Acta Agraria Kaposváriensis (2006) Vol 10 No 3, 109-120
Kaposvári Egyetem, Állattudományi Kar, Kaposvár
University of Kaposvár, Faculty of Animal Science, Kaposvár

A partícionálás hatása fuzzy asszociatív osztályozók teljesítményére

Pach[1] F.P., Gyenesei[2] A., Németh[1] S., Árva[1] P., Abonyi[1] J.
[1]Pannon Egyetem, Folyamatmérnöki Tanszék, Veszprém, 8200, Egyetem u. 10.
[2]Department of Knowledge and Data Analysis, Unilever Research Vlaardingen, The Netherlands

ÖSSZEFOGLALÁS

Az osztályozás az egyik legszélesebb körben használt adatbányászati technika. Egy osztályozási feladatnál az egyik legfontosabb tulajdonság az osztályozás pontossága, azonban sok alkalmazási területen rendkívül fontos szempont, hogy az osztályozási eredmény áttekinthető, egyszerűen értelmezhető legyen. A fuzzy szabály alapú osztályozó rendszerek a felhasználók számára könnyen értelmezhető „Ha…Akkor" típusú szabályok formájában tartalmazzák az osztályozáshoz feltárt összefüggéseket. A cikk egy fuzzy asszociációs szabály alapú osztályozási módszert javasol, illetve annak továbbfejlesztési lehetőségeit ismerteti. Mivel az osztályozási teljesítményt a numerikus attribútumok partícionálása jelentősen befolyásolhatja, ezért két eltérő jellegű adatsorra alkalmazva részletesen megvizsgáltunk fuzzy és éles partícionáló módszereket is. Az elvégzett vizsgálatok alapján megállapítható, hogy a fuzzy csoportosítási módszerek alkalmazásával nagyobb osztályozási teljesítmény érhető el.
(Kulcsszavak: fuzzy logika, asszociációs szabály, osztályozás, partícionálás, csoportosítás)

ABSTRACT

The effect of partitioning for the performance of fuzzy associative classifiers

F.P. Pach[1], A. Gyenesei[2], S. Németh[1], P. Árva[1], J. Abonyi[1]
[1]Pannon University, Department of Process Engineering, H-8200, Veszprém, Egyetem u. 10.
[2]Department of Knowledge and Data Analysis, Unilever Research Vlaardingen, The Netherlands

Classification is one of the most popular and extensively applied techniques in data mining. The efficiency of a classification model is evaluated by two parameters, namely the accuracy and interpretability of the model. This paper proposes a fuzzy association rule-based classifier methodology that meets both criteria. Using the fuzzy concept, the obtained model is easily understandable and interpretable for the users. Since the accuracy of a classification model can be largely affected by the partitioning of numerical attributes, this paper discusses several fuzzy and crisp partitioning techniques. The effect of partitioning methods is examined on different case studies. The results of analysis show that classifier methods with fuzzy clustering based partitioning serve higher classification performance.
(Keywords: fuzzy logic, association rule, classification, partitioning, clustering)

BEVEZETÉS

Egy osztályozási feladat adathalmaza olyan mintákat (rekordokat) tartalmaz, amelyek a bemeneti változókon felvett értékekből és egy osztálycímkéből állnak. Az osztályozási feladat lényege, hogy az adott probléma ismert (ún. tanító) adatmintáit felhasználva, egy olyan becslési

modellt identifikáljunk, amellyel az ismeretlen minták kimenetét (az osztálycímkét) meghatározott pontossággal tudjuk megbecsülni a minták bemeneti változókon felvett értékeiből. Az osztályozásról részletesebben olvashatunk Abonyi könyvében *Abonyi* (2006).

Az utóbbi évtizedben *Agrawal* (1993) cikke nyomán a gyakori elemhalmazok és az általános „Ha…akkor" ($X \rightarrow Y$) formátumú asszociációs szabályok feltárására számos algoritmust fejlesztettek ki. A gyakori elemhalmazok kereséséről, illetve az asszociációs szabályok generálásáról részletesen olvashatunk Abonyi János könyvének 6. fejezetében, ahol a fejezet végén egy alapos irodalmi összefoglalást is elhelyeztünk *Abonyi* (2006).

A feltárt szabályok általános célú felhasználása mellett, az egyik fő kutatási irányt az asszociatív osztályozók képviselik: *Liu* (1998), *Dong* (1999), *Meretakis* (1999), *Liu* (2000), *Wang* (2000), *Li* (2001), *Yin* (2003), *Zimmermann* (2004).

