Az adatbányászat osztályozási eljárásainak alkalmazása a vektoros térinformatikában

Elek István, az ELTE IK Térképtudományi
és Geoinformatikai Tanszékének docense,
az MTA Térképészeti és Térinformatikai kutatócsoportjának tagja

Bevezetés

A tematikus térkép attribútum adatokból levezethető csoportosítás eredményét jeleníti meg. A legtöbb esetben éppen az a cél, hogy egy kiválasztott objektumcsoport (layer, feature class) egy speciális adatféleség valamely tulajdonsága alapján legyen osztályozva (pl. előre megadott csoportszám, megadott vagy számított értékhatárok stb.).

Vegyünk egy egyszerű példát. Tegyük fel, hogy rendelkezünk egy olyan adatbázissal, amely a magyarországi településekről 50 adatot tartalmaz, mint például *lakosság szám, munkanélküliség, elvándorlás, odavándorlás, ipari létesítmények száma, ezek árbevétele (nyeresége), helyi adóbevétel, kulturális létesítmények száma, legalább egy idegen nyelvet beszélők száma, diplomások száma, funkcionális analfabéták száma* stb.

Egy ilyen értékes adatbázisból sokféle tematikus osztályozás és térkép készítése lehetséges, pontosabban ahány adatféleség, annyiféle tematikus térkép, annyiféle osztály. Ez a funkcionalitás a legtöbb esetben elegendőnek mondható.

Felmerülhet a kérdés, hogy mi van akkor, ha olyan osztályozási eljárást szeretnénk végrehajtani az adatsorunkon, amely minden adatot figyelembe vesz. Gondoljuk meg, hogy amikor egyetlen adatféleség (mint például a lakosság szám) alapján osztályozunk, akkor önkényesen elhagyjuk a többi adatot, vagyis az osztályozást nagyon jelentős adatvesztés után hajtjuk csak végre. Ezáltal akár elvileg meg is lehetne kérdőjelezni az eljárás korrektségét. A gyakorlati példánál maradva, a települések várossá nyilvánítása sem csak egyetlen ismérv alapján történik, hanem több adat egyidejű figyelembevétele által.

A következőkben megvizsgáljuk, hogy hogyan állíthatnánk elő a térinformatikai szoftverek jelenlegi adottságai mellett egy minden adatot figyelembe venni képes tematikus osztályozást és térképet.

A klaszterezésről általában

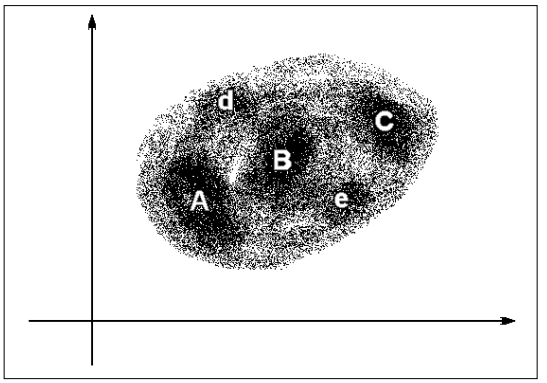
A nagy tömegű adathalmazokban való eligazodás meglehetősen bonyolult feladat, hiszen soha nem látott méretű adatbázisok jöttek és folyamatosan jönnek létre. Az adatok keresése, legyűjtése mögött gyakran valamilyen interpretációs szándék húzódik meg, amelyre az adatok szegmentálása, (tematikus) csoportokba foglalása teremti meg a lehetőséget. Mint tudjuk, az adat a gondolkodó ember fejében válik információvá, és ezt a folyamatot nagymértékben elősegítheti az adatok csoportosítása, tekintve, hogy javítja az áttekinthetőséget. Az informatika egyik modern ága az adatbányászat, amely nem kisebb feladatot tűzött ki, mint a nagyméretű adatbázisokban való eligazodást. A geoinformatikát a szakmai zsargon nem sorolja az adatbányászat témakörébe, pedig az adatbázisok mérete, az adatok sokfélesége és a grafika teremtette nehézségek ezt akár indokolhatnák is. Nem véletlen, hogy a térinformatikában – különösen a raszteres adatmodellt követő esetekben, mint amilyen az űrfotók feldolgozása – az adatbányászatban kiemelkedően fontos eljárás, a *klaszter analízis*, fontos szerepet játszik. Röviden összefoglaljuk a *klaszter analízis* főbb ismérveit, a dimenziócsökkentés lehetőségeit, valamint összehasonlítjuk a vektoros térinformatikában elterjedt tematikus kartográfiai modell és a többdimenziós osztályozások közötti hasonlóságot.

