Acta Agraria Kaposváriensis (2010) Vol 14 No 3, 173-182
Kaposvári Egyetem, Állattudományi Kar, Kaposvár
University of Kaposvár, Faculty of Animal Science, Kaposvár

Support vector gépek alkalmazása hitelpontozó kártyák fejlesztésében

Szücs I.
Szent István Egyetem, Gazdaság- és Társadalomtudományi Kar, Gazdálkodás- és Szervezéstudományi Doktori Iskola
2103 Gödöllő, Páter Károly u. 1.

ÖSSZEFOGLALÁS

Az adatbányászati algoritmusok fejlődésének egy újabb állomását a statisztikai tanuláselméleti kutatások során kialakult support vector gépek jelentik. Az eljárás igen természetes módon közelíti meg az osztályozás problémáját, és komoly hatékonyság-növelést ígér az osztályozási feladatok tanulásában. A Bázel 2 tőkemegfelelőségi szabályozás által megkívánt paraméterek becslésében és a hitelpontozó kártyák fejlesztésében azonban kevéssé terjedt még el a módszer használata. A tanulmányban a support vector gépek alkalmazásának lehetőségét mutatom be egy hitelintézet hitelügyleteinek bedőlési valószínűségének becslésén keresztül, külön kiemelve a modell üzleti értelmezhetőségének kérdését és a várható veszteségre gyakorolt hatását.
(Kulcsszavak: adatbányászat, support vector machine, Bázel II)

ABSTRACT

Using support vector machine in credit scorecard development
I. Szücs
Szent István University, GTK, GSZDI, H-2103 Gödöllő, Páter Károly u. 1.

Support vector machines, came into alive from statistical learning theory, means a new stage in the evoution of datamining algorithms. The method handle the problem of classification in a very natural way, and promise growth is the efficiency of learning statistical patterns. However in Basel II required parameter estimation and in credit scorecard development support vector machines are not wide range used. In this paper it will be shown how to use SVMs for calculating probability of default, placing emphasis on business understanding and impact on expected loss calculation.
(Keywords: datamining, support vector machine, Basel II)

BEVEZETÉS

Az új bázeli tőke-megfelelőségi szabályozás (*Basel 2*, 2004; *Basel 2*, 2005) hatására már minden jelentősebb bank hitelpontozó kártyák (scorecard) segítségével minősíti ügyleteit, s azok eredményét fontos paraméterként veszi figyelembe üzleti döntéseinek meghozatala során. Ez alapján döntik el, hogy egy hiteligénylést befogadnak-e vagy elutasítanak. A befogadott ügyletek esetében pedig a hiteltörlesztés teljes ideje alatt vizsgálják, mekkora valószínűséggel válik nem-fizetővé az ügylet. Ezen számítások eredménye a provízió és a tőketartalék képzésében fejti ki hatását, ami méltán jelzi az előrejelzések pontosságának és megbízhatóságának szükségességét.

A bankok többsége ma még a logisztikus regressziót alkalmazza a hitelpontozó kártyák fejlesztésére, bár már jó ideje rendelkezésre állnak a statisztikai tanuláselmélet újabb, és nagyobb pontosságú becsléseket eredményező eljárásai.

A tanulmány a support vector gépek (SVM) alkalmazásának lehetőségét mutatja be hitelpontozó kártyák fejlesztésére. A módszer segítségével egy publikusan elérhető adatbázis adatait felhasználva kerül fejlesztésre egy hitelpontozó kártya, a jó és rossz ügyletek osztályozási problémájának megoldására. A modell eredményi statisztikai módszerekkel kerülnek ellenőrzésre, továbbá bemutatásra kerül, hogyan befolyásolja a modell alkalmazása a várható veszteség mértékét.

A modell gazdasági értelmezhetősége az egyik legfontosabb kérdésként vetődik fel a support vector gépek hitelpontozó kártya fejlesztésére való alkalmazását illetően.

ANYAG ÉS MÓDSZER

Felhasznált adatok

A tanulmányhoz a Kaliforniai Egyetem gondozásában lévő és kutatási célból szabadon hozzáférhető UCI Machine Learning Repository (http://archive.ics.uci.edu/ml/) credit screening adatbázisa került felhasználásra (*Quinlan*, 1987, 1992). Az adatbázis eredetileg 690 hiteligénylés adatait tartalmazza az ügylet később tapasztalt minősítésével együtt. A változók átnevezésre, a változók által felvett értékek transzformálásra kerültek. Ennek oka, hogy azokból semmilyen következtetést ne lehessen levonni az adatokat szolgáltató pénzintézetről.