Az asszociatív osztályozó modell az adott osztályozási probléma tanító adathalmazán feltárt asszociatív osztályozó szabályok egy halmazán alapul. Egy asszociatív osztályozó szabály a következő alakban írható fel: $X \rightarrow C_j$, ahol a szabály előzmény része (X) valamely attribútumokból és azok értékeiből, a következmény rész, pedig egy osztálycímkéből (C_j) áll. Az osztályozó előállítására a legelterjedtebb megközelítés, hogy első lépésben valamely algoritmussal asszociatív osztályozó szabályokat tárnak fel, majd egy másik módszerrel kiválasztják az osztályozási célra leginkább megfelelő szabályokat. Mindkettő lépésnek nagy szerepe van a végleges osztályozási teljesítmény alakulásában.

Korábbi munkáinkban már bemutattunk egy olyan fuzzy asszociációs szabálykereső módszert, amellyel többek között kompakt, áttekinthető (szerkeszthető) és pontos osztályozó modelleket is identifikálhatunk *Pach* (2005), *Pach* (2006). A módszerhez kapcsolódó alapfogalmakat, főbb jelöléseket a függelékben helyeztük el. Módszerünk főbb lépései az *1. ábrán* láthatóak.

1. ábra

Fuzzy asszociatív osztályozó módszer fő lépései

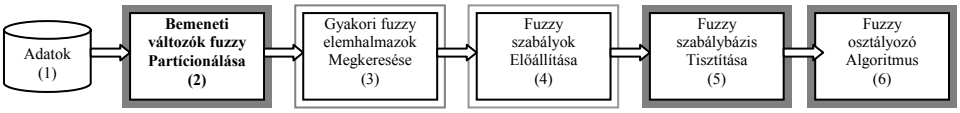

Figure 1. Main steps of the proposed fuzzy associative classification method

Data(1), Fuzzy partitioning of input variables (2), Searching of frequent fuzzy item sets(3), Generation of fuzzy rules(4), Cleaning of fuzzy rule base(5), Fuzzy classification algorithm(6)

A kifejlesztett módszer továbbfejlesztési lehetőségei (az *1. ábrán* kiemelve) a következők:

1. megfelelő bemenet partícionálási technika kiválasztása
2. hatékonyabb szabálybázis tisztító algoritmus létrehozása
3. pontosabb osztályozási eljárás kifejlesztése, alkalmazása.

Cikkünk fő célja az első fejlesztési lehetőség körüljárása, vagyis a szóba jöhető partícionálási technikák bemutatása, illetve az osztályozási teljesítményre gyakorolt hatásuk vizsgálata. Az osztályozó modellek bemeneti változói, attribútumai folytonos, vagy diszkrét értékűek lehetnek, a kimeneti attribútum (az osztály) viszont az adathalmaz valamely kategorikus változója (pl. testmagasság: alacsony, magas) lehet. Az asszociatív osztályozók többségénél a szabályok feltárásához valamilyen gyakori

elemhalmaz keresési módszert, például az Apriori algoritmust alkalmazzák *Agrawal* (1994). Tehát a folytonos attribútumokat tartalmazó adathalmazok esetében, az attribútum elemek (mint diszkrét értékek) meghatározására valamilyen partícionálási módszerre van szükség. Alapvetően kétféle megközelítés lehetséges:

I. kategóriák előzetes, manuális beállítása (pl. alacsony, közepes, magas)
II. adatvezérelt, automatikus felosztás

Amíg bizonyos területeken megfelelő eredményt nyújthat valamely manuális módszer alkalmazása, sok esetben (például az attribútumok nagy száma, vagy éppen az előzetes ismeretek hiányában) nélkülözhetetlen egy adat vezérelt, automatikus felosztási technika alkalmazása. Alapvető követelmény, hogy a szóba jöhető technikák közül, az osztályozási teljesítményt legjobban növelő módszert válasszuk. Cél továbbá az is, hogy az adott módszer gyorsan és áttekinthető módon generálja le a partíciókat.

Az alkalmazott felosztás, a partíciók átfedése alapján *éles*, vagy *fuzzy* lehet. A legegyszerűbb éles technika az egyenlő intervallum felosztás (*equal interval width*), amelynél egy attribútumon a partíciókat meghatározott számú, megegyező hosszúságú intervallumokkal definiáljuk. A módszer hátránya, hogy nem veszi figyelembe az adatok eloszlását, így gyakran eredményezhet ritka (kis mintával rendelkező) partíciókat *Catlett* (1991). Az egyenlő gyakoriságú intervallumokra (*equal frequency intervals*) történő felosztás viszont már figyelembe veszi az adatok eloszlását is. Ennek megfelelően az intervallumok hossza eltérő, azaz k intervallum és m adatpont esetén, minden intervallumba m/k adatpont tartozik. E módszer már sokkal hatékonyabb felosztást eredményezhet, azonban fő hátránya, hogy egy osztályozási problémánál nem használja fel a rendelkezésre álló osztálycímke információt, tehát egy *nem felügyelt* partícionáló technika. A *felügyelt* felosztási módszerek viszont a partíciók meghatározásánál figyelembe veszik az osztálycímkék eloszlását is, így a meghatározott partíciók magát az osztályozási problémát is tükrözik.