Egy adathalmaz pontjainak az adatrekordok hasonlósága alapján történő diszjunkt csoportokba sorolását klaszterezésnek nevezzük. A csoportosítás jósága alapvetően két dolgon múlik: a jó hasonlóság definíción és egy jó algoritmuson, amely a hasonlóságon alapulva valamilyen kritériumok alapján megállapítja a klasztereket.

Sokszor használjuk az osztályozás kifejezést is, ami majdnem ugyanazt jelenti, mint a klaszterezés. Míg a klaszterezés nem felügyelt csoportosítás, addig az osztályozás felügyelt. Ebben az

összefüggésben a felügyelt jelző azt jelenti, hogy a csoportok minőségi paraméterei előre definiáltak, míg a nem felügyelt esetben nem tudjuk, hogy milyen minőségi osztályba fognak tartozni az előálló csoportok, sőt ezek határai sem tudhatók előre.

Kétdimenziós esetben ábrázolva az adatpontokat, már szemrevételezéssel is el tudunk különíteni csoportokat az adatok sűrűsödése alapján (*1. ábra*). Látható, hogy az egy csoportba sorolt pontok egymáshoz közel vannak, illetve saját csoportjuk közepéhez közelebb, mint bármelyik másik csoport közepéhez.



1. ábra
Szemrevételezéssel is elvégezhető az osztályozás a egyszerűbb esetekben. Jól látható, hogy az *A, B, C* jelűek nagyobb, míg a *d, e* jelű csoportok kisebb jelentőségűek. Az is látható azonban, hogy a szemrevételezéses eljárás nem egzakt.

Több dimenzióban természetesen az effajta szemrevételezéses módszerek már nem használhatók. A korrekt osztályozás csakis klaszterezési algoritmusok révén valósítható meg. Itt terjedelmi korlátok okán csak két fő algoritmus csoportot említünk, mert egyébként sokféle eljárás létezik.

A távolság mátrix

A hasonlóság definiálásának egy kézenfekvő módja az euklideszi távolság fogalom. Jelöljön *u, v* két adatpontot. Az Euklideszi-távolságuk

$$d(u,v) = \left(\sum_{i=1}^{d} [u_i - v_i]^2 \right)^{1/2}$$

Jelöljük $d_{i\,j}$-vel az *i*-edik és *j*-edik adatpontok közötti távolságot. (Természetesen d_{ii}=0.) Írjuk fel az összes lehetséges távolságot egy mátrixban:

$$\mathbf{D} = \begin{pmatrix} d_{11} & d_{12} & \dots & d_{1n} \\ d_{21} & d_{22} & \dots & d_{2n} \\ \vdots & \vdots & \ddots & \\ d_{n1} & d_{n2} & \dots & d_{nn} \end{pmatrix}$$

A hasonlóság a távolság mintájára definiálható, hiszen az attribútum adatokból valamilyen függvény szerint hasonlósági mutatók képezhetők. Az *n* adat attribútum hasonlóságát definiáló mutatókat mátrixba rendezve kapjuk meg a hasonlósági mátrixot. A függvény ismeretében a hasonlósági mátrix tehát kiszámolható, de *n* nagy száma esetén előfordulhat olyan nagyméretű mátrix, amely lehetetlenné teszi a klaszteresés gyakorlati végrehajtását. Ilyenkor alkalmazzuk a dimenziócsökkentő eljárásokat, de erről a problémáról majd később szólunk.