Az adatbázis összesen 16 változót tartalmaz, melyeknek típus szerinti megoszlása:

- 6 folytonos változó (C_1 … C_6)
- 9 nominális változó (N_1 … N_9)
- 1 célváltozó - A hitelintézet jó / rossz ügylet definíciójának megfelelően

Az adatbázisban lévő folytonos változók főbb statisztikai jellemzőit az *1. táblázat* mutatja.

1. táblázat

A folytonos független változók statisztikai jellemzői

	C_1	C_2	C_3	C_4	C_5	C_6
Várható érték (1)	31,57	4,76	2,22	2,40	184,01	1017,39
Standard hiba (2)	0,46	0,19	0,13	0,19	6,68	198,35
Medián (3)	28,46	2,75	1,00	0,00	160,00	5,00
Módusz (4)	22,67	1,50	0,00	0,00	0,00	0,00
Minta varianciája (5)	142,99	24,78	11,20	23,65	30208,79	27145169,08
Csúcsosság (6)	1,12	2,27	11,20	50,83	19,50	214,67
Ferdeség (7)	1,15	1,49	2,89	5,15	2,72	13,14
Tartomány (8)	66,50	28,00	28,50	67,00	2000,00	100000,00
Minimum (9)	13,75	0,00	0,00	0,00	0,00	0,00
Maximum (10)	80,25	28,00	28,50	67,00	2000,00	100000,00
Darabszám (11)	678,00	690,00	690,00	690,00	677,00	690,00

Table 1: Descriptive statistics of continous independent variables

Expected Value(1), Standard error(2), Median(3), Modus(4), Variance(5), Kurtosis(6), Skewness(7), Range(8), Minimum(9), Maximum(10), Count(11)

Az eredeti mintában a Jó / Rossz ügylet megoszlás alakulását a *2. táblázat* mutatja.

2. táblázat

Jó és Rossz ügyletek megoszlása a mintában

	Minta elemszám (1)	**Százalék (2)**
Jó ügylet (3)	307	44,49%
Rossz ügylet (4)	383	55,51%
Összes hiteligénylés (5)	690	

Table 2: Distribution of Good and Bad applications int he sample

Piece of applications int he sample(1), Percent(2), Good application(3), Bad application(4), All application(5)

A tanulmánynak nem tárgya a hiányzó értékek kezelésének problematikája. Mivel a hiányzó értékkel rendelkező rekordok száma csupán 24, így a modellezés során kialakított állományokból azok törlésre kerültek, hogy a hiányzó adatok pótlásának módszere ne befolyásolja az eredményeket. A modellezéshez az alapállományból tréning és teszt állomány került kialakításra, 2:1 arányban:

- tréning állomány: a modell fejlesztéséhez használt halmaz
- teszt állomány: a modell eredményességének visszaméréséhez használt minta

Az így keletkezett állományok jó / rossz ügylet megoszlását mutatja a *3. táblázat*.

3. táblázat

A modellezés során használt állományok jó/rossz ügylet megoszlásai

	Teljes fejlesztési minta (1)		**Tréning állomány (2)**		**Teszt állomány (3)**	
	Elemszám (4)	Százalék (5)	Elemszám	Százalék	Elemszám	Százalék
Jó ügylet (6)	299	44,89%	200	45,05%	99	44,59%
Rossz ügylet (7)	367	55,11%	244	54,95%	123	55,41%
Összes hiteligénylés (8)	666		444		222	

Table 3: Good / Bad distributions in different sets used in modeling

Total development set(1), Training set(2), Test set(3), Piece of applications int he sample(4), Percent(5), Good application(6), Bad application(7), All application(8)

Új változók képzése

A support vector gépek független változóinak értékeire vonatkozóan a következő elvárások fogalmazhatóak meg (*Chih et al.*, 2008):

- csak numerikus értékek lehetnek
- a hatékonyság érdekében érdemes a változókat azonos skálára transzformálni

- A nominális változókat kategóriánként külön változókba érdemes transzformálni. Túl sok kategória esetén a statisztikai szempontból hasonló csoportok összevonása lehetséges.

