Numerikus adatterek hasonlósági elemzése

NAGY ISTVÁN (analog@logana.com, www.logana.com)

Összefoglalás

E tanulmány bemutatja a numerikus adatok hasonlósági elemzésének alapelvét, módszertanának alapjait és alkalmazását.

Az alkalmazás során a cél a numerikus adatterek reprezentatív objektumainak, és az ezekhez "elegendő" mértékben hasonló (vagyis a reprezentált) objektumoknak a kijelölése, továbbá az ezeket "elegendő" mértékben jellemző attribútumok, és azok "megfelelő" értéktartományának meghatározása, és mindez oly módon, hogy a reprezentált objektumok összessége "elegendő" mértékben fedje le az eredeti adatteret.

Kulcsszavak

Numerikus adattér, numerikus transzformáció, hasonlósági tér, hasonlósági műveletek, hasonlósági transzformáció, határhasonlóság, hasonlósági térredukció, hasonlósági osztályozás, attribútum-redukció.

TARTALOM

BEVEZETŐ

A növekvő mennyiségű, és egyre szélesebb körből gyűjtött adatok feldolgozása és elemzése új szemléletet igényel. A hasonlósági (asszociatív) elemzés alapgondolata abból a felismerésből indul ki, mely szerint a vizsgált objektumok diszjunkt osztályozása a háttérben lévő folyamatok megismerésének a legfontosabb módszertani akadálya. Ha azonban megengedjük az olyan osztályképzést, melynek eredményeként egyes objektumok akár több osztály elemei is lehetnek (ezt fogjuk hasonlósági osztályozásnak nevezni), akkor lehetőségünk lesz felfigyelni olyan folyamatokra, melyek az általunk már felismertek hátterében azok hatását módosítják.

Az alapfeladatunk a "helyettesíthetőség" vizsgálata lesz, vagyis az, hogy egy kapcsolatrendszerbeli objektum milyen feltételekkel helyettesíthető egy másikkal oly módon, hogy a kapcsolatrendszerben elfoglalt helye egy megadott "tűréshatáron" belül maradjon.

Tekintsük például a kirándulóbusz-feladatot, mely szerint a cél meghatározni egy kisvárosban azokat a helyeket, ahol célszerű megállnia a kirándulókat összeszedő busznak azzal a feltétellel, hogy az embereknek az egyes gyűjtőhelyekre ne kelljen 10 percnél többet sétálniuk. Nyil-

ván megállhat minden egyes kiránduló házánál (így eljutunk az iskolabusz-feladathoz, ahol az energia-idő takarékosság jegyében már az a fő kérdés, hogy milyen sorrendben szedje össze a diákokat), de a 10 perc ~ 1 km feltételezéssel a kirándulók lakcímeinek és a város térképének ismeretében kialakíthatunk célszerű gyűjtőpontokat. Ekkor azt mondhatjuk, hogy az egyes kirándulók otthonainak halmazát (nevezzük ezt vizsgált halmaznak) helyettesíthetjük a kisebb elemszámú gyűjtőhalmazzal.

Ebben az esetben még az is érdekes, hogy a gyűjtőhalmazban jórészt olyan elemek lesznek, melyek a vizsgált halmaznak nem is elemei. Persze a feladat megfogalmazható oly módon is, hogy a gyűjtőhalmaz mindenképpen részhalmaza legyen a vizsgált halmaznak.

Az alapfeladat tehát a gyűjtőhalmaz kijelölése, ám adott gyűjtőhalmaz esetén egy másik érdekes feladat annak meghatározása, hogy egy adott kiránduló otthona mely gyűjtőponttal (illetve gyűjtőpontokkal) helyettesíthető, feltételezve az 1 km-nél nem nagyobb odasétálási távolságot (a határtávolságot)...

A fentiek alapján az objektumok hasonlósági elemzésére az általános módszer a következő.

1.) *Az objektumtér létrehozása*

Ennek során az első feladat a vizsgált objektumoknak, és a vizsgálati céloknak a meghatározása, majd ezek alapján a vizsgált objektumok (adott vizsgálat szempontjából fontos) tulajdonságainak, az úgynevezett attribútumoknak a kijelölése, végül az attribútumértékek begyűjtése (mérésből, vagy már létező különböző adatforrásokból), ezáltal létrehozva az objektumteret (az "őstáblát").

2.) *A numerikus adattér létrehozása*

A hasonlósági elemzés alapvető környezete a vizsgált objektumok hasonlósági tere (lásd alább), ennek létrehozásához a legegyszerűbb út, ha az (általában) nem csupán numerikus adatokat tartalmazó objektumtérből létrehozunk egy kizárólag numerikus adatokat tartalmazó adatteret az úgynevezett *numerikus transzformáció* segítségével.

3.) *A hasonlósági tér létrehozása*

A numerikus adattérben az objektumok között különböző *hasonlósági függvények* (numerikus hasonlóságok) értelmezhetőek az attribútumok jellege, valamint a feldolgozás szempontjai szerint. Válasszuk ki az adott vizsgálathoz célszerű hasonlósági függvényt, majd ennek segítségével transzformáljuk az adatteret hasonlósági térré (*hasonlósági transzformáció*). Jelöljünk ki továbbá egy úgynevezett *határhasonlóság* értéket, melynek a további feldolgozás szempontjából meghatározó jelentősége lesz. (A hasonlóság algebrai értelmezésére ekkor úgy tudunk áttérni, hogy két objektumot akkor tekintünk hasonlónak, ha a rájuk vonatkozó hasonlósági függvény értéke eléri, vagy meghaladja az adott határhasonlóság értékét.)

