FOGALOMHÁLÓ ALAPÚ OSZTÁLYOZÁSI MÓDSZEREK

CLASSIFICATION METHOD BASED ON CONCEPT LATTICE ARCHITECTURE

*Kovács László**

ABSTRACT

The classification is a central task in data mining and knowledge engineering. There is a huge literature on the classic methods of classification. The paper presents an approach on development of a novel algorithm using concept lattice structure. In the classification based on Formal Concept Analysis (FCA), the class label is treated as special category attribute. The FCA provides an efficient mechanism to perform generalization of the attribute structure. The decision tree generated from the lattice provides better results than the traditional methods for a test grammar induction problem.

1. OSZTÁLYOZÁS FELADATKÖRE

Az számítógépi intelligencia problémaköreinek egyik fő eleme az osztályozás feladata, melynek során az objektumokat előre definiált osztálykódokhoz rendeljük. Az osztályozás feladatánál feltesszük, hogy az objektumok egy magadott tulajdonsághalmazzal jellemezhetőek és az objektum osztálykódja az objektum tulajdonságaitól függ. Az osztályozás célja az ismeretlen kapcsolat leíró f függvény mellett egy olyan

$$g : O \to 2^C$$

osztályozó függvény meghatározása, melyre

$$E(f, g, S) \to \min$$

teljesül, ahol az E(.,.) függvény a hibafüggvényt, S a tanítóhalmazt és C az osztálykódok halmazát jelöli. Az így kapott g függvény felhasználható az f függvény közelítésére is a teljes O halmazon. A hibát egy adott objektumnál valós értékű osztályozó függvények mellett rendszerint az eltérésnégyzettel mérjük:

$$E(f, g) = \sum_{o \in S} (f(o) - g(o))^2 \ .$$

Az osztályozó függvény statisztikai alapú meghatározása egy felügyelt tanítással történik. A felügyelt tanítás során a tanítóhalmaz

$$T = \{(o, f(o)) \mid o \in S\}$$

alakú, tehát a mintában minden objektumhoz ismertek az odatartozó kategória értékek. Az ismeretlen g osztályozó függvény meghatározására alapvetően kétféle módszer terjedt el:

• particionáló módszer, melyben az objektumok vektorterét különböző kategóriákhoz rendeljük;
• osztályvalószínűségi függvény alapú, amikor egy P(c|o) valószínűségi értéket határozzunk meg minden érintett kategóriára.

Az irodalomban az osztályozás területe egy aktívan vizsgált terület, igen sok osztályozási algoritmus változatot dolgoztak ki. A teljesség igénye nélkül megemlítjük az elterjedtebb algoritmusokat. Az egyik csoportba tartoznak a döntési fa alapú algoritmusok, mint az ID3 [1] és a C4.5. A szabályalapú működést megvalósító módszerek közül kiemelhetőek a CN2, CL2 [2] módszerek. A tiszta statisztikai elveken működő eljárásokhoz tartozik a CART, a Bayes osztályozók [3] és a genetikus algoritmus alapú változatok. A nemlineáris modellek közül kiemelhető a neurális háló alapú modell [4] és a nemlineáris regressziós eljárás. A kvadratikus programozás alapú osztályozók közé tartozik az SVM [5] módszer. Az induktív logika eszközrendszerével dolgozó osztályozók közé tartozik a [6]-ben javasolt rendszer.

2. FOGALOMHÁLÓ ARCHITEKTÚRÁJA

A fogalomhálók módszere a formális fogalom analízis (FCA: formal concept analysis)[7] elméletéhez kapcsolódik. A rendszer kiindulása egy objektumhalmaz, ahol mindegyik objektum egy megadott tulajdonsághalmazból vehet fel jellemzőket. Ezen halmazra vonatkozólag fogalom alatt objektumok és tulajdonságok halmazainak olyan párosát értjük, melyre teljesül, hogy az objektumok mindegyike rendelkezik a megadott tulajdonsággal,

**egyetemi docens, Miskolci Egyetem Általános Informatikai Tanszék*

és a halmazon kívüli objektumoknál a megadott tulajdonsághalmaz együtt már nem fordul elő; a tulajdonsághalmaz tartalmazza az objektumok minden közös tulajdonságát. Ha a tulajdonsághalmaz szerinti jelölését vesszük, akkor az egyes fogalmak között a leíró tulajdonsághalmaz szerinti tartalmazási relációt kivetíthetjük a kapcsolódó fogalmak közötti tartalmazási relációra. Ezáltal egy részben rendezést kapunk a fogalmakra, s az is belátható, hogy a kapott struktúra egy háló. A kapott hálót nevezik fogalomhálónak.