Ezen elvárásoknak megfelelően a következő változó transzformációk kerültek elvégzésre:

- A folytonos független változók 0-1 értéktartományra való transzformálása, lineáris leképezéssel: új_változó = (régi_változó – min_érték) / (max_érték – min_érték)
- A nominális független változók értékei kategóriánként a WOE (Weight Of Evidence) alapján csoportosításra kerültek. Majd annyi 0/1 változó került kialakításra, ahány csoport képződött az adott kategória értékeiből. Így az eredeti 9 nominális változóból 23 darab változó került kialakításra.
- A függő- vagy célváltozó eredeti adatbázisban található + / - értékei, +1 / -1 értékekre lettek transzformálva.

A nominális változók értékeinek csoportosítása az egyes kategóriákhoz rendelhető WOE érték alapján történt, ami az adott kategória relatív kockázatát számszerűsíti:

$$WOE_{attribútum} = \ln \frac{p^{nemesemény}_{attribútum}}{p^{esemény}_{attribútum}} \tag{1}$$

ahol: $p^{esemény}_{attribútum} = \frac{n^{esemény}_{attribútum}}{N^{esemény}}$ és $p^{nemesemény}_{attribútum} = \frac{n^{nemesemény}_{attribútum}}{N^{nemesemény}}$,

$n_i^{esemény}$ és $n_i^{nemesemény}$ a rossz illetve jó ügyletek számát jelöli.

A WOE érték segítségével az egyes csoportokban lévő jó/rossz ügyletek arányára alapján lehet eldönteni, érdemes-e a két kategóriát egybe vonni vagy sem. A WOE=0 azt jelenti, hogy az adott csoportban a jó és rossz ügyletek aránya megegyezik. A WOE értékében 0-tól távolodva egyre nagyobb arányban vannak a jó vagy a rossz ügyletek az adott csoportban. Az *1. ábra* a WOE értékét mutatja a csoporton belüli esemény bekövetkezésének valószínűsége függvényében.

1. ábra

WOE értékének alakulása az esemény bekövetkezési valószínűségének függvényében.

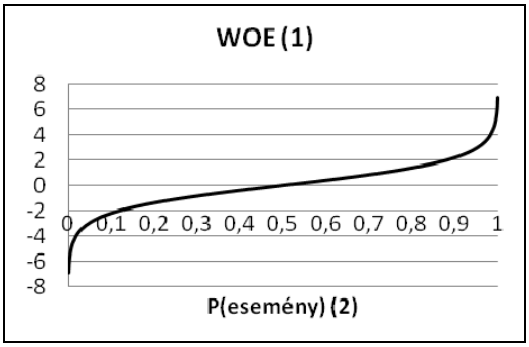

Figure 1: WOE as the function of probability of event.

Weight Of Evidence(1), Probability of Event(2)

Support Vector Gépek

A Support Vector Gépek (Support Vector Machine - SVM) alkalmazása a statisztikai tanuláselméleti kutatások révén terjedt el. (*Vapnik*, 1998) Mára széles körben alkalmazzák osztályozási és regreszsiós problémák megoldására.

Az eljárás alapja, hogy az adott tréning halmazbeli x_i, y_i párokra, ahol $x_i \in R^n$ és $y_i \in \{-1,+1\}$, a következő optimalizálási probléma megoldását keressük:

$$\min_{w,b,\xi} \frac{1}{2} w^T w + C \sum_{i=1}^{l} \xi_i , \tag{2}$$

$$y_i \left(w^T \phi(x_i) + b \right) \geq 1 - \xi_i , \text{ ahol } \xi_i \geq 0 . \tag{3}$$

A Φ függvény segítségével az x_i pontokat egy nagyobb, akár végtelen dimenziós térbe transzformáljuk. Az SVM egy lineárisan szeparáló hipersíkot keres ebben a magasabb dimenziójú térben. A gyakorlati életben tökéletesen szeparáló hipersíkot találni még ebben a magas dimenziószámú térben sem kivitelezhető, így a C paraméteren keresztül teszünk engedményt, valamekkora hiba vétésére. A $K(x_i, x_j) = \Phi(x_i)^T \Phi(x_j)$ függvény a kernel függvény. Számos kernel függvény terjedt el a használatban. Jelen tanulmányban az RBF (Radial Basis Function) kernel függvény került alkalmazásra:

$$K(x_i, x_j) = \exp(-\gamma || x_i - x_j||^2), \text{ ahol } \gamma>0. \tag{4}$$