4.) *Hasonlósági térredukció*

A hasonlósági térredukció a hasonlósági tér olyan, minél kisebb számú, úgynevezett *reprezentatív* objektumainak kijelölése, melyek a hozzájuk "eléggé hasonló" (legalább határhasonlóságú) objektumokkal együtt a teljes teret (vagy legalábbis annak "elegendően nagy" részét) lefedik.

5.) *Hasonlósági osztályozás*

A hasonlósági osztályozás során a redukált hasonlósági tér minden eleméhez kijelöljük az eredeti hasonlósági tér összes, hozzá "eléggé hasonló" (legalább határhasonlóságú) elemét. Az így létrehozott, úgynevezett hasonlósági osztályok fontos tulajdonsága, hogy részben átfedőek lehetnek.

Már itt jelezzük, hogy a gyakorlatban az attribútumok megfelelő csoportosításával, vagy esetleg a meglévők alapján új attribútumok származtatásával esetenként többdimenziós

hasonlósági teret, illetve még egy dimenzió mentén is több hasonlósági osztályozást is hozhatunk létre (lásd Többdimenziós hasonlósági terek).

6.) *Attribútum-redukció*

Az attribútum-redukció során kijelöljük azon attribútumokat, és ezek azon értéktartományait, melyek "megfelelő" mértékben jellemzik az egyes hasonlósági osztályokat.

7.) *Hasonlósági tér bővítése, hasonlósági osztályba-sorolás*

A gyakorlatban az objektumok hasonlósági tere általában folyamatosan bővül, ám egy nagyméretű hasonlósági térben már az új objektumok beillesztése is igen nagy számítási kapacitást igényel, a legnagyobbat azonban mindenképpen a bővített tér újratervezése (az újraosztályozás). Igen fontos feladat ezért ezek csökkentése természetesen ügyelve a kvázi-optimális megoldás kézbentartására.

A hasonlósági elemzés célja az ilyen típusú feladatok értelmezése, kitűzése és célszerű megoldása.

Megjegyezzük, hogy egy adott alkalmazásban talán megkérdőjelezhető egy konkrét hasonlósági transzformáció, a határhasonlóság, és más, úgynevezett minősítő paraméterek értékeinek *önkényes* kijelölése, ám éppen ez ad új lehetőséget az elemzési szempontoknak a korábbi elemzési tapasztalatok alapján történő iteratív érvényesítésére, új összefüggések, folyamatok felismerésére, az adatok hatékony kiértékelésére és redukált tárolására.

... Merre induljanak tehát az öregek? Nyilván a legközelebbi gyűjtőponthoz. No, de mit tegyenek azok, akik kettő, vagy több gyűjtőponttól is hasonló távolságra vannak? Nos, hát az ő esetük a legérdekesebb a számunkra. Ők ugyanis már egyéb szempontokat is figyelembe vehetnek a távolságon kívül. Mehetnek például a kedvenc boltjuk felé, vagy a lejtősebb úton.

Tulajdonképpen ez az egész tanulmány róluk szól. Azokról, akik több gyűjtőhalmazhoz, vagyis több hasonlósági osztályhoz tartoznak.

Megjegyzések a szakirodalomhoz

Már a fenti áttekintés alapján is megállapítható, hogy az objektumok hasonlósági vizsgálata (osztályozása, redukálása, elemzése) egy fontos módszer. A szakirodalom a hasonlóság fogalmát (akár fizikai, akár nyelvészeti, akár genetikai hasonlóságnak is nevezi), szinte kizárólag geometriai (vagy arra visszavezethető) értelemben használja (azaz algebrailag elvárja a tranzitivitást is), pedig ez a hasonlóság, algebrailag csupán ekvivalencia. Mi az általános hasonlósággal foglalkozunk (mely természetesen speciális esetben lehet akár ekvivalencia is).

1. NUMERIKUS ÉS HASONLÓSÁGI ADATTÉR

1.1. Numerikus adattér

Legyen $E = \{e_1,...,e_m\}$ az *objektumok halmaza*, $A = \{A_1,...,A_n\}$ pedig az *attribútumok halmaza*, továbbá z egy kétváltozós numerikus függvény az E és az A halmazok felett, azaz

$$z : E \times A \rightarrow \mathbf{R},$$

ahol **R** a racionális számok halmaza.

Ekkor a

$$Z = \langle E, A, z \rangle$$

struktúrát *m-elemű n-dimenziós numerikus adattérnek*, rövidebben *adattérnek*, vagy egyszerűen csak *térnek*, a z függvényt pedig a Z tér *adatfüggvényének* nevezzük. Ennek valamely $e_i \in E$ objektumhoz, és $A_k \in A$ attribútumhoz tartozó értékét

$$z(e_i, A_k)$$

módon jelöljük, tehát $z(e_i, A_k) \in A_k$

Bevezetjük az *Obj objektum-függvényt*

$$Obj(z) = E,$$

valamint az *Attr attribútum-függvényt*

$$Attr(z) = A$$

módon, melyek jelentését kiterjesztjük

$$Obj(Z) = Obj(z), \text{ és}$$

$$Attr(Z) = Attr(z)$$

módon, továbbá a Z tér z adatfüggvényére bevezetjük az alábbi jelölést:

$$Dat(Z) = z.$$

Egy Z numerikus adatteret célszerű egy olyan (azonos módon jelölt) *Z adatmátrixként* (adattáblaként) ábrázolni, melynek az $e_i \in Obj(Z)$ objektumhoz tartozó sorában az $A_k \in Attr(Z)$ attribútumhoz tartozó úgynevezett *objektumelem*:

$$Z[e_i, A_k] = z(e_i, A_k).$$

Ennek megfelelően a Z adatmátrix e_i objektumhoz tartozó *sora*, vagyis *objektum-vektora*:

$$Z[e_i, A] = \langle z(e_i, A_1),..., z(e_i, A_n) \rangle,$$

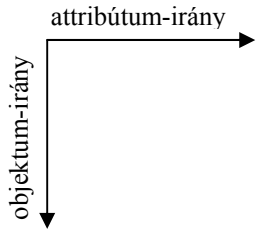
és az A_k attribútumhoz tartozó *oszlopa*, vagyis *attribútum-vektora*:

$$Z[E, A_k] = \langle z(e_1, A_k),..., z(e_m, A_k) \rangle,$$

ahol $n = \mu(Attr(z))$, $m = \mu(Obj(z))$, és μ a halmazszámosság jele.

Ezt a mátrixot esetenként a sorai (az objektumai) szerint, esetenként pedig az oszlopai (az attribútumai) szerint fogjuk feldolgozni, és ennek alapján beszélünk majd objektum-irányú, illetve attribútum-irányú adatfeldolgozásról. Mindezt szemléletesen az alábbi ábra mutatja be:

Z	A_1	...	A_k	...	A_n
e_1	...	...	...	...	...
...	...	...	...	...	...
e_i	...	...	$Z[e_i, A_k]$	...	...
...	...	...	...	...	...
e_m	...	...	...	...	...



A Z mátrix *objektum-irányú feldolgozásához* az objektum-vektorokat, az *attribútum-irányú feldolgozásához* pedig az attribútum-vektorokat fogjuk használni.

1.2. Numerikus adattér létrehozása

Az objektumtér feldolgozásának első lépése a numerikus adattér létrehozása.

A numerikus adatteret olymódon hozzuk létre az objektumtérből, hogy az objektumteret a nem-numerikus attribútumai mentén numerikussá transzformáljuk (*numerikus transzformáció*). Ez egyszerűen elvégezhető, ha a nem-numerikus attribútumokon a véges számú attribútum-értékek valamilyen alkalmazói prioritás szerint szekvenciálisan rendezhetők, és e rendezett sor egyik végét *preferált értéknek* (vagyis az adott alkalmazás szempontjából kitüntetettnek) tekinthetjük. Ekkor 0-tól kezdve (ezt a legkisebb prioritású attribútum-értékhez rendelve) növekvően hozzájuk rendelhetjük a nemnegatív egész számokat.

A továbbiakban a numerikus adattér (adatmátrix) mindkét irányában (tehát objektum- és attribútum-irányban is) fogunk elemzéseket végezni.

1.3. Hasonlósági tér létrehozása

Az objektumtér feldolgozásának következő lépése egy hasonlósági tér létrehozása a (numerikus) adattérből az objektum-vektorok összehasonlítása alapján (*hasonlósági transzformáció*).

A hasonlósági térben meghatározzuk azon objektumokat, melyek "elegendő mértékben" reprezentálják az eredeti adatteret (*hasonlósági redukció*), és az ezen objektumok által reprezentált környezeteket (*hasonlósági osztályozás*).

A továbbiakban *antimetrikus hasonlósági terekkel* foglalkozunk, tehát az $E = \{e_1,...,e_m\}$ objektum-halmazon értelmezett $S = \langle E,\sigma\rangle$ *hasonlósági térnek* a

$$\sigma : E\times E \rightarrow [0,1]$$

közelségi függvény a *hasonlósági függvénye*.

Megjegyzés (Összevont és többdimenziós hasonlósági terek)

A továbbiakban az n-dimenziós adattérből egy egydimenziós, úgynevezett *összevont hasonlósági teret* fogunk létrehozni. Ez azt jelenti, hogy az objektumok E halmazán minden $e_i,e_j\in E$ objektum esetén az $A = \{A_1,...,A_n\}$ attribútum-halmaz összes attribútumára vonatkozóan egyetlen $\sigma(e_i, e_j)$ objektum-hasonlósági függvényt értelmezünk.

Nagyszámú attribútum esetén azonban e módszer alkalmazásakor előfordulhat, hogy némely (és esetleg éppen a vizsgált jelenség szempontjából fontos) attribútum "eljelentéktelenedik", vagyis ezek értékei nem jelennek meg elég szemléletesen a hasonlóságokban. Ekkor célszerű áttérni a *többdimenziós hasonlósági terek* használatára (lásd később).

1.4. Hasonlósági transzformációk

A $Z = \langle E, A, z\rangle$ numerikus adatteret az $S = \langle E,\sigma\rangle$ hasonlósági térre leképező *hasonlósági transzformációt* az objektumok halmazán értelmezett *objektum-hasonlósági függvénnyel* definiáljuk, melyet az egyszerűség érdekében szintén a σ betűvel jelölünk.

Az $e_i,e_j\in E$ *objektumok hasonlóságát* többféleképpen is definiálhatjuk. Az alábbiakban bemutatunk néhányat, ahol $e_i,e_j\in E$, $A = \{A_1,...,A_n\}$, és a 0/0 alakú törtkifejezéseket 0-val helyettesítjük.