Kontextus alatt egy K(G,M,I) hármast értünk, ahol G az objektumok halmaza, M a tulajdonságok halmaza és I egy reláció G és M vonatkozásában. Az $A \subseteq G$ halmazokra egy deriválási operátor értelmezett, melynek jelentése:

$$A' = \{a \in M | \forall g \in A: gIa\}$$

és ennek analógiájára a tulajdonság halmazokra is felveszünk egy hasonló operátort:

$$B' = \{g \in G | \forall a \in B: gIa\}.$$

A deriválási operátorra igaz, hogy minden A_i-re

$$(\cup_i A_i)' = \cap_i A'_i$$

illetve minden B_i-re

$$(\cup_i B_i)' = \cap_i B'_i$$

teljesül. A C(A,B) párost fogalomnak nevezik a K kontextusra vonatkozólag, ha

$$A \subseteq G;\ B \subseteq M;\ A' = B;\ B' = A$$

teljesül. A fogalomnak az objektum részét nevezik extent-nek, a tulajdonság részét pedig intent-nek. A K(G,M,I)-re vonatkozó fogalmak halmazát jelöljük Φ-vel. Ekkor a Φ elemei között egy rendezési reláció értelmezhető a következő módon:

$$C_1 \leq C_2 \Leftrightarrow A_1 \subseteq A_2,$$

ahol C1 és C2 tetszőleges fogalmak. Belátható, hogy minden (C1,C2) fogalompárosra teljesül, hogy:

$$C_1 \wedge C_2 \in \Phi, C_1 \vee C_2 \in \Phi.$$

Ez alapján a $\Lambda(\Phi,\leq)$ páros hálónak (lattice) tekinthető, melyet fogalomhálónak neveznek. A háló felépítésének algoritmusát elemzi többek között a [10],[12] publikációk.

3. FCA ALAPÚ OSZTÁLYOZÁS

Az osztályozó algoritmusok célja osztálykategóriák objektumokhoz való rendelése az objektumok tulajdonságai alapján. A fogalomháló alkalmas eszköz arra, hogy az összes lényeges, zárt attribútum csoportra megvizsgáljuk az osztályhovatartozás jellegét [11]. A fogalomhálók osztályozási feladatokra történő alkalmazására tett első javaslatok között kiemelhető Zhao [8] dolgozata. A felvázolt modellben az objektum attribútumok közül ez egyiket osztálykategóriaként kezeljük. A klasszifikációs szabály leírja a kategória címkének a logikai formuláktól való függőségének mikéntjét.

A Zhao által kidolgozott modellben egy adott K(G,M, I) kontextus esetén, a klasszifikációs szabály $f \Rightarrow c$ alakú. A szabályban az f jelöli a logikai formulákat, míg a c szimbólum az a_c kategória attribútum egy lehetséges értékét, tehát egy osztálykategóriát jelöl. A megadott szabály jósága a konfidencia és az általánosíthatóság mérőszámaival határozható meg, ahol

$$\text{conf}(f \Longrightarrow c) = \frac{|m(f \cap (a_c = c))|}{|m(f)|};$$

$$\text{generality}(f \Longrightarrow c) = \frac{|m(f)|}{|O_K|}.$$

Minél nagyobb a konfidencia érték annál pontosabb az osztályozás eredménye. A konzisztens osztályozási szabály olyan osztályozási szabály, ahol a konfidencia értéke 1, azaz

$$|m(f \cap c)| = |m(f)|$$

teljesül. Az m(f) szimbólum azon objektumok halmazát jelöli, melyek kielégítik az f predikátumot. A (X, f) konjunktív fogalmat konzisztens fogalomnak nevezik, ha az egy egyedi kategória címkét indukál és konfidencia értéke 1. A legáltalánosabb konzisztens fogalom olyan konzisztens fogalom, ahol a szülő fogalmak egyike sem konzisztens fogalom. Be lehet bizonyítani, hogy a legáltalánosabb konzisztens fogalmak az univerzumok lefedését alkotják. Emiatt a legáltalánosabb konzisztens fogalmak elegendőek az osztályozási feladatok elvégzésére.