Az SVM tanításához kizárólag a tréning állomány került felhasználásra. A túltanulás elkerülése érdekében ún kereszt-validáció került alkalmazásra, melynek során a tréning állomány 5 részre lett osztva, majd a tanulás iteratív módon 5-ször egymás után, 4 tanulási és 1 ellenőrző minta segítségével történt.

Hogy a szeparáló hipersíktól való távolsággal arányos osztályba tartozási valószínűséget lehessen képezni az SVM regressziós eljárásként került alkalmazásra. Ezt követően a becslés eredménye a logisztikus transzformáció segítségével {0,1} tartományba lett leképezve. Az így kapott becsült értékekkel az ügyletek között egy sorrendet ehet felállítani, ami az alapját képezi a modellek kiértékelésénél alkalmazott ROC illetve LIFT görbéknek.

Szoftver

A modellezés az R-Project gondozásában lévő, R statisztikai programozói környezetben történt. A szoftver kutatási és oktatási célokra ingyenesen használható. Az SVM futtatásához az e1071, a modell kiértékeléséhez a ROCR csomagok kerültek felhasználásra. (*R-Project*)

Kiértékelés módszertana

A modell kiértékelése a szeparálóképesség mérésével történt (*Sobehart et al.*, 2000, *Engelmann et al.*, 2003)

- ROC (Receiver Operating Characteristic)
- AUC (Area Under Curve – görbe alatti terület)
- LIFT görbe
- Kolmogorov-Smirnov statisztika

A ROC görbe ábrázolása a következőképpen történt (*Sobehart and Keenan*, 2001):

- a vízszintes (FAR - False Alarm Rate) tengelye: a tévesen „rossz” ügyfélnek sorolt ügyfelek aránya az összes „jó” ügyfélhez viszonyítva, adott becsült valószínűség mellett. FAR(C)=F(C)/NND, ahol F(C) azon „jó” ügyfelek száma, akik tévesen „rossz” ügyfélnek lettek minősítve, NND a mintában lévő összes „jó” ügyfél száma.

- Függőleges (HR - Hit Rate) tengelye: a helyesen „rossz" ügyfélnek sorolt ügyfelek aránya az összes „rossz" ügyfél számához viszonyítva, adott becsült valószínűség mellett. HR(C) = H(C) / ND , ahol H(C) az adott C „cutoff" pontnál helyesen „rossz" ügyfélnek minősített ügyfelek száma, ND a mintában lévő összes „rossz" ügyfél száma.

2. ábra

ROC

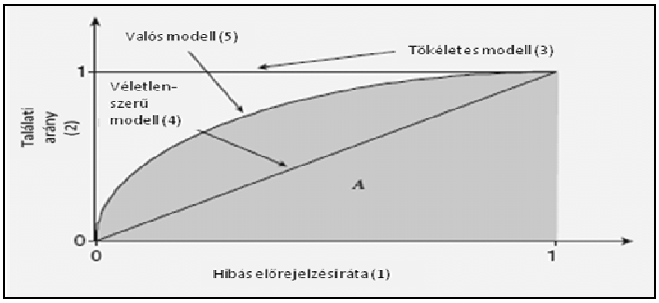

Figure 2: Receiver Operating Curve

False alarm rate(1),Hit rate(2), Perfect model(3), Random modell(4), Rating modell(5)

Amennyiben egy számmal szeretnénk jellemezni a modell szeparáló képességét, úgy a görbe alatti terület (AUC) ennek egy lehetséges megközelítése:

$$AUC = \int_{FAR=0}^{1} HR(FAR)d(FAR) \tag{5}$$

Szeparáló-képességgel nem rendelkező modellek esetén A=0,5, míg tökéletes modellekre A=1. A gyakorlatban alkalmazott modellek esetén nyilván 0,5 és 1 közötti értéket vesz, és a vizsgált modell annál jobb, minél közelebb van az A értéke 1-hez.