Lényeges, hogy képesek legyünk megállapítani egy klaszterező eljárás jóságát. Egyik jósági mutató lehet egy előre megállapított, minta klaszter állomány előállításának képessége (vagyis hányszor téved az eljárás). Egy másik lényeges minőségi mutató a kiugró, kívülálló pontok meghatározásának képessége. Ez azért fontos, mert ha a kiugró pontok mérési hibából erednek, akkor felismerésük esetén ki tudjuk hagyni őket a számításokból, de az is lehet, hogy a pontok kívülállása éppen egy fontos jelenségre hívja fel a figyelmet, tehát külön klaszterben történő kezelésük különösen fontos.

Particionáló eljárások

A particionáló eljárások iterációs elven működnek, vagyis úgy érik el a megfelelő klaszterek megállapítását, hogy a folyamatosan kapott újabb és újabb klaszteréket pontosítják. Egy pillanatnyilag kialakult klasztert valamilyen módon tömören reprezentálunk (pl. súlypont), a klaszterek és az adatpontok közti távolságokat kiszámítjuk. Az adatpontokat újra particionáljuk, és a pontokat a hozzájuk legközelebb eső klaszterhez rendeljük. Az eljárás akkor áll le, amikor a soron következő iteráció után a partíció egy megadott értéknél kisebb mértékben változik.

Hierarchikus eljárások

A hierarchikus eljárások az adatelemeket fába rendezik el. Az adatok a fák leveleiben helyez-

kednek el, míg a fa minden belső pontja megfelel egy klaszternek. Ezek a klaszterek a fában alattuk lévő pontokat tartalmazzák. Egyes eljárások abból indulnak, hogy kezdetben minden adat egy külön klaszter, és az eljárás előrehaladása során a klaszterek csökkentése révén találja meg megfelelő felosztást. Más eljárások éppen fordítva, abból indulnak ki, hogy kezdetben egyetlen klaszter van, amely valamennyi adatpontot tartalmazza. Az eljárás az előrehaladása során – valamely leállási kritérium szerint – megáll, és az így kapott klaszterek jelentik a végeredményt. Akár a felhalmozó, akár a lebontó jellegű algoritmusokat nézzük, számos eljárás létezik az adatok klaszterekbe történő beosztásához.

Dimenziócsökkentés

Az ismertetett eljárások valamennyi rendelkezésre álló adatot figyelembe vesznek a csoportosítás elvégzéséhez. Az algoritmusok működési mechanizmusából látható, hogy a távolságok és a klaszter középpontok állandó számítása rendkívüli számítási igényt támaszt, különösen olyankor, amikor sokdimenziós az adatrendszer. Nem ritka, hogy több tíz adatféleség áll rendelkezésünkre objektumonként, amely elfogadhatatlan mértékben megnövelheti a számítási időt. Ilyen esetekben alkalmazunk dimenziócsökkentő eljárásokat.

A dimenziócsökkentés egyik lehetséges módja a főkomponens analízis, amely a többváltozós matematikai statisztika egy széleskörűen elterjedt eljárása. Legyen p számú megfigyelési egységünk, amelyek egyenként n számú adatot tartalmaznak (p számú megfigyelési vektorunk van).

$\mathbf{x}^1$	$\mathbf{x}^2$	...	$\mathbf{x}^p$
x_1^1	x_1^2	...	x_1^p
x_2^1	x_2^2	...	x_2^p
$\vdots$			$\vdots$
x_n^1	x_n^2	...	x_n^p

Tekintsük az x^j vektorokat valószínűségi változóknak, a vektorok elemeit a valószínűségi változók realizációinak. Standardizáljuk a változókat:

$$\tilde{x}_i^j = \frac{x_i^j - \overline{x}^j}{s^j}$$

ahol x^j a j-edik vektor elemeinek átlaga (a várható érték becslése) és s^j az empirikus szórása. Így tehát 0 várható értékűvé és 1 szórásúvá tettük a valószínűségi változóinkat. Ezek után számítsuk ki az adatrendszerünk korrelációs mátrixát:

$$\underline{\mathbf{R}} = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1p} \\ r_{21} & r_{22} & \dots & r_{2p} \\ \vdots & \vdots & \ddots & \\ r_{p1} & r_{p2} & \dots & r_{pp} \end{pmatrix}$$

ahol r_{ij} az i és j-edik megfigyelési egységek korrelációs együtthatója.