1.) Ha az adattér attribútumai azonos típusúak, akkor használhatjuk az alábbi kifejezést:

$$\sigma(e_i, e_j) = \frac{\sum_{A_k \in A} Min\left(z(e_i, A_k), z(e_j, A_k)\right)}{\sum_{A_k \in A} Max\left(z(e_i, A_k), z(e_j, A_k)\right)}.$$

2.) Ha az adattér attribútumai különböző típusúak, akkor célszerűbb az alábbi kifejezés:

$$\sigma(e_i, e_j) = \frac{1}{\mu(A)} \sum_{A_k \in A} \frac{Min\left(z(e_i, A_k), z(e_j, A_k)\right)}{Max\left(z(e_i, A_k), z(e_j, A_k)\right)},$$

illetve, ha zérus nem fordulhat elő a $z(e_1, A_k)$ értékek között, akkor az alábbi kifejezés:

$$\sigma(e_i, e_j) = \prod_{A_k \in A} \frac{Min\left(z(e_i, A_k), z(e_j, A_k)\right)}{Max\left(z(e_i, A_k), z(e_j, A_k)\right)}.$$

3.) Ha az adattér attribútumai csak két (például "igen", "nem") értéket vehetnek fel, és mindkét értéket azonos súllyal szeretnénk figyelembe venni az elemzés során, akkor jól használható az alábbi kifejezés:

$$\sigma(e_i, e_j) = \frac{1}{\mu(A)} \sum_{A_k \in A} b(e_i, e_j, A_k) \text{ , ahol}$$

$$b(e_i, e_j, A_k) = \begin{cases} 1, \text{ ha } z(e_i, A_k) = z(e_j, A_k) \\ 0, \text{ egyébként} \end{cases}.$$

4.) A metrikus térben szokásos távolság-képlet, és a multiplikatív metrikus transzformáció alapján adódik:

$$\sigma(e_i, e_j) = \frac{1}{1 + \sqrt{\sum_{A_k \in A} \left(z(e_i, A_k) - z(e_j, A_k)\right)^2}}.$$

2. HASONLÓSÁGI ADATREDUKCIÓ

Az alábbi példa bemutat egy, vásárlási szokásokat felderítő, hasonlóság-alapú elemzést.

2.1. Az objektumtér előkészítése feldolgozásra

.1. A mintapélda eredeti adattáblája és numerikus adattere

Az elemi vásárlások alábbi adattáblájában egy vásárlást (egy vásárlót) egy sor reprezentál, és az *i*-edik sor *k*-adik oszlopában álló "+" jel azt jelenti, hogy az *i*-edik vásárlásban szerepelt a *k*-adik termék.

A *numerikus Z adatteret* az eredeti *U objektumtérből* (az "őstáblából") úgy hozzuk létre, hogy az objektumtér adatértékeit leképezzük a {0,1} értékhalmazba, és az 1 értékhez rendeljük az adattábla *preferált* (vagyis az alkalmazás szempontjából kitüntetett, "+" módon jelölt) adatértékeit.

A $Z = \langle E, A, z \rangle$ numerikus adattér tehát az objektumok (a vásárlások) $E = \{e_1,...,e_{11}\}$ halmazából, az attribútumok (a termékek) $A = \{A_1,...,A_5\}$ halmazából, és a $z : E \times A \rightarrow \{0,1\}$ adatfüggvényből épül fel. Az így létrehozott *Z adatteret* az alábbi (szintén) *Z* jelű *adatmátrixszal* reprezentáljuk, és ebben az e_i objektum A_k attribútumhoz tartozó értékét $Z[e_i, A_k]$ módon jelöljük.

Az eredeti objektumtér, "őstábla" (*U*):

Ssz	Krumpli	Dinnye	Pelenka	Sör	Toll
1	+	+			
2			+	+	
3	+				
4	+	+			
5		+	+		
6				+	+
7	+		+	+	
8	+	+	+	+	
9	+	+		+	
10	+		+		
11			+	+	

A numerikus tér adattáblája (*Z*):

Z	A_1	A_2	A_3	A_4	A_5
e_1	**1**	**1**	**0**	**0**	**0**
e_2	**0**	**0**	**1**	**1**	**0**
e_3	**1**	**0**	**0**	**0**	**0**
e_4	**1**	**1**	**0**	**0**	**0**
e_5	**0**	**1**	**1**	**0**	**0**
e_6	**0**	**0**	**0**	**1**	**1**
e_7	**1**	**0**	**1**	**1**	**0**
e_8	**1**	**1**	**1**	**1**	**0**
e_9	**1**	**1**	**0**	**1**	**0**
e_{10}	**1**	**0**	**1**	**0**	**0**
e_{11}	**0**	**0**	**1**	**1**	**0**

A továbbiakban a fenti adatmátrix mindkét irányában (tehát objektum- és attribútum-irányban is) fogunk elemzéseket végezni.

.2. A hasonlósági tér létrehozása

A $Z = \langle E, A, z \rangle$ numerikus adattér feldolgozásának első lépése egy $S = \langle E, \sigma \rangle$ hasonlósági tér létrehozása az adattér objektumaiból az objektum-vektorok összehasonlítása alapján. Mivel e példában az adattér attribútumai azonos típusúak, így a korábban bemutatottak közül az

$$\sigma(e_i, e_j) = \frac{\sum_{A_k \in A} Min\big(Z[e_i, A_k], Z[e_j, A_k]\big)}{\sum_{A_k \in A} Max\big(Z[e_i, A_k], Z[e_j, A_k]\big)}$$

objektum-hasonlósági függvényt választjuk, ahol $e_i, e_j \in E$, $A = \{A_1,...,A_5\}$, és a 0/0 alakú törtkifejezéseket 0-val helyettesítjük.