EREDMÉNY ÉS ÉRTÉKELÉS

Az SVM eredményének kiértékelése kizárólag a teszt állományon történt.

Az SVM modell statisztikai kiértékelése

A modell szeparáló-képességének vizualizálására szolgáló ROC látható a *3. ábrán*.

A 45 fokos egyeneshez képest lényegesen a bal felső sarokhoz húzódó görbe alakja azt mutatja, hogy a modell erősen szeparálja a jó és rossz ügyleteket. A görbe alatti terület nagysága: AUC = 0,92, ami szintén az erős szeparáló-képességre utaló érték. A gyakorlatban előforduló modellek esetén a 0,8-es ROC alatti terület már bevezethető modellnek tekinthető.

A LIFT görbe azt mutatja meg, hogy ha a becsült osztályba tartozási valószínűség szerinti top x%-ot választanánk ki, akkor a véletlenszerű kiválasztáshoz képest hányszor több esemény fordulna elő a kiválasztott mintában. Mint a *4. ábrán* látható a legjobb 50%-át kiválasztva az ügyleteknek, a jó ügylet arány 1,6-szorosára nő a véletlenszerű kiválasztáshoz képest.

3. ábra

Az SVM kiértékelése a ROC segítségével

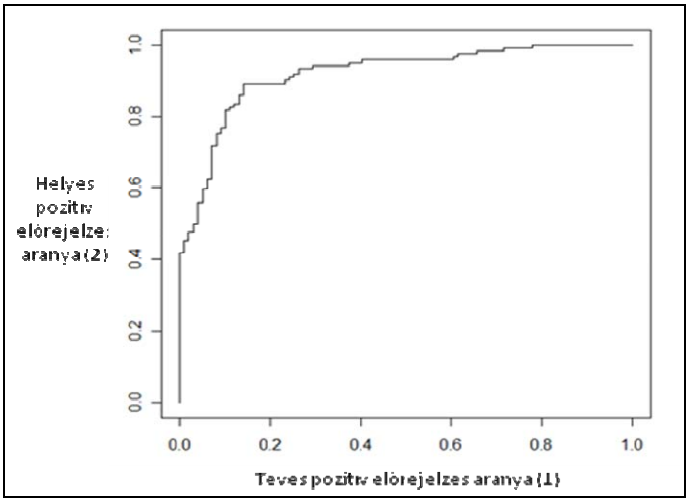

Figure 3: ROC of SVM model

False positive rate (1), True positive rate (2)

4. ábra

Az SVM kiértékelése LIFT görbe segítségével

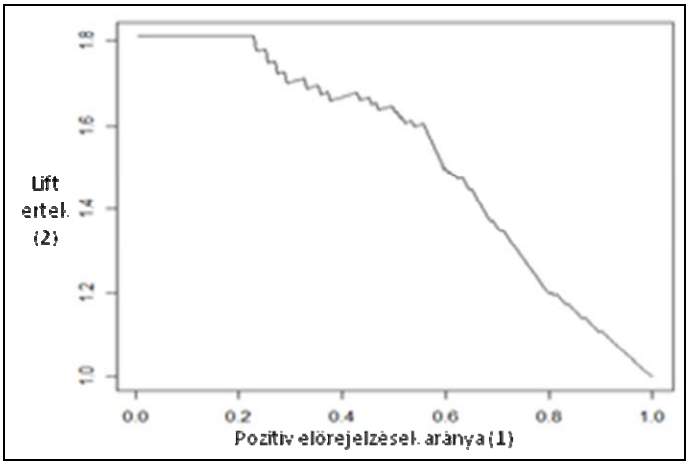

Figure 4: LIFT curve of SVM model

Rate of positive predictions (1), Lift value (2)

Az *5. ábrán* látható, hogy a jó ügyletek sűrűségfüggvénye az elvárt módon, az alacsony pontszámok mellett meredeken nő, majd lassan közelít a maximális, mintában található értékhez. A rossz ügyfelek sűrűségfüggvénye épp fordítva viselkedik. Az alacsony pontszámok mellett alacsony meredekségű, majd egy határt elérve a lehetséges maximális meredekséget veszi fel.

A két függvény különbségeként előállítható KS görbe is az elvárt alakot veszi fel. Látható, hogy a modell szeparálóképességének maximumát a pontszám szerinti 43%-nál veszi fel, értéke 72,17%.

