbe alatti terület nagysága. Minél nagyobb valamely csődmodell ROC-görbe alatti területe, annál megbízhatóbban jelzik a valószínűségi értékek az outputkategóriába tartozást.

Mivel a csődmodellek célja új megfigyelések megfelelő klasszifikálása, a modellek teljesítményének sorrendjét a tesztelő mintán kell megítélni.

5. ábra A négy csődmodell ROC-görbéje a tanulási és a tesztelő mintán



Megjegyzés. Itt és a 3. táblázatnál ROC (receiver operating characteristic): kumulált besorolási pontosság.

3. táblázat

A ROC-görbe alatti területek a tanulási és a tesztelő mintán

Minta	Döntési fa	Logisztikus regresszió	Neurális háló	CBR
Tanulási	0,79	0,82	0,87	0,85
Tesztelő	0,73	0,78	0,82	0,75

A ROC-görbe alatti területtel mért modellteljesítmény alapján mind a tanulási, mind a tesztelő mintán a legmagasabb besorolási pontosságot a neurális háló modellje érte el (87, illetve 82 százalék). A tanulási mintán a második legjobb előrejelző képességgel a CBR-modell rendelkezett (85%), a tesztelő mintán azonban e modell eredménye már csak 75 százalék volt; ezáltal a logisztikus regressziós modellnél gyengébb, ugyanakkor a döntési fánál jobb becslőképesség jellemzi jelen empirikus vizsgálat szerint. Vagyis, az általunk összeállított hazai csődelőrejelzési adatbázison az empirikus eredmények a neurális háló fölényét támasztották alá, ami nem szokatlan az utóbbi években publikált, egyes klasszifikációs módszerek teljesítményét összehasonlító empirikus vizsgálatok alapján.

Hasonló sorrendre és következtetésre juthatunk a kétmintás K–S- (Kolmogorov–Smirnov-) próba eredményeiből. A tesztelő mintán a neurális háló modellje 54,1, a logisztikus regresszió 47,6, a CBR 42,6, míg a döntési fa 42,2 százalékos

K–S-értékkel jellemezhető, vagyis a neurális háló fölénye ezzel a modellteljesítmény-indikátorral is igazolható.

Tanulságos képet mutat az egyes modellek által becsült csődvalószínűség a tesztelő mintán. (Lásd a 4. táblázatot.)

4. táblázat

Becsült csődvalószínűségek és a megfigyelések megoszlása a tesztelő mintán posterior valószínűségsávonként

Posterior valószínűségsáv	Átlagos becsült valószínűség a tesztelő mintán				Megfigyelések megoszlása a tesztelő mintán (%)			
	Döntési fa	Logisztikus regresszió	Neurális háló	CBR	Döntési fa	Logisztikus regresszió	Neurális háló	CBR
0,95–1,00	1,0000	0,9892	0,9812	1,0000	0,70	8,84	12,56	9,30
0,90–0,95	0,9231	0,9314	0,9237	0,9333	2,56	3,26	3,26	2,33
0,85–0,90	0,8771	0,8734	0,8767	0,8667	21,63	4,42	4,42	4,19
0,80–0,85	n. a.	0,8250	0,8234	n. a.	0,00	4,19	2,79	0,00
0,75–0,80	n. a.	0,7764	0,7704	0,8000	0,00	5,58	3,72	4,19
0,70–0,75	0,7480	0,7242	0,7251	0,7333	8,14	2,33	6,51	4,65
0,65–0,70	n. a.	0,6875	0,6784	0,6667	0,00	0,70	2,79	4,19
0,60–0,65	0,6429	0,6235	0,6238	n. a.	6,51	1,40	2,09	0,00
0,55–0,60	n. a.	0,5750	0,5696	0,6000	0,00	0,93	2,79	3,26
0,50–0,55	n. a.	0,5158	0,5280	0,5333	0,00	1,16	3,02	4,19
0,45–0,50	0,4545	0,4716	0,4744	0,4667	0,23	3,49	2,33	4,42
0,40–0,45	n. a.	0,4225	0,4267	n. a.	0,00	4,19	5,35	0,00
0,35–0,40	n. a.	0,3767	0,3720	0,4000	0,00	9,77	4,88	6,98
0,30–0,35	0,3276	0,3260	0,3162	0,3333	4,88	13,02	6,05	11,40
0,25–0,30	0,2486	0,2746	0,2768	0,2667	55,35	19,30	7,91	14,19
0,20–0,25	n. a.	0,2306	0,2201	n. a.	0,00	11,40	5,58	0,00
0,15–0,20	n. a.	0,1814	0,1730	0,2000	0,00	1,86	7,67	12,09
0,10–0,15	n. a.	0,1212	0,1273	0,1333	0,00	1,40	7,91	9,53
0,05–0,10	n. a.	0,0794	0,0776	0,0667	0,00	2,56	8,37	3,95
0,00–0,05	n. a.	0,0408	n. a.	0,0000	0,00	0,23	0,00	1,16
Összesen					*100,00*	*100,00*	*100,00*	*100,00*