Számos felügyelt és nem felügyelt módszer létezik. A módszerek egyik lehetséges csoportosítását az *1. táblázatban* tüntettük fel *Dougherty* (1995). Az osztálycímke felhasználása mellett, a tanítás során az attribútumok figyelembe vétele alapján megkülönböztetünk globális, illetve lokális algoritmusokat. Amíg a lokális módszerek az adott probléma minden attribútumára külön-külön állapítják meg a partíciókat, addig a globális algoritmusok a partíciók meghatározásakor az összes attribútumot együttesen veszik figyelembe.

Mivel fő célunk egy olyan asszociatív osztályozási módszer kifejlesztése, amellyel áttekinthető, könnyen értelmezhető, ugyanakkor pontos osztályozó modelleket határozhatunk meg, az értelmezhetőség alapvető szempont kell, hogy legyen már a felosztási módszer kiválasztásában is. A fuzzy logika felhasználása mellett szól, hogy alkalmazásával a feltárt szabályok természetesebb módon reprezentálhatóak a felhasználó számára, illetve sokkal robosztusabb (hamis, inkonzisztens és a hiányzó adatok megfelelő kezelése) osztályozó modellek létrehozása teszi lehetővé. Cikkünkben éppen ezért főleg fuzzy felosztási módszereket vizsgálunk.

A VIZSGÁLT PARTÍCIONÁLÁSI TECHNIKÁK

A fejezetben négy felosztási módszert ismertetünk. Elsőként egy globális, nem felügyelt, fuzzy technikát, a Ruspini-típusú felosztás jellemzőit mutatjuk be. Ezt követően, a fuzzy Gustafson-Kessel (GK) csoportosítási algoritmus alkalmazhatóságát tárgyaljuk, amely egy nem felügyelt, lokális felosztási módszer *Gustafson* (1979). Majd egy további csoportosítási algoritmust, a felügyelt Gath-Geva (GG) algoritmust ismertetjük, amely a globális módszerek közé tartozik *Gath* (1989). Végül a fuzzy felosztási módszerekkel összevetve a C4.5 algoritmus által alkalmazott éles felosztást elemezzük.

1. táblázat

Partícionáló módszerek egy lehetséges csoportosítása

	Globális (1)	**Lokális (2)**
Felügyelt (3)	1RD adaptív kvantáló ChiMerge D-2 Fayyad and Irani / Ting felügyelt MCC becslő érték max *Gath-Geva csoportosítás (fuzzy)*	vektor kvantálás hierarchikus maximum entrópia Fayyad and Irani *C4.5*
Nem felügyelt (4)	egyenlő intervallum felosztás egyenlő gyakoriságú intervallumok nem felügyelt MCC *Ruspini-típusú felosztás (fuzzy)*	k-means csoportosítás *Gustafson-Kessel csoportosítás (fuzzy)*

Table 1. A possible categorization of the partitioning techniques

Global (1), Local (2), Supervised (3), Unsupervised (4)

Ruspini-típusú felosztás alkalmazása

Ahogy már a bevezető fejezetben említettük, az egyik legegyszerűbb partícionálási módszer, amikor az attribútumokat egyenletes felosztással partícionáljuk. A fuzzy logika alkalmazásával egy adatpont az adott attribútumon felvett értékével minden (tagsági függvénnyel definiált) partícióhoz más-más tagsági függvény értékekkel egyszerre tartozik. Az egyenletes fuzzy felosztást, Ruspini-típusú partícionálásnak hívjuk (*2. ábra*) és alap esetben háromszög alakú tagsági függvényeket alkalmazunk, ahol a háromszög csúcspontjait rendre *a*, *b*, illetve *c* jelöli.

2. ábra

Ruspini-típusú attribútum partícionálás

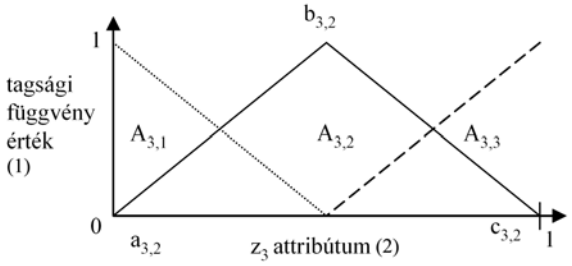

Figure 2. The Ruspini-type partitioning of an attribute

Membership value(1), Attribute z_3(2)

Az ábrán is jól látható, hogy minden tagsági függvény *a* és *c* pontjai a szomszédos tagsági függvények *b* pontjához tartozó értéknél találhatóak. Ezt a felosztást alkalmazva minden attribútumra teljesül, hogy egy adatpont (*x*) tagsági függvény értékeinek ($A_{j,i}(x_{j,k})$) összege egy attribútumon belül mindig egyenlő eggyel:

$$\sum_{i=1}^{n_i} A_{j,i}(x_{j,k}) = 1, \forall\ j,k,\ \text{a}hol\ \ a_{j,i} = b_{j,i-1}\ \text{és}\ c_{j,i} = b_{j,i+1}, \tag{1}$$