5. ábra

Az SVM modell kiértékelése a Kolmogorov-Smirnov statisztika segítségével

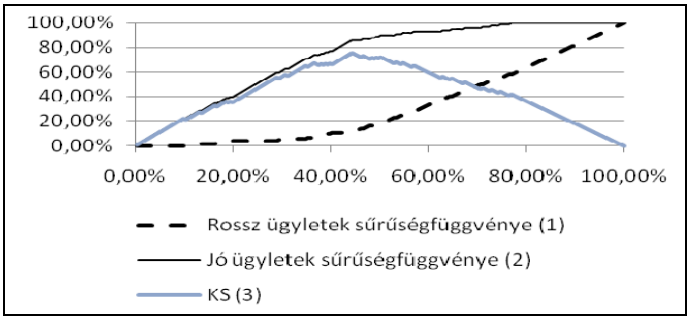

Figure 5: Kolmogorov – Smirnov statistic of the SVM model

Bad applications cumulative density function(1), Good applications cumulative density function(2), Kolmogorov-Smirnov statistic(3)

Modell gazdasági kiértékelése

A bedőlés valószínűségének pontos becslése a várható veszteség miatt képzendő provízión és a nem-várható veszteség kezelésére képzendő tőketartalékon keresztül fejti ki hatását:

- Várható veszteség = PD * LGD
- Nem várható veszteség = $LGD * N\{(1-R)^{-1/2}*N^{-1}(PD)+(R/(1-R))^{1/2}N^{-1}(0,999)\}-PD*LGD$

A beengedési arány megválasztásával a provízió és a tőketartalék mértéke egyaránt befolyásolható. A *6. ábra* 20%-os LGD-t feltételezve mutatja, a beengedett portfolió rossz ügyleteinek arányát és a várható veszteség mértékét. A számításokhoz R értéke 0,04, ami a rulírozó hiteltermékre vonatkozóan érték az új bázeli tőkekövetelmény szabályozásának megfelelően (*Basel 2*, 2004).

Modell gazdasági értelmezhetősége

A support vector gépek – és a neurális hálózatok - kitűnő tanulási képességgel rendelkeznek. A koordináta-transzformáció matematikai szempontból szükséges és fontos lépése a tanulásnak, hisz az új térben nyílik lehetőség a lineáris szeparáló sík egyenletének meghatározására. A módszer legnagyobb hátránya annak „fekete doboz" jellege. Nehezen tudjuk megmondani, melyik változó milyen mértékben és milyen módon befolyásolta a kimenetet. Többdimenziós függvények esetén maga a kérdés is

igen nehezen tehető fel - elég egy domborzati térképre gondolni, és megpróbálni megválaszolni, hogy a szélességi vagy hosszúsági fokok befolyásolják jobban a magassági szintet. A „fekete doboz" jellegű modellezés gyakorlati életben való elterjedésének épp az a korlátja, hogy tapasztalat szerint a kiinduló adatbázisok számos olyan zajjal, szennyeződéssel vannak tele, melyek jelenlétét a szakterületi szakértők nem tudják előre megmondani, s a modellezés iteratív jellege adja a lehetőséget a zajok, szennyeződések észrevételére. Egy-egy modell bevezetésének, cseréjének kérdése pedig súlyos következményekkel jár, így számos esetben szívesebben választják a kevésbé jól tanuló, de azt biztosan teljesítő modelleket a pontosabb becslést ígérő, de nagyobb bevezetési kockázattal járó modellek helyett.

6. ábra

A legjobb x% ügylet kiszűrése mellett elérhető rossz ügylet arány és várható veszteség