Megjegyezzük, hogy bár mindegyik attribútum csak két értéket vehet fel ("van", "nincs"), mégsem célszerű az objektum-hasonlósági példa-függvények közül a 3.-at választani, mivel két vásárlás közötti hasonlóság szempontjából nagyobb a jelentősége annak, ha egy termék mindkettőben szerepel, mint annak, ha egyikben sem.

2.2. Objektum-redukció, és osztályozás

.1. Hasonlósági osztályozás és hasonlósági térredukció

Az objektum-minősítő paraméterek önkényes beállítása

Az objektum-minősítő vektor:

$q = \langle \sigma_{min}, \mu_{min}, C_{min} \rangle$,

ahol legyen

a határhasonlóság: $\sigma_{min} = 0{,}6$,

a minimális hasonlósági osztály mérete: $\mu_{min} = 3$, és

a minimális hasonlósági lefedettség: $C_{min} = 0{,}7$.

Megjegyzés

E paraméterek önkényes megadásánál figyelembe vesszük a korábbi elemzések tapasztalatait, valamint a kijelölt határhasonlóság alapján létrejött hasonlósági osztályokat.

Egy eljárás a hasonlósági térredukcióra

A fentiek alapján a hasonlósági osztályozás során tehát arra törekszünk, hogy adott hasonlósági tér, és adott határhasonlóság esetén minél kevesebb reprezentatív elemből, minél kevesebb hasonlósági osztályból építsük fel a hasonlósági teret. Ennek érdekében különböző algoritmikus módszereket használhatunk, melyek segítségével a redukálandó (a "maradék") tér egyre kisebbé válik. Az alábbiakban bemutatunk egy lehetséges módszert a hasonlósági térredukcióra:

1. *lépés*

   Először jelöljük ki reprezentatív objektumként a redukálandó ("maradék") tér azon e_m objektumát, melynek σ_{min} határhasonlóságú hasonlósági környezetéhez a legnagyobb összesített hasonlóság tartozik az eredeti S térben, vagyis, melyre az

   $$\sum_{e_i \in Env(S, e_m, \sigma_{min})} \sigma(e_m, e_i)$$

   érték a legnagyobb.

2. *lépés*

   Ha több ilyen is van, akkor közülük azt az e_k objektumot válasszuk ki először, amelyik a legnagyobb (σ_{min} határhasonlóságú) hasonlósági környezettel rendelkezik, vagyis amelyik esetén a

   $\mu(Env(S, e_k, \sigma_{min}))$

   érték a legnagyobb.

3. *lépés*

   Ha még ezekből is több van, akkor közülük egy olyat válasszunk ki először, mely a legnagyobb hasonlósági értékkel rendelkezik a saját (σ_{min} határhasonlóságú) hasonlósági környezetének valamelyik elemével.

4. *lépés*

Az *n*-edik reprezentatív objektum kiválasztása után a következőt az eredeti térnek a kiválasztottakon kívüli (a maradék) térrészéből választjuk ki az összes 3-ik lépés szerintiek közül (tehát ez a legbelső ciklus), majd az összes 2-ik lépés szerintiek közül, végül az összes 1-ső lépés szerintiek közül (ez tehát a külső ciklus), mindaddig, amíg a redukálandó tér üressé nem válik.

A hasonlósági osztályozás eredménye

A feladatra elvégezve a hasonlósági osztályozási eljárást, az *S* tér *reprezentatív objektumainak halmaza*:

$$E_{Rep} = \{e_3, e_5, e_6, e_7, e_9\},$$

és így az $e_m \in E_{Rep}$ reprezentatív objektumnak a σ_{min} határhasonlósághoz tartozó *hasonlósági környezete* az

$$E(e_m) = Env(S, e_m, \sigma_{min})$$

kifejezés alapján:

$$E(e_3) = \{e_3\},\ E(e_5) = \{e_5\},\ E(e_6) = \{e_6\},\ E(e_7) = \{e_2, e_7, e_8, e_{10}, e_{11}\},\ E(e_9) = \{e_1, e_4, e_8, e_9\}.$$

Mivel a minimális hasonlósági osztály méretére vonatkozó ($\mu_{min} = 3$) feltételnek csak az $E(e_7)$, és az $E(e_9)$ hasonlósági környezetek felelnek meg, ezért a *minősített reprezentatív objektumok halmaza*:

$$E_{qRep} = \{e_7, e_9\},$$

és így az E_{Rep} halmazbeli reprezentatív objektumoknak a σ_{min} határhasonlósághoz, és a μ_{min} minimális hasonlósági osztály mérethez tartozó *minősített hasonlósági osztályai* az

$$S_q(e_m) = S \| Env(S, e_m, \sigma_{min})$$

kifejezés alapján:

$$S_q(e_7) = S \| \{e_2, e_7, e_8, e_{10}, e_{11}\},\quad S_q(e_9) = S \| \{e_1, e_4, e_8, e_9\}.$$

Mivel a *minősített lefedő résztér*:

$$S_q = S_q(e_7) \cup S_q(e_9) = S \| \{e_1, e_2, e_4, e_7, e_8, e_9, e_{10}, e_{11}\},$$

így a *hasonlósági lefedettség*:

$$C = \mu(Obj(S_q)) / \mu(Obj(S)) = 8 / 11 = 0{,}73.$$

Mivel ez az érték nagyobb a minimális hasonlósági lefedettség ($C_{min} = 0{,}7$) értékénél, így a *minősített hasonlósági osztályozás*

$$\Omega_{qS} = \{S_q(e_7), S_q(e_9)\}$$

eredményét elfogadhatjuk.