Határozzuk meg a korrelációs mátrix sajátértékeit és sajátvektorait, vagyis oldjuk meg a következő sajátérték egyenletet:

$$\underline{\mathbf{R}}\mathbf{v} = \lambda\mathbf{v}$$

Számítsuk ki a főkomponenseket a következő módon, legyen a j-edik főkomponens a következő:

$$C_i^j = \sum_p x_i^p v_p^j$$

ahol $i=1,n$ és $j=1,p$.

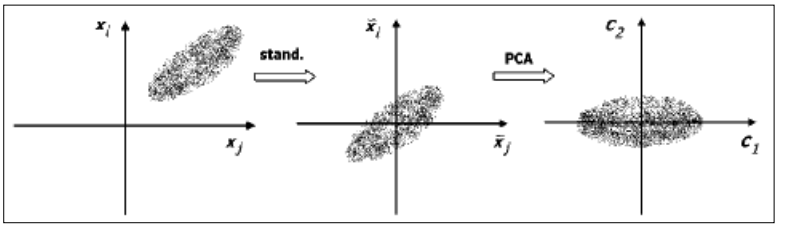
A főkomponensek ortogonális rendszert alkotnak, vagyis korrelálatlanok, azaz korrelációs mátrixuk

$$\underline{\mathbf{R}}_\mathbf{C} = \begin{pmatrix} \lambda_1 & & & 0 \\ & \lambda_2 & & \\ & & \ddots & \\ 0 & & & \lambda_p \end{pmatrix}$$

A $\underline{\mathbf{R}}_\mathbf{C}$ fontos tulajdonsága, hogy a főkomponensek és a standardizált változók összvarianciája azonos:

$$\sum_{j=1}^{p} \lambda_j = \sum_{i=1}^{p} \tilde{s}_i^2 = \sum_{j=1}^{p} s_j^2 = p$$

Amint látható, a főkomponensek kiszámításával nagymértékben átrendeztük a varianciákat, mivel, ha ez lehetséges volt, összevontuk őket az első (néhány) főkomponensben. Az eljárás főbb mozzanatainak geometria jelentését a *2. ábra* mutatja.



2. ábra
A folyamat geometria jelentése:
a standardizálás 0 várható értékűvé és 1 empirikus szórásúvá teszi a változókat, a főtengely transzformáció pedig beforgatja a pontfelhőt a legnagyobb változások irányába eső tengelyek irányába

Abban az esetben, ha például az első főkomponens képes magába sűríteni a megfigyelési egységek varianciáinak nagy részét, akkor megtehetjük, hogy az egész adatrendszert csak az első főkomponensével helyettesítjük. Ezzel végül is nem követünk el túl nagy hibát, viszont jelentős mértékben csökkentjük az adatrendszer dimenziószámát, ezzel adatszámát, és így meggyorsítottuk, megkönnyítettük egy soron következő eljárás, például a klaszter analízis működését.

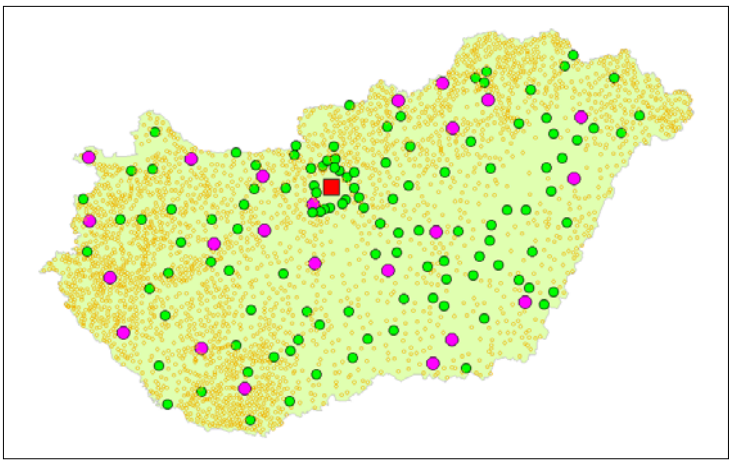
Felmerülhet a kérdés, hogy mikor nem használható az első főkomponens a teljes adatrendszer helyettesítésére? Ha a korrelációs mátrix diagonális, vagyis a változók korrelálatlanok, akkor biztosan nem. Ebben az esetben minden további számítás értelmetlen.