Gustafson-Kessel csoportosítási algoritmus alkalmazása

A fuzzy Gustafson-Kessel csoportosítási algoritmus egy nem felügyelt lokális felosztási módszernek felel meg. Minden z_j bemeneti adatra meghatározza a v^j csoportközéppontokat és az $U^j \in [0,1]^{n_j \times N}$ partíciós mátrixot, ahol a mátrix elemei a $z_{j,k}$ adatpont tagsági függvény értékeit reprezentálják az *i*-edik csoportra vonatkozóan ($k = 1,...,N$ és $i=1,...,q_j$, ahol q_j a csoportok számát jelöli). A létrejövő csoportok közvetlenül felhasználhatóak az adatpontok fuzzy értékeinek a meghatározásához, például: $A_{j,i}(x_{j,k})=U^j_{k,i}$. A fuzzy halmazok ($A_{j,i}(x_{j,k})$) reprezentálásra viszont érdemes paraméterezett, például trapéz alakú tagsági függvényeket definiálnunk (*3. ábra*).

3. ábra

Attribútum partícionálás Gustafson-Kessel csoportosítással

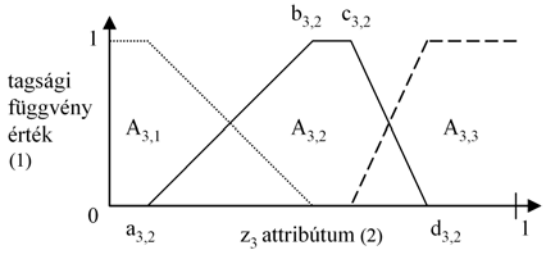

Figure 3. The Gusfason-Kessel clustering based partitioning of an attribute

See Figure 2

Minden trapéz tagsági függvény négy paraméterrel jellemezhető, mégpedig a trapéz csúcspontjaival, melyek rendre: *a*, *b*, *c* és *d*. A csúcspontokra a Ruspini-módszernél ismertettet elv alapján teljesül az alábbi állítás:

$$\sum_{i=1}^{n_i} A_{j,i}(x_{j,k}) = 1, \forall\ j,k, \textbf{i.e.}\ a_{j,i} = c_{j,i-1}\ \textbf{and}\ d_{j,i} = b_{j,i+1}, \tag{2}$$

Gath-Geva csoportosítási algoritmus alkalmazása

Az $A_{j,i}$ ($x_{j,k}$) fuzzy halmazok reprezentálására a háromszög és a trapéz tagsági függvényeken kívül Gauss függvény is alkalmazható:

$$A_{j,i}(x_{j,k}) = \exp\left(-\frac{1}{2}\cdot\frac{(x_{j,k}-v_{j,i})^2}{\sigma_{j,i}^{\ 2}}\right), \tag{3}$$

ahol $v_{j,i}$ a Gauss függvény várható értékét, $\sigma_{j,i}^{\ 2}$ pedig a szórásnégyzetét (variancia) jelöli. A felügyelt Gath-Geva csoportosítási algoritmussal, az egyes attribútumokra vonatkozóan Gauss tagsági függvény formájában határozhatjuk meg a partíciókat:

$$A_{j,i}(x_j;a,b,c,d) = \max\left(0, \min\left(\frac{x_j - a}{b-a}, 1, \frac{d - x_j}{d-c}\right)\right), \quad (4)$$

A csoportosítási algoritmus által szolgáltatott Gauss tagsági függvényeket pedig, amennyiben háromszög tagsági függvényekre szeretnénk transzformálni, az alábbi összefüggéseket alkalmazhatjuk:

$$a_{j,i} = v_{j,i} - 3 \cdot \sigma_{j,i}, \quad b_{j,i} = c_{j,i} = v_{j,i} \quad d_{j,i} = v_{j,i} + 3 \cdot \sigma_{j,i} \quad (5)$$

A szélső tagsági függvényeket trapézosíthatjuk, azaz a $b_{j,1}$=0 $c_{j,i}$=1 paraméter értékeket alkalmazzuk (példa a *4. ábrán*).

4. ábra

Attribútum partícionálás Gath-Geva csoportosítással

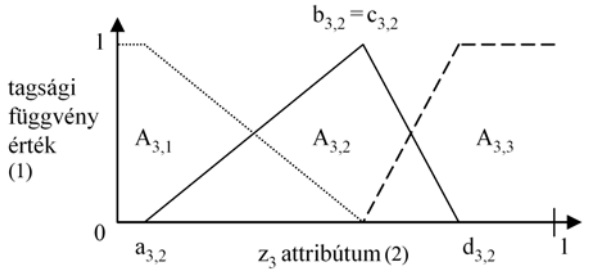

Figure 4. The Gath-Geva clustering based partitioning of an attribute

See Figure 2

C4.5 alapú felosztás alkalmazása

Az ID3 (Interactive Dichotomizer 3), vagyis az interaktív felosztó az egyik legelterjedtebb döntési fa előállítására használt algoritmus *Quinlan* (1986). Továbbfejlesztett verziója a C4.5 algoritmus, mely a döntési fa csomópontjaiban a minták felosztásához mindig a legnagyobb információnyereséggel járó vágást hajtja végre. Mivel az algoritmus mohó, vagyis lokálisan az optimumra törekszik, ezért előfordulhat, hogy az eredmény, azaz a végleges partícionálás globálisan nem lesz optimális.