A halmaz felépítésének brute-force megközelítésében, az összes fogalom hálóját építik el először. Ezt követően a konzisztens illetve a legáltalánosabban konzisztens fogalmak válogatjuk ki.. Az algoritmus az alábbi lépéseket tartalmazza:

1. A konzisztens klasszifikációs szabályok generálása az alábbi algoritmus alapján:
   (a) ciklus az összes kategória címkére (c).
   (b) a (c|φ) valószínűség értékének meghatározása.
   (c) a legvalószínűbb kategória kiválasztása.
2. Konzisztens definíciós pár generálása minden szabályhoz.
3. A konjunktív konzisztens fogalmak halmazának előállítása.
4. Azon konjunktív konzisztens fogalmak eliminálása, melyek nem a legáltalánosabb fogalmak közé tartoznak.

A tapasztalatok [9] azt mutatják, hogy a fogalmi hálón alapuló osztályozási módszerek jobb osztályozási pontosságot tudnak elérni, mint az ID3 alapú döntési fák. A háló alapú megközelítés előnye, hogy a formulák minden lehetséges zárt csoportja feldolgozásra kerül a legjobb kategória korrelációs értékkel bíró formulák meghatározásakor. Hátránya, hogy a háló felépítése nagyon költséges folyamat.

A fenti elképzelésből kiindulva egy olyan hibrid modell került kidolgozásra hoztam létre, amelyben a fogalomhálóból egy döntési fa generálódik, s az így kapott fa közvetlenül felhasználható az objektumoknak az attribútumaik szerinti osztályozására.

Jelölje Λ a megadott K(G,M,I) kontextushoz tartozó fogalomhálót. Bővítsük az attribútumok körét a kategória attribútummal. Az így kapott fogalomhálót jelöljük Λ*-vel. Adott Λ=({Ai,Bi},≤) hálóhoz a kibővített Λ*= ({Ai,Bi},≤,c,d,s) fogalomháló tartalmaz egy

$s: T \rightarrow N^+$ támogatottsági értéket;

$c: T \rightarrow C$ kategória cimkét;

$d: T \rightarrow C$ default kategória cimkét.

A T szimbólum az Λ háló fogalmainak halmazát jelöli, a támogatottság a fogalomtól kisebb (≤ reláció) dominált fogalmak számossága. A nem atomi fogalmaknál a c értéke a dominált fogalmak kategóriáinak metszete. A default kategória értéke egyenlő azon nem üres kategóriával, amely a dominált fogalmak között a legnagyobb támogatottsággal rendelkezik. Az osztályozásnál a maximálisan konzisztens fogalmak attribútum elemei alapján határozhatjuk meg a bejövő attribútum halmazhoz rendelhető osztálykategóriát.

Az Λ* hálóban a következő szabályok érvényesülnek:

- Ha egy adott c kategóriához csak egyetlen χ maximálisan konzisztens fogalom létezik, melynek intenzió része χ_a , akkor

$$\chi_a \Leftrightarrow c$$

teljesül, azaz c akkor és csak akkor érvényes, ha χa érvényes. Így a kapcsolódó döntési fában egyetlen csomópont elegendő ezen kategóriához.

- Ha egy adott c kategóriához a χ egy konzisztens fogalom, melynek intenzió része χa, akkor

$$\chi_a \Rightarrow c$$

teljesül. A kapcsolódó döntési fában a χa –hez egy olyan csomópont fog tartozni, melynek egyetlen homogén c-hez kapcsolódó leszármazottja van.

A generált döntési fa előnye, hogy könnyen érthető a külső szakemberek által is és jól követi az emberi gondolkodásmódot.

A döntési fa generáló algoritmus további sajátossága, hogy azon csomópontok, melyek nem konzisztensek, a dominált csomópontok alapján legvalószínűbb kategória kódot kapják meg jellemzőként. A generált fa hatékonyabb kezelésére az elkészült fában egyes elemek eliminálásra kerülnek. A redukció azon gyerekeket eliminálja, melyek osztálykategória jelzője megegyezik a szülő kategóriájával.