Egy másik kézenfekvő kérdés, hogy ha például a megfigyelési egységeink mérési értékek (pl. digitális képek, fizikai mennyiségek), akkor miért tekintjük őket valószínűségi változóknak? Ennek pusztán az az oka, hogy először a többváltozós matematikai statisztika használta ezt az eljárást dimenziószám csökkentésre. A probléma leírható algebrai módszerekkel is, és így a valószínűségi megközelítés mellőzhető.

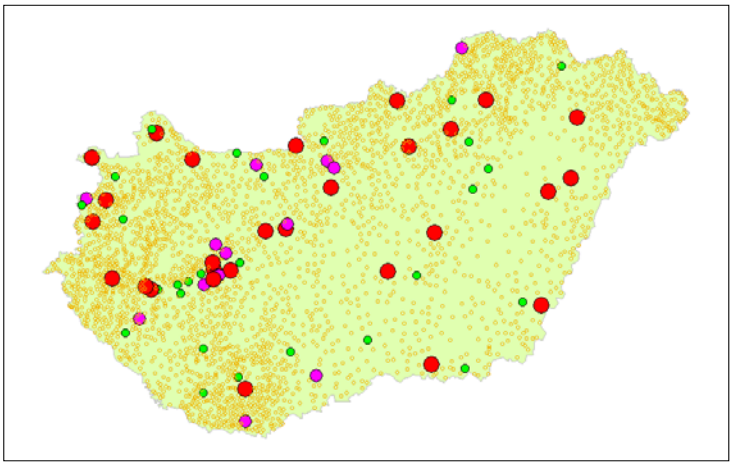
A tematikus térkép, mint egy klaszterezési eljárás eredménye

Az előbbi matematikai okfejtések célja az volt, hogy belássuk: a klaszterképzés révén megkönnyítjük, előkészítjük adataink értelmezését, és mindezt olyan eszközökkel tesszük, amelyeknek tulajdonságait előre meg tudjuk mondani, vagyis jóságukat képesek vagyunk minősíteni.

Az adatok osztályozásának hétköznapi módja a térinformatikában a tematikus térkép készítése. Ilyenkor a csoportosítás eredményét tematikus térképen jelenítjük meg, vagyis a csoportok állapotát szimbolizáló értékeket (fizikai mennyiségek, kategória változók stb.) valamilyen grafikus szimbólum vagy stílus (vonaltípus, kitöltési mintázat, szín) formájában mutatjuk meg a térképen. A geoinformatika gyakorlati alkalmazói – statisztikusok, elemzők, szociológusok stb. – jól ismerik az adatok láttatásának ezt a módját. Az informatikai piacon elérhető szoftverek kiváló tematikus térképkészítő képességekkel rendelkeznek, de az objektumok többdimenziós leírásából származó nehézségeket általában úgy oldják meg, hogy csak egyetlen változó értékeit jelenítik meg, ami persze dimenziócsökkentés, de egyáltalán nem optimális módon, hiszen az adatrendszer varianciáira nincs tekintettel, önkényesen hagy figyelmen kívül adatokat. Nyilvánvalóan megbízhatóbb az a csoportosítás, amely a tematikus felosztást nem egy önkényesen kiragadott adatféleség alapján végzi, hanem optimális adatvesztés révén hoz létre



3. ábra
Az állandó népesség területi megoszlását mutató tematikus térkép



4. ábra
A településenkénti vendégéjszakák számát mutató tematikus térkép