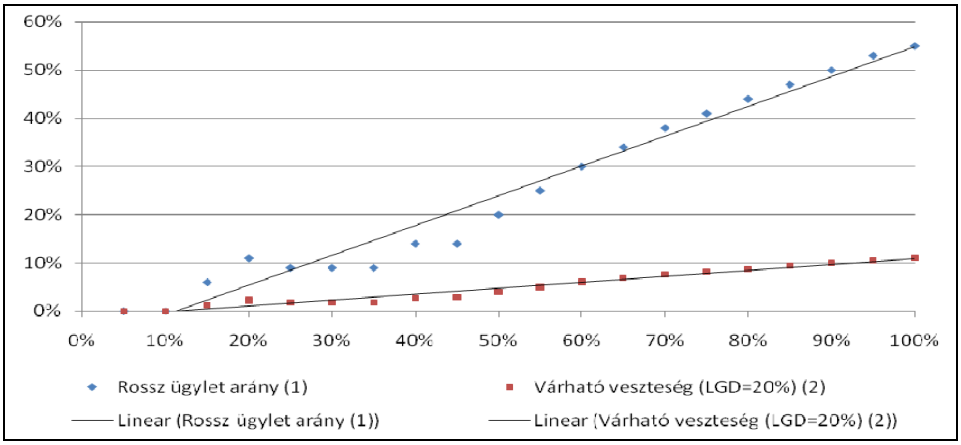

Figure 6: Bad ratio and expected loss in case of th ebest X% of applications

Bad application ratio(1), Expected Loss(2)

KÖVETKEZTETÉSEK

A tanulmányban bemutatott support vector gépek matematikailag jól megalapozott, hatékony eszközei az osztályozási problémák megoldásának. Ezen témakörbe tartoznak az igénylési illetve viselkedési hitelpontozó kártyák, melyek egy adott ügylet igénylésének pillanatában illetve a hiteltörlesztés folyamán jelzik előre a bedőlés valószínűségét. A modellek helyességének ellenőrzése statisztikai szempontból részletesen elvégezhető, azonban a fekete doboz jellegnek köszönhetően a gazdasági szempontú vizsgálatok nehézkesen eszközölhetőek.

A provízió és tőketartalék képzésen keresztül a bedőlés valószínűségének becslése nagymértékben befolyásolja a pénzintézetek versenyképességét, így a pontosabb becsléseket eredményező módszerek alkalmazása várható a jövőben. A modellek közgazdasági megfontolások szerinti vizsgálata egyrészt természetes igény, másrészt a Bázel 2 előírások által támasztott követelmény. Ennek következtében egyre fontosabbá

válik a modellek gazdasági vizsgálhatóságának kutatása (*Szücs*, 2007) és a gazdasági környezet beépítése a modellezés folyamatába (*Szücs és Pitlik*, 2007).

IRODALOM

Basel Comittee on Banking Supervision (2004). Basel II: International Convergence of Capital Measurement and Capital Standards: a Revised Framework

Basel Comittee on Banking Supervision (2005). Studies on Validation of Internal Rating Systems

Hsu, C.W., Chang, C.C., Lin, C.J. (2008): A Practical Guide to Support Vector Classification, Department of Computer Science, National Taiwan University, Taipei 106, Taiwan, http://www.csie.ntu.edu.tw/~cjlin

Engelmann B., Hayden E., Tasche D. (2003). Testing rating accuracy. www.risk.net

Quinlan S. (1987): Simplifying decision trees, Int J Man-Machine Studies 27. 221-234. p.

Quinlan S. (1992): C4.5: Programs for Machine Learning, Morgan Kaufmann

R-Project: http://www.r-project.org

Sobehart J., Keenan S., Stein R. (2000). Validation methodologies for default risk models. Credit, 51-56 p.

Sobehart J., Keenan S. (2001). Measuring default accurately, Risk, S31-S33, 2001. March

Szücs I. (2007). Unstable regions in the scorecards' input space, Business Sciences – Symposium for Young

Szücs I., Pitlik L. (2007): Lakossági termékvásárlási modellek és viselkedési hitelpontozó kártyák fejlesztése makrogazdasági peremfeltételekkel, Acta Agraria Kaposváriensis 11. 2. 153-163. p.

Researchers, Budapest Tech of Hungary, 181-186 pp. ISBN: 978 963 71 54 64 5

Vapnik V. (1998): Statistical learning theory, John Wiley & Sons, Inc., USA

Levelezési cím (*Corresponding author*):

Szücs Imre
Szent István Egyetem, Gazdálkodás- és Szervezéstudományi Doktori Iskola
2103 Gödöllő, Páter Károly u. 1.
Szent István University, GTK, GSZDI
H-2103 Gödöllő, Páter Károly u. 1.
Tel.: +36-70-311-9770
e-mail: icsusz@gmail.com