A módszer nem feltétlenül használja fel az összes rendelkezésre álló attribútumot, ezért partícionálási szempontból az eredmény lehet, hogy nem teljes. Az *5. ábrán* az előző partícionálási módszerekkel szemben, négy attribútum (z_1-z_4) partícionálási eredményeit láthatjuk. Az első két attribútum esetében nem történt tényleges partícionálás, mert az algoritmus csak a harmadik és negyedik változókat vette figyelembe az osztályozás során. Ezeknél rendre az egy ($d_{3,1}$), illetve kettő éles vágás ($d_{4,1}$, $d_{4,2}$) alapján értelmezhető kettő ($A_{3,1}$, $A_{3,2}$), illetve három ($A_{4,1}$-$A_{4,3}$) partíciót láthatjuk. Ez ilyen eseteknél vagy elhagyjuk a C4.5 által az osztályozási szempontból feleslegesnek jelölt attribútumokat (amilyen z_1 és z_2) vagy, pedig más technikával partícionáljuk azokat. Mi az előbbi módszert alkalmaztuk, tehát csak a C4.5 által felhasznált attribútumokkal generáltunk asszociatív osztályozási szabályokat. A módszer és a többi partícionálási technika teljesítményeinek összevetését a következő fejezetben olvashatjuk.

5. ábra

Attribútum partícionálás a C4.5 algoritmussal

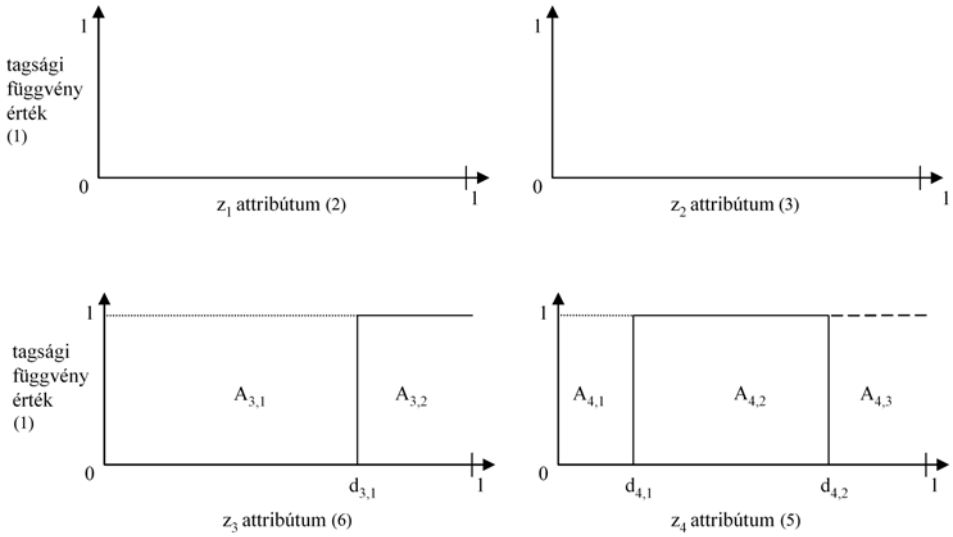

Figure 5. Partitioning of attributes by C4.5 method

Membership value(1), Attribute z_1(2), Attribute z_2(3), Attribute z_3(4), Attribute z_4(5)

ALKALMAZÁSI PÉLDÁK

A bemutatott partícionálási módszerek alkalmazási lehetőségét, illetve az osztályozási teljesítményére gyakorolt hatásukat két közismert osztályozási teszt adatsor esetén vizsgáltuk meg. Mindkettő osztályozási probléma (Iris, Wine) letölthető a *http://www.ics.uci.edu/~mlearn/MLRepository.html* címen. Az osztályozási teljesítményeket mindegyik osztályozási problémánál tízszeres keresztvalidálással határoztuk meg, vagyis az összes osztályozási mintát permutáltuk, majd a teljes halmazt tíz, azonos számú mintát tartalmazó részhalmazra osztottuk. Egy tanítási folyamatban a részhalmazok közül mindig kilencet használtunk fel a módszerek tanítására, és a kimaradó egy halmaz volt a teszthalmaz. Tíz esetben végeztük el a tanítást, így a teljesítményre kapott értékek a tíz teszt átlagolt eredményei. A pontosság mellett vizsgáltuk az egyes módszerek által generált szabálybázisok komplexitását, a szabályok és a feltételek számát. Az alkalmazott osztályozási módszerek a következők voltak:

Az első számú osztályozónál (Osztályozó 1.) a feltárt szabályok egy halmazával történik az osztályozás. Egy adatminta esetén a szabálybázis minden szabályára kiszámítható, hogy mennyire illeszkedik az adott mintára. Ezt a szabály *tüzelési erősségének* (*firing strength*) hívjuk, és a *k*-adik mintára az adott *j*-edik szabály fuzzy tagsági függvény értékeinek a minimumával képezzük:

$$\beta_j(\mathbf{x}_k) = \min(t_k(z_i)), \quad \langle z_i : c_{i,j} \rangle \in \langle Z : C \rangle \tag{6}$$

Legegyszerűbb esetben minden osztályra meghatározzuk a szabály bázis által adott pontszámokat, az összes szabály tüzelési erősségének és fuzzy konfidenciájának (FC_j) a szorzatával: $w_e = \sum_j (\beta_j \cdot FC_j)$, ahol az osztályok: $e = 1,\ldots,C$, majd a legtöbb pontszámmal rendelkező osztály lesz a becsült osztálycímke. Amennyiben osztályonként összevonjuk ($cover_e$) az illeszkedő szabályok tüzelési együtthatóját, és figyelembe vesszük az illeszkedő szabályok számát ($rules_e$), úgy ezek hányadosát súlytényezőként alkalmazhatjuk az osztályok pontozásánál:

$$w_e = \left(\sum_j (\beta_j \cdot FC_j) \right) \cdot \frac{\mathrm{cov}\, er_e}{rules_e} \quad \rightarrow \hat{y} = \arg\max\,(w_e) \tag{7}$$

A második osztályozónál (Osztályozó 2.) a feltárt szabályok közül mindig csak a legerősebb szabály határozza meg az adott minta osztálycímkéjét. Aktuálisan azt a szabályt tekintjük a legerősebbnek, amelynek tüzelési együtthatójának és fuzzy konfidenciájának a szorzata maximális értékű:

$$w_e = \max\,(\beta_j \cdot FC_j) \quad , j = 1,\ldots,M \rightarrow \hat{y} = \arg\max\,(w_e) \tag{8}$$

A kiválasztott teszt adatsorokkal megvizsgáltuk mindkettő osztályozó teljesítményét, hogyan befolyásolja azok osztályozási pontosságát az alkalmazott partícionáló módszer. Emellett a *2.* és a *3. táblázatban* feltüntettük az egyes technikák által generált szabálybázisok bonyolultságát is.

2. táblázat

Osztályozási teljesítmények az Iris adatsor esetén

Partícionálás (1)	**Osztályozó 1 (2)**	**Osztályozó 2 (3)**	**Szabályok (4)**	**Feltételek (5)**
Ruspini	94.67	94.67	5.8	10.9
GK	94.00	94	3.9	8.5
GG	**96.00**	**96.67**	**5.2**	**9.7**
C4.5	93.33	93.33	6.1	8.8

Table 2. Classification performances on the Iris dataset

Partitioning technique(1), Classifier 1(2), Classifier 2(3), Number of rules(4), Number of conditions(5)

3. táblázat

Osztályozási teljesítmények a Wine adatsor esetén

Partícionálás (1)	**Osztályozó 1 (2)**	**Osztályozó 2 (3)**	**Szabályok (4)**	**Feltételek (5)**
Ruspini	93.26	**94.34**	67.8	172.3
GK	88.20	91.56	48.2	124.9
GG	**93.30**	91.63	44.5	130.3
C4.5	92.02	92.02	**10.1**	**27.1**

Table 3. Classification performances on the Wine dataset

See Table 2

Az Iris adatsor esetében a Gath-Geva csoportosítási algoritmussal mindkettő osztályozási módszer esetén megfelelő pontosságú osztályozási eredményt kaptunk. A szabályok és a feltételek száma alapján megállapítható, hogy az alkalmazott módszerrel kompakt osztályozási modelleket sikerült létrehozni.

Ugyanez sajnos nem mondható el a Wine adatsor esetében, ahol kizárólag a C4.5 algoritmus által meghatározott partíciókat felhasználva sikerült kompakt méretű és megfelelő pontosságú asszociatív osztályozókat előállítani. Ennek egyik fő oka lehet, hogy a Wine adatsor egy 13 attribútumot tartalmazó osztályozási probléma, melyben a minták száma az attribútumok számához képest csekély (178). Tehát amennyiben az összes attribútumot felhasználjuk az asszociatív osztályozó létrehozásakor, úgy a szabálytisztító algoritmus (mely a ϕ korrelációs együttható alapján működik) csak kisebb hatékonysággal képes megtisztítani a szabálybázist.