Megjegyzés. A táblázat értékei kerekítés miatt nem adják ki a 100,00-ot.

Ötszázalékos valószínűségsávonként elemezve az átlagos becsült valószínűségeket és a megfigyelések megoszlását, megállapítható, hogy a leginkább kiegyensúlyozott csődvalószínűség-eloszlással a neurális háló modellje rendelkezik, ezzel egyidejűleg arányaiban a neurális háló becsülte a legtöbb 95 és 100 százalék közötti csődvalószínűséget, ami – miután a legtöbb megfigyelt csődesemény a tesztelő mintán is

ebbe a valószínűségi sávba tartozik – együtt jár a magasabb klasszifikációs képességgel, különösen az elsőfajú hiba tekintetében. A döntési fa specialitásai (annyiféle csődvalószínűséget becsül, ahány döntési szabály benne létezik) némileg értelmezhetetlenné teszi az összehasonlítást a másik három módszerrel. Szembeötlő ugyanakkor a hasonlóság a CBR és a logisztikus regresszió becsült valószínűségeloszlása között, hiszen mindkét modell a döntési fához hasonlóan legnagyobb arányban 15 és 35 százalék közötti csődvalószínűséget becsült a tesztelő mintában szereplő megfigyelések esetén.

5. Következtetések

A tanulmány azt vizsgálta, hogy mennyire lehet létjogosultsága a napjainkban növekvő népszerűségre szert tevő CBR-módszernek a magyarországi mikrovállalkozások csődelőrejelzésében, klasszifikációs erő szempontjából összevetve a CBR-t a szakirodalomban és a gyakorlatban leggyakrabban alkalmazott három csődelőrejelzési eljárással (a döntési fával, a logisztikus regresszióval és a neurális hálóval).

A tanulmány áttekintést adott a CBR fejlődéstörténetéről, módszertanáról, alkalmazási feltételeiről, a csődelőrejelzés területén rendelkezésre álló nemzetközi empirikus vizsgálatok tapasztalatairól, valamint gyakorlati kihívásairól. Az empirikus vizsgálat 1 828 hazai mikrovállalkozás 2017. évben megfigyelt csődeseményének bekövetkezését vizsgálta azok 2015. és 2016. évi beszámolóinak adataiból számított pénzügyi mutatók alapján.

Összességében megállapítható, hogy érdemes figyelmet szentelni a CBR-módszernek a csődelőrejelzéshez hasonló klasszifikációs problémák megoldása során. A tesztelő mintában szereplő megfigyelések esetén azonban a CBR-modell előrejelző ereje elmaradt a neurális háló és a logisztikus regresszió teljesítményétől; ezáltal jelen empirikus vizsgálat alapján nem javasolt CBR-rel felváltani a gyakorlatban széles körben alkalmazott eljárásokat.

Irodalom

AAMODT, A. – PLAZA, E. [1994]: Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Communications*. Vol. 7. No. 1. pp. 39–59.