A hasonlósági tér táblázatos szemléltetése

Az alábbiakban az egyes hasonlósági objektum-osztályokat az úgynevezett *hasonlósági táblá-jukkal* adjuk meg.

Eredeti numerikus adattábla:

Z	A_1	A_2	A_3	A_4	A_5
e_1	**1**	**1**	**0**	**0**	**0**
e_2	**0**	**0**	**1**	**1**	**0**
e_3	**1**	**0**	**0**	**0**	**0**
e_4	**1**	**1**	**0**	**0**	**0**
e_5	**0**	**1**	**1**	**0**	**0**
e_6	**0**	**0**	**0**	**1**	**1**
e_7	**1**	**0**	**1**	**1**	**0**
e_8	**1**	**1**	**1**	**1**	**0**
e_9	**1**	**1**	**0**	**1**	**0**
e_{10}	**1**	**0**	**1**	**0**	**0**
e_{11}	**0**	**0**	**1**	**1**	**0**

Hasonlósági tér adattáblája:
(objektum-irányú feldolgozás)

S	e_1	e_2	e_3	e_4	e_5	e_6	e_7	e_8	e_9	e_{10}	e_{11}
e_1	**1,00**	0,00	0,50	**1,00**	0,33	0,00	0,25	0,50	**0,67**	0,33	0,00
e_2	0,00	**1,00**	0,00	0,00	0,33	0,33	**0,67**	0,50	0,25	0,33	**1,00**
e_3	0,50	0,00	**1,00**	0,50	0,00	0,00	0,33	0,25	0,33	0,50	0,00
e_4	**1,00**	0,00	0,50	**1,00**	0,33	0,00	0,25	0,50	**0,67**	0,33	0,00
e_5	0,33	0,33	0,00	0,33	**1,00**	0,00	0,25	0,50	0,25	0,33	0,33
e_6	0,00	0,33	0,00	0,00	0,00	**1,00**	0,25	0,20	0,25	0,00	0,33
e_7	0,25	**0,67**	0,33	0,25	0,25	0,25	**1,00**	**0,75**	0,50	**0,67**	**0,67**
e_8	0,50	0,50	0,25	0,50	0,50	0,20	**0,75**	**1,00**	**0,75**	0,50	0,50
e_9	**0,67**	0,25	0,33	**0,67**	0,25	0,25	0,50	**0,75**	**1,00**	0,25	0,25
e_{10}	0,33	0,33	0,50	0,33	0,33	0,00	**0,67**	0,50	0,25	**1,00**	0,33
e_{11}	0,00	**1,00**	0,00	0,00	0,33	0,33	**0,67**	0,50	0,25	0,33	**1,00**

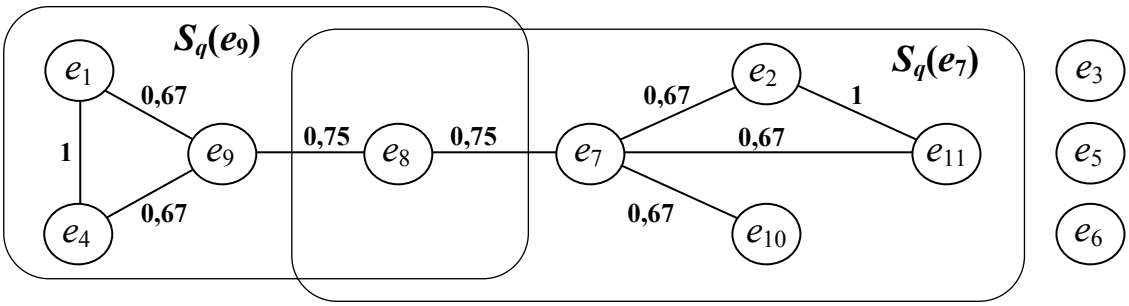
Hasonlósági táblák:

$Z_q(e_9)$	A_1	A_2	A_3	A_4	A_5
e_1	1	1	0	0	0
e_4	1	1	0	0	0
e_8	1	1	1	1	0
***e*₉**	**1**	**1**	**0**	**1**	**0**

$Z_q(e_7)$	A_1	A_2	A_3	A_4	A_5
e_2	0	0	1	1	0
***e*₇**	**1**	**0**	**1**	**1**	**0**
e_8	1	1	1	1	0
e_{10}	1	0	1	0	0
e_{11}	0	0	1	1	0

A hasonlósági tér ábrázolása hasonlósági gráfként

A hasonlósági (súlyozott) gráfban tehát a szemléletesség érdekében a határhasonlóságnál kisebb hasonlóságú (súlyú) éleket, a hurokéleket, és az élek irányait nem tüntetjük fel.