6. ábra

Példa a Gath-Geva csoportosítási algoritmussal generált fuzzy asszociatív osztályozó szabálybázis megjelenítésére

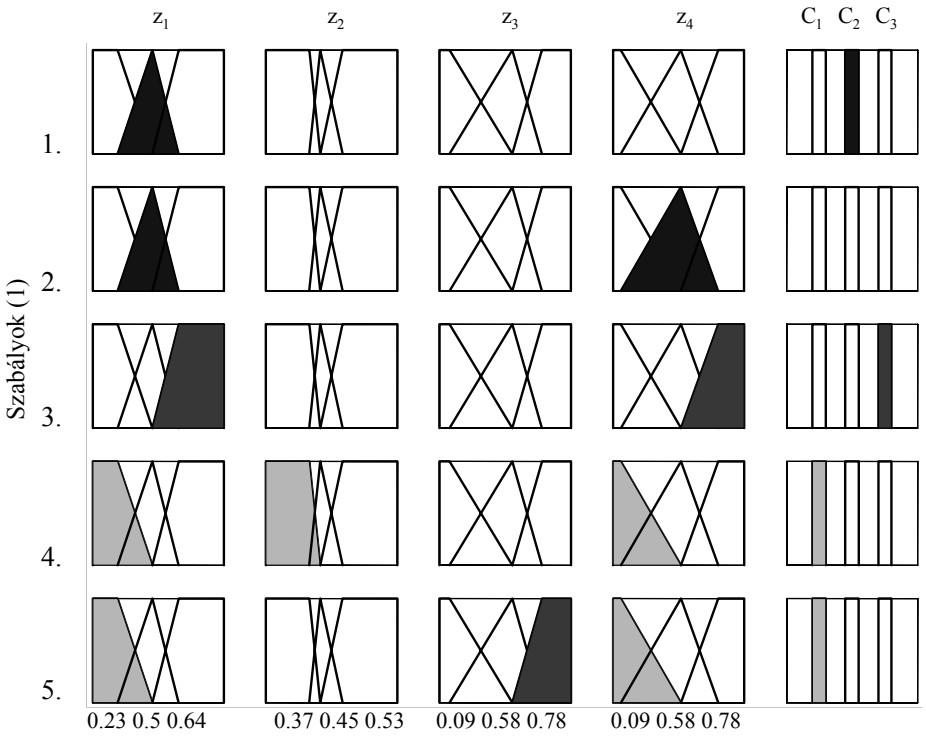

Figure 6. Example for visualization of a fuzzy associative classifiaction rulebase generated by the Gath-Geva based partitiong

Rules(1)

Következésképpen, az egyik ígéretes továbbfejlesztési irány (a hatékonyabb szabálybázis tisztítás mellett) az osztályozási szempontból felesleges attribútumok

kiszűrése lehet. Egy előzetes osztályozással, automatikusan, adatvezérelt módon, vagy akár vizualizációs eszközök segítségével, manuális úton is kiválaszthatóak a felesleges változók. A *6. ábrán* egy lehetséges szabálybázis ábrázolási technikát láthatunk, ahol az egyes sorok reprezentálják a szabályokat, az oszlopok, pedig az attribútumokat, illetve az egyes szabályok attribútumra vonatkoztatott tartalmát (a satírozott halmazok). Az utolsó oszlopban a szabályok kimenete látható. Jól látható, hogy az első szabály csak az első attribútumról hordoz információt, nevezetesen *„Ha z_1 attribútum értéke közepes, akkor C_2 osztály"*. Az attribútumoknál az egyes partíciók (fuzzy halmazok) osztópontjait is feltüntethetőek az ábrán (lásd 5. szabálynál), így a konkrét érték tartományok is jól láthatóak minden szabály esetén. Az ábrázolással egyértelműen megállapítható, hogy egy szabálybázisnál melyik attribútumokat használhatjuk fel az osztályozáshoz, melyik az osztályozási szempontból felesleges változó (amelyik egyik szabálynál sincsen besatírozva), illetve, hogy van-e ismétlődés az osztályokra vonatkozóan (több szabály is azonos kimenettel, pl. *6. ábrán* 1. és 2. illetve a 4. és 5. szabályok.) Az értelmezhetőségnek köszönhetően a kiválasztáson túl, a szabálybázisok szerkeszthetősége is lehetségessé válik. Így az elkészült algoritmusok, egy adott alkalmazási terület (például folyamatmérnökség) szakértőivel együttműködve hatékony (technológiai) döntéstámogatási szoftver eszköz alapjait is képezhetik.

KONKLÚZIÓ

Az ismertetett fuzzy asszociatív osztályozó algoritmus egyik fő előnye, hogy az osztályozó szabályok előállításához nem igényli a tanítási minták előzetes lefedettség vizsgálatát, amely nagy számú minta esetén (mint sok asszociatív osztályozó algoritmusnál) jelentős számítási teljesítményt követelne. A cikkben tárgyalt négy partícionáló módszer közül ígéretes eredményeket sikerült elérnünk a felügyelt Gath-Geva csoportosítási algoritmus alkalmazásával az Iris adatsor esetén. A felügyelt csoportosítási algoritmus kiegészítéseként egy attribútum kiválasztó eszközzel, egyrészt csökkenthető lenne a gyakori elemhalmazok (mely az asszociatív osztályozók szűk keresztmetszete), illetve a szabályok keresési ideje, másrészt a kompaktabb szabálybázissal, növelhető lenne az osztályozási teljesítmény is (ahogy arra a C4.5 algoritmus partícionálási módszere a Wine osztályozási problémánál rávilágított). Az eredmények és a fejlesztési lehetőségek figyelembe vételével, a jövőben további, részletesebb, több adatsorra kiterjedő vizsgálatot folytatunk, majd az eredményeket hazai és nemzetközi fórumokon is publikáljuk.

KÖSZÖNETNYILVÁNÍTÁS

A szerzők ezúton szeretnék kifejezni köszönetüket a Vegyészmérnöki Intézet Koordinációs Kutatási Központjának (VIKKK-III/1 projekt) és az OTKA-nak (T 049534) a támogatásért.

IRODALOM

Abonyi, J. (2006). Adatbányászat a hatékonyság eszköze, Gyakorlati útmutató kezdőknek és haladóknak, Computerbooks

Agrawal, R., Srikant R. (1994). Fast algorithm for mining association rules in large databases. In: Proceedings of The 20th International Conference on Very Large Data Bases, Santiago, Chile, 487–499.

Agrawal, R., Imielinski T., Swami, A. (1993). Database mining: A performance perspective. IEEE Transactions on Knowledge and Data Engineering, 5. 6. 914–925.

Catlett, J. (1991). Megainduction: machine learning on very large databases, PhD thesis, University of Sydney

Dong, G., Zhang, X., Wong, L., Li, J. (1999). CAEP: classification by aggregating emerging patterns. In: Proceedings of The Second International Conference on Discovery Science (DS '99), Tokyo, Japan, 30-42.

Dougherty, J., Kohavi R., Sahami, M. (1995). Supervised and Unsupervised Discretization of Continuous Features, In: Proceedings of The Twelfth International Conference on Machine Learning, Tahoe City, CA, USA, 194-202

Gath, I., Geva A.B., (1989). Unsupervised optimal fuzzy clustering, IEEE Transactions on Pattern Analysis and Machine Intelligence, 11 (7), 773-780.

Gustafson, D.E., Kessel, W.C. (1979). Fuzzy clustering with fuzzy covariance matrix, In: Proceedings of The IEEE Conference on Decision and Control, San Diego, CA, 761-766.

Liu, B., Hsu, W., Ma, Y. (1998). Integrating classification and association rule mining. In:: Proceedings of The Fourth International Conference on Knowledge Discovery and Data Mining (KDD '98), New York City, USA, 80-86.

Liu, B., Ma, Y., Wong C. K. (2000). Improving an Association Rule Based Classifier, In: Proceedings of the 4th European Conference on Principles of Data Mining and Knowledge Discovery, (PKDD 2000), Lyon, France, 504-509.

Meretakis, D., Wuthrich, B. (1999). Extending Naive Bayes Classifiers Using Long Itemsets, In: Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining (KDD '99), San Diego, USA, 165-174.

Pach, F.P., Gyenesei, A., Németh, S., Árva, P., Abonyi, J. (2006). Fuzzy Association Rule Mining for the Analysis of Historical Process Data, Acta Agraria Kaposváriensis, szerkesztés alatt

Pach, F.P., Gyenesei, A., Németh, S., Árva, P., Abonyi, J. (2006). Fuzzy Association Rule Mining for Model Structure Identification, Applications of Soft Computing: Recent Trends, Part VI Identification and Forecasting, Springer, 261-271.

Quinlan, J.R., (1986). Induction on decision trees. Machine Learning, 1. 1. 81-106.

Wang, K., Zhou, S., He, Y. (2000). Growing decision tree on support-less association rules. In: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD'00), Boston, MA, USA, 265-269.

Yin, X., Han, J., (2003). CPAR: Classification based on predictive association rules, in Proceedings of the Third SIAM International Conference on Data Mining (SDM'03), San Francisco, CA, USA

Zimmermann, A., Raedt L. D. (2004). CorClass: Correlated Association Rule Mining for Classification, Discovery Science, 7th International Conference, Padova, Italy, 60-72.

Levelezési cím (*Corresponding author*):

János Abonyi
University of Pannónia, Department of Process Engineering
H-8201, Veszprém, P.O. Box 158
Pannon Egyetem, Folyamatmérnöki Tanszék
8201, Veszprém, Pf. 158.
Tel.: 36-88-624-447, Fax: +36-88-624-171
e-mail: abonyij@fmt.uni-pannon.hu