4. KISÉRLETI EREDMÉNYEK

A kidolgozott módszer tesztelésére egy, a szavak ragozását végző feladat került kiválasztásra. A szavak ragozása egy osztályozási feladatként is értelmezhető, ahol a rag megfelel az osztálykódnak. Az induló kísérlet csak minimális, pár százas méretű tanító halmazzal dolgozott. A kísérletben a FCA alapú módszer mellett FSA és HMM alapú módszerek kerültek implementálásra. A mintarendszer fejlesztése Java nyelven az Oracle JDeveloper eszköz segítségével történt. A betanítás sikerességét az U és T kontroll halmazokkal ellenőrizzük. A kis minta esetén a kontroll halmazok is kisebb méretűek. A T halmazban a már betanított szavak szerepelnek, míg az U halmaz a még ismeretlen szavakat tartalmazza. A minta U halmaz a következő szavakat fogta össze:

U = {gatya, labda, tanár, krumpli, fej, korong, csapat,köd, ló, korom}

A ragozott alakok ekkor a következők:

R = {gatyát, labdát, tanárt, krumplit, fejet, korongot, csapatot, ködöt, lovat, kormot)

Az FST módszernél igen rossz eredmény született az U halmaz vizsgálatánál, hiszen ezekre még nem kapott példát a modell:

R’ = {g, l, tanítót, k, fecskét, k, csattot, kö,

lócát, k}

A HMM esetében is hasonló rossz eredmények születtek a kis tanító halmaz esetében. A saját háló alapú módszernél viszont már előfordultak jó becslések is:

R' = {gatyát, labdát, tanárt, krumplit, fejt, korongot, csapatot, ködt, lót, koromet}

Látható, hogy itt 60%-os pontosságot generált a modell. Az is észrevehető, hogy a tévesen ragozott alakok halmaza is sokkal közelebb áll a humán gondolkodáshoz, mint a másik két módszernél kapott eredménylista.

1. ábra. Találati pontosságok összehasonlítása

A pontosság mellet a háló alapú módszer másik előnye, hogy viszonylag gyorsan tud döntést hozni. A transzformált alak előállítása itt lényegesen hatékonyabb, mint a HMM esetében, ahol dinamikus algoritmussal kell meghatározni a legkisebb összköltségű utat.

5. ÖSSZEFOGLALÁS

A tanulmány a fogalomháló alapú osztályozási rendszer működési elvét mutatja be kiegészítve annak döntési fával kombinált változatával. A kidolgozott módszer előnye az általánosítási operátor erőteljessége. A mintarendszerben a módszer hatékony osztályozási eszköznek bizonyult.

6. KÖSZÖNETNYÍLVÁNÍTÁS

A bemutatott kutató munka a TÁMOP-4.2.1.B-10/2/KONV-2010-0001 jelű projekt részeként az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósul meg.

7. IRODALOM

[1] QUINLAN J. .: Induction of decision trees. Machine Learning, 1:85–106, 1986

[2] CLARK P, NIBLETT T.: The CN2 induction algorithm. Machine Learning;3(4): 261–83., 1989

[3] FRIEDMAN N, GEIGER D, GOLDSMIDT M.: Bayesian network classifiers. Machine Learning; 29 (2):131–63, 1997

[4] RUMELHART DE, HINTON GE, WILLIAMS RJ.: Learning internal representations by error propagation. Parallel distributed processing: explorations in the micro-structure of cognition. MIT Press; pp. 318–63., 1986

[5] VAPNIK V.N.: The nature of statistical learning theory. New York: Springer; 2000

[6] DZEROSKI S.: Inductive logic programming and knowledge discovery in databases. Advances in knowledge discovery and data mining. p. 117–52., 1996

[7] GANTER B. AND WILLE R.: Formal Concept Analysis, Mathematical Foundations, Springer Verlag, 1999.

[8] ZHAO Y. –YAO Y.: Classification based on logical concept analysis, Proceedings of 19th Conference on the Canadian Society for Computational Studies of Intelligence (Canadian AI 2006), Québec City, Québec, Canada, June 7-9, pp. 419-430, 2006

[9] FERRE, S., RIDOUX, O.: A logical generalization of formal concept analysis, 2001

[10] KOVACS L.: Efficiency Analysis of Building Concept Lattice, Proceedings of 2nd ISHR on Computational Intelligence, Budapest, 2001

[11] KOVACS L. , BARANYI P.: Document Clustering based on Concept Lattice, Proc. of IEEE SMC, Tunisia, pp 67-75, 2002

[12] VALTCHEV. P., MISSAOUI.,R. : Building concept (Galois) lattices from parts: generalizing the incremental methods., Proc. of ICCS01, pp 290-303. 2001