AHN, H. – KIM, K. J. [2009]: Bankruptcy prediction modeling with hybrid case-based reasoning and genetic algorithms approach. *Applied Soft Computing*. Vol. 9. No. 2. pp. 599–607. https://doi.org/10.1016/j.asoc.2008.08.002

ALTMAN, E. I. [1968]: Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The Journal of Finance*. Vol. 23. No. 4. pp. 589–609. https://doi.org/10.1111/j.1540-6261.1968.tb00843.x

BROWN, C. E. – GUPTA, U. [1994]: Applying case-based reasoning to the accounting domain. *Intelligent Systems in Accounting, Finance and Management*. Vol. 3. No. 3. pp. 205–221. https://doi.org/10.1002/j.1099-1174.1994.tb00066.x

BRYANT, S. M. [1997]: A case-based reasoning approach to bankruptcy prediction modelling. *Intelligent Systems in Accounting, Finance and Management*. Vol. 6. No. 3. pp. 195–214. https://doi.org/10.1002/(SICI)1099-1174(199709)6:3<195::AID-ISAF132>3.0.CO;2-F

BUTA, P. [1994]: Mining for financial knowledge with CBR. *AI Expert*. Vol. 9. No. 2. pp. 34–40.

GOODWIN, C. J. [2009]: *Research in Psychology – Methods and Design*. John Wiley & Sons. Hoboken.

HAJDU O. [2003]: *Többváltozós statisztikai számítások.* Központi Statisztikai Hivatal. Budapest.

HAJDU O. [2004]: A csődesemény logit-regressziójának kismintás problémái. *Statisztikai Szemle*. 82. évf. 4. sz. 392–422. old.

HAJDU, O. – VIRÁG, M. [2001]: A Hungarian model for predicting financial bankruptcy. *Society and Economy*. Vol. 23. Nos. 1–2. pp. 28–46. https://doi.org/10.2307/41468499

JARMULAK, J. – CRAW, S. – ROWE, R. [2000]: Self–optimising CBR retrieval. In: *Titsworth, F. M.* (ed.): *Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence*. IEEE Computer Society. Vancouver. pp. 376–383.

JO, H. – HAN, I. – LEE, H. [1997]: Bankruptcy prediction using case-based reasoning, neural network and discriminant analysis for bankruptcy prediction. *Expert Systems with Applications*. Vol. 13. No. 2. pp. 97–108. https://doi.org/10.1016/S0957-4174(97)00011-0

KAISER, H. F. [1960]: The application of electronic computers for factor analysis. *Educational and Psychological Management*. Vol. 20. No. 1. pp. 141–151. https://doi.org/10.1177/001316446002000116

KOLODNER, J. L. [1993]: *Case-based Reasoning.* Morgan Kaufmann. San Mateo.

KRISTÓF T. [2005]: A csődelőrejelzés sokváltozós statisztikai módszerei és empirikus vizsgálata. *Statisztikai Szemle*. 83. évf. 9. sz. 841–863. old.

KRISTÓF T. [2008]: A csődelőrejelzés és a nem fizetési valószínűség számításának módszertani kérdéseiről. *Közgazdasági Szemle*. LV. évf. Május. 441–461. old.

KRISTÓF, T. – VIRÁG, M. [2012]: Data reduction and univariate splitting. Do they together provide better corporate bankruptcy prediction? *Acta Oeconomica*. Vol. 62. No. 2. pp. 205–227. https://doi.org/10.1556/AOecon.62.2012.2.4

KSH (KÖZPONTI STATISZTIKAI HIVATAL) [2017]: *A vállalkozások teljesítménymutatói kis- és középvállalkozási kategória szerint* (*2013–*). http://www.ksh.hu/docs/hun/xstadat/xstadat_eves/i_qta005.html#

KSH [2018]: A regisztrált gazdasági szervezetek száma, 2017. *Statisztikai Tükör*. Április 13. https://www.ksh.hu/docs/hun/xftp/gyor/gaz/gaz1712.pdf

LI, H. – YU, J. L. – YU, L. A. – SUN, J. [2014]: The clustering-based case-based reasoning for imbalanced business failure prediction: A hybrid approach through integrating unsupervised process with supervised process. *International Journal of Systems Science*. Vol. 45. No. 5. pp. 1225–1241. https://doi.org/10.1080/00207721.2012.748105

LIU, H. – ZHANG, S. [2012]: Noisy data elimination using mutual *k*-nearest neighbor for classification mining. *Journal of Systems and Software*. Vol. 85. No. 5. pp. 1067–1074. https://doi.org/10.1016/j.jss.2011.12.019