.2. A hasonlósági osztályozás eredményének alkalmazása a numerikus adattérre

Az S hasonlósági térre kapott hasonlósági osztályozás eredményeit alkalmazva a Z numerikus adattérre, az alábbiakat kapjuk:

A Z numerikus adattér *minősített reprezentatív résztere* (adott határhasonlóságra és minősítő paraméterekre vonatkozóan) a

$$Z_{qRep} = Z \| E_{qRep}$$

kifejezés alapján:

$$Z_{qRep} = Z \| \{e_7, e_9\},$$

és a *minősített hasonlósági objektum-osztályainak halmaza*:

$$\Omega_{qZ} = \{Z_q(e_7), Z_q(e_9)\}, \text{ ahol}$$

$$Z_q(e_7) = Z \| \{e_2, e_7, e_8, e_{10}, e_{11}\}, \quad Z_q(e_9) = Z \| \{e_1, e_4, e_8, e_9\}.$$

Ekkor az adattér *minősített lefedő résztere*:

$$Z_Q = Z_q(e_7) \cup Z_q(e_9),$$

a *lefedetlen résztere* pedig:

$$Z_N = Z \setminus Z_q = Z \| \{e_3, e_5, e_6\}.$$

2.3. Attribútum-irányú feldolgozás: Attribútum-redukció

A numerikus adattér feldolgozásának második lépése azon attribútumok kijelölése, és ezek azon attribútum-értékeinek meghatározása, melyek "elegendő" mértékben jellemzik az egyes hasonlósági osztályokat. A hasonlósági osztályok jellemző attribútum-értékeinek meghatározására különböző statisztikai jellemzőket használhatók, ezek közül mi az egyszerű átlagot választottuk. (A továbbiakban a korábban bevezetett jelölésekre fogunk hivatkozni.)

Reprezentatív objektum attribútum-átlaga, attribútum-referenciája

A $Z_q(e_m) \in \Omega_{qZ}$ minősített hasonlósági objektum-osztályban egy $A_k \in A$ attribútum *minősített hasonlósági osztályátlaga*:

$$Avg\left(Z_q(e_m), A_k\right) = \frac{\sum_{e_i \in Obj(Z_q(e_m))} z(e_i, A_k)}{\mu(Obj(Z_q(e_m)))}.$$

A $Z_q(e_m)$ objektum-osztályban az A_k attribútumot a

$$Ref(Z_q(e_m), A_k)$$

attribútum-referencia értékkel jellemezzük, ahol

$$Ref \in \{Ref_F, Ref_R\}, \quad \text{és}$$

a.) az *egzakt referencia-érték* használata esetén:

$$Ref_F(Z_q(e_m), A_k) = Avg(Z_q(e_m), A_k),$$

b.) a *közelített referencia-érték* használata esetén:

$$Ref_R(Z_q(e_m), A_k) = x, \quad \text{ahol}$$

$$Abs(x - Avg(Z_q(e_m), A_k)) = Min(\{Abs(y - Avg(Z_q(e_m), A_k)) \mid y \in A_k\}).$$

Megjegyzés

Fontos megjegyezni, hogy egy hasonlósági osztályban egyáltalán nem szükségszerű, hogy a reprezentatív elem képviselje az attribútum-referencia értéket.

Attribútum-deviancia

A $Z_q(e_m)$ objektum-osztályban az $A_k \in A$ attribútum $Dev(Z_q(e_m), A_k)$ *attribútum-deviancia* értéke megadja ezen osztálybeli objektumok A_k attribútum-értékeinek a $Ref(Z_q(e_m), A_k)$ referenciától való eltérésük (e referenciához viszonyított) relatív átlagát, azaz

$$Dev(Z_q(e_m), A_k) = \frac{\sum_{e_i \in Obj(Z_q(e_m))} Abs\left(z(e_i, A_k) - Ref\left(Z_q(e_i), A_k\right)\right)}{\mu(Obj(Z_q(e_m))) * Ref\left(Z_q(e_m), A_k\right)}.$$

Megjegyzés

Az attribútum-deviancia segítségével tehát minősíteni tudjuk az attribútum-referenciát.

Minősített attribútum, határdeviancia, attribútum-minősítő paraméter

Az A_k attribútum a $Z_q(e_m)$ objektum-osztálynak a $Ref(Z_q(e_m), A_k)$ attribútum-referencia értékkel jellemzett, és egy Dev_{max} *határdevianciával minősített attribútuma*, ha

$$Dev(Z_q(e_m), A_k) \le Dev_{max},$$

ahol a Dev_{max} határdevianciát *attribútum-minősítő paraméternek* is fogjuk nevezni, és ez (az objektum-minősítő paraméterekhez hasonlóan) az adattér feldolgozója által megadott *önkényes* érték.

Minősített referencia résztér, términősítő vektor

Először bevezetjük a *términősítő vektort*

$$Q = \langle \sigma_{min}, \mu_{min}, C_{min}, Dev_{max} \rangle$$

módon, mely az eddigi objektum-minősítő paramétereken (σ_{min} a határhasonlóság, μ_{min} a minimális hasonlósági osztály mérete, és C_{min} a minimális hasonlósági lefedettség) kívül tartalmazza a Dev_{max} határdevianciát, mint attribútum-minősítő paramétert.

Ezek után valamely $e_m \in E_{qRep}$ reprezentatív objektumnak egy Q términősítő vektorra vonatkozó *minősített attribútum-halmaza*:

$$A_Q(e_m) = \{A_k \mid A_k \in A, Dev(Z_q(e_m), A_k) \le Dev_{max}\},$$

és a Z numerikus adattérnek az e_m reprezentatív objektumra, valamint a Q términősítő vektorra vonatkozó *minősített referencia résztere*:

$$Y_{QRef}(e_m) = \{\langle e_m, A_k, Ref(Z_q(e_m), A_k)\rangle \mid A_k \in A_Q(e_m)\},$$

ahol a *Ref* lehet egzakt, vagy közelített referencia-érték (az alkalmazói igény szerint),

és *minősített referencia résztér-rendszere*:

$$Y_{QRef} = \{Y_{QRef}(e_m) \mid e_m \in E_{qRep}\}.$$

Megjegyzés

1.) Az $Y_{QRef}(e_m)$ minősített referencia résztereket *fiktív részstereknek* nevezzük, ha az egzakt referencia-érték meghatározást választjuk olyan esetben, amikor az attribútumok értékkészlete nem tartalmazza e referencia-értékeket. (Ennek oka lehet például az, hogy az értékkészlet diszkrét elemekből áll, lásd például a kirándulók gyűjtőpontjainak halmazát.)