MAMMADLI, S. [2017]: Financial time series prediction using artificial neural network based on Levenberg–Marquardt algorithm. *Procedia Computer Science*. Vol. 120. Special Issue. pp. 602–607. https://doi.org/10.1016/j.procs.2017.11.285

MATIGNON, R. [2007]: *Data Mining Using SAS Enterprise Miner*. John Wiley & Sons. Hoboken. https://doi.org/10.1002/9780470171431

NYITRAI T. [2014]: Validációs eljárások a csődelőrejelző modellek teljesítményének megítélésében. *Statisztikai Szemle*. 92. évf. 4. sz. 357–377. old.

NYITRAI T. [2015]: Hazai vállalkozások csődjének előrejelzése a csődeseményt megelőző egy, két, illetve három évvel korábbi pénzügyi beszámolók adatai alapján. *Vezetéstudomány*. 46. évf. 5. sz. 55–65. old.

NYITRAI T. – VIRÁG M. [2017]: A pénzügyi mutatók időbeli tendenciájának figyelembevétele logisztikus regresszióra épülő csődelőrejelzési modellekben. *Statisztikai Szemle*. 95. évf. 1. sz. 5–28. old. https://doi.org/10.20311/stat2017.01.hu0005

OHLSON, J. [1980]: Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*. Vol. 18. No. 1. pp. 109–131. https://doi.org/10.2307/2490395

PARK, C. S. – HAN, I. [2002]: A case-based reasoning with the feature weights derived by analytic hierarchy process for bankruptcy prediction. *Expert Systems with Applications*. Vol. 23. No. 3. pp. 255–264. https://doi.org/10.1016/S0957-4174(02)00045-3

SCHANK, R. C. [1982]: *Dynamic Memory: A Theory of Reminding and Learning in Computers and People*. Cambridge University Press. New York.

STEIN, R. M. [2005]: The relationship between default prediction and lending profits: Integrating ROC analysis and loan pricing. *Journal of Banking & Finance*. Vol. 29. No. 5. pp. 1213–1236. https://doi.org/10.1016/j.jbankfin.2004.04.008

VIRÁG M. – HAJDU O. [1996]: Pénzügyi mutatószámokon alapuló csődmodell-számítások. *Bankszemle*. 15. évf. 5. sz. 42–53. old.

VIRÁG, M. – FIÁTH, A. [2010]: *Financial Ratio Analysis*. Aula Kiadó. Budapest.

VIRÁG, M. – NYITRAI, T. [2014]: Is there a trade-off between the predictive power and the interpretability of bankruptcy models? The case of the first Hungarian bankruptcy prediction model. *Acta Oeconomica*. Vol. 64. No. 4. pp. 419–440. https://doi.org/10.1556/AOecon.64.2014.4.2

YIP, A. Y. N. [2006]: Business failure prediction: A case-based reasoning approach. *Review of Pacific Basin Financial Markets and Policies*. Vol. 9. No. 3. pp. 491–508. https://doi.org/10.1142/S021909150600080X

ZURADA, J. – LONIAL, S. [2005]: Comparison of the performance of several data mining methods for bad debt recovery in the healthcare industry. *The Journal of Applied Business Research*. Vol. 21. No. 2. pp. 37–54. https://doi.org/10.19030/jabr.v21i2.1488

Summary

Development in science has enabled the improvement of bankruptcy prediction models through several data-driven and artificial intelligence-based methods. One of such promising methods is CBR (case-based reasoning). The aim of this study is to consider the applicability of CBR on a sample of Hungarian micro enterprises, within the framework of classic bankruptcy prediction, by comparing the classification power of CBR to the three most frequently applied bankruptcy prediction techniques (decision tree, logistic regression, neural networks). The empirical research examined the occurrence of bankruptcy events (initiating bankruptcy, liquidation, or forced deregistration procedure) for 1,828 Hungarian micro enterprises in 2017, using financial ratios calculated from their 2015 and 2016 annual reports. Overall, it can be concluded that it is worthwhile to consider CBR methodology in solving classification problems similar to bankruptcy prediction. However, based on the empirical research, the predictive power of the developed CBR model underperformed the accuracy of neural networks and logistic regression on observations in the testing sample.