2.) Valamely $e_i, e_j \in E_{qRep}$ reprezentatív objektumokhoz tartozó $Z_q(e_i)$ és $Z_q(e_j)$ minősített hasonlósági objektum-osztály e*kvivalens*, ha az $Y_{QRef}(e_i)$ és $Y_{QRef}(e_j)$ minősített referencia részstereik ugyanazokra az attribútumokra vonatkoznak, és az egyes attribútumokra vonatkozó referencia-értékek is megegyeznek. Az ekvivalens hasonlósági objektum-osztályok létrejöttének oka általában a túl magas határhasonlóság-érték.

A határdeviancia önkényes beállítása

A határdeviancia legyen

$$Dev_{max} = 0{,}4.$$

Tehát a términősítő vektor:

$$Q = \langle \sigma_{min}, \mu_{min}, C_{min}, Dev_{max} \rangle.$$

2.4. A feldolgozás eredménye

A Z numerikus adattér attribútumaira mindegyik hasonlósági objektum-osztályban meghatározva az attribútum-átlag és az attribútum-deviancia értékeket az alábbi eredményeket kaptuk (az úgynevezett *attribútum-minősítő táblákba* kiemelve az attribútum-átlag és az attribútum-deviancia értékeket):

Hasonlósági táblák:

$Z_q(e_9)$	A_1	A_2	A_3	A_4	A_5
e_1	1	1	0	0	0
e_4	1	1	0	0	0
e_8	1	1	1	1	0
e_9	**1**	**1**	**0**	**1**	**0**

$\Rightarrow$

Attribútum-minősítő táblák:

	A_1	A_2	A_3	A_4	A_5
Avg	**1,00**	**1,00**	0,25	0,50	0,00
Dev	**0,00**	**0,00**	1,50	1,00	####

$Z_q(e_7)$	A_1	A_2	A_3	A_4	A_5
e_2	0	0	1	1	0
e_7	**1**	**0**	**1**	**1**	**0**
e_8	1	1	1	1	0
e_{10}	1	0	1	0	0
e_{11}	0	0	1	1	0

$\Rightarrow$

	A_1	A_2	A_3	A_4	A_5
Avg	0,60	0,20	**1,00**	0,80	0,00
Dev	1,00	2,00	**0,00**	0,50	####

Mivel jelen esetben a megfelelő minősítésű attribútum-átlag értékek az attribútumok értéktartományainak elemei, így azok egyúttal attribútum-referencia értékeknek is tekinthetők.

A *minősített reprezentatív objektumok halmaza* (az objektum-redukció alapján):

$$E_{qRep} = \{e_7, e_9\},$$

az ehhez tartozó *minősített hasonlósági környezetek*:

$$E_q(e_7) = \{e_2, e_7, e_8, e_{10}, e_{11}\},\ E_q(e_9) = \{e_1, e_4, e_8, e_9\}.$$

Az eredmény

A fentiek alapján a Q términősítő vektorhoz tartozó *minősített referencia részterek* (az attribútum-redukció eredményeként):

$$Y_{QRef}(e_7) = \{\langle e_7, A_3, 1\rangle\},$$

$$Y_{QRef}(e_9) = \{\langle e_9, A_1, 1\rangle, \langle e_9, A_2, 1\rangle\},$$

és így a *minősített referencia résztér-rendszer*:

$$Y_{QRef} = \{Y_{QRef}(e_7), Y_{QRef}(e_9)\}.$$

Az eredmény értelmezése

Mit is jelent ez az eredeti vásárlási feladatra (az objektumtérre) vonatkozóan?

Az adatgyűjtés aktuális állapotában (vagyis a jelenlegi "őstábla" alapján) a vásárlók két fő csoportba sorolhatók; az egyikbe a pelenkát, a másikba a krumplit és dinnyét vásárlók tartoznak. Mivel pedig a két csoport nem csupán nem diszjunkt, hanem elég erősen össze is kapcsolódik (lásd az e_8 vásárlót, aki 75%-os mértékű hasonlósággal kötődik mindkét csoporthoz a hasonlósági térben), ezért az esetleges leárazási akciók során, vagy a termékeknek a boltban való elhelyezése során e kapcsolatot célszerű figyelembe venni.

Végül pedig (tegyük fel a kérdést) mi lehet az a háttérjelenség, ami a hasonlósági osztályozás e feladatbeli jellemzőjeként további vizsgálatra érdemesnek tűnhet? (Talán: a NŐ.)

IRODALOM

A szakirodalmi szokásokkal ellentétben e tanulmány végén nem sorolunk fel tételes irodalomjegyzéket.

Ennek egyik oka az, hogy munkánk újdonságjellege miatt (lásd korábban; Megjegyzések a szakirodalomhoz) nem tudunk olyan hivatkozást megadni, melyhez érdemben kapcsolódhatnánk. A másik ok, hogy az e tanulmányban példaként hivatkozható területek (a távolság-, és hasonlóság-függvény definíciók, a faktoranalízis, illetve a hagyományos, ekvivalencia-reláció szerinti, azaz diszjunkt osztályozás) munkái tömegesen találhatók meg mind a szakfolyóiratokban, mind az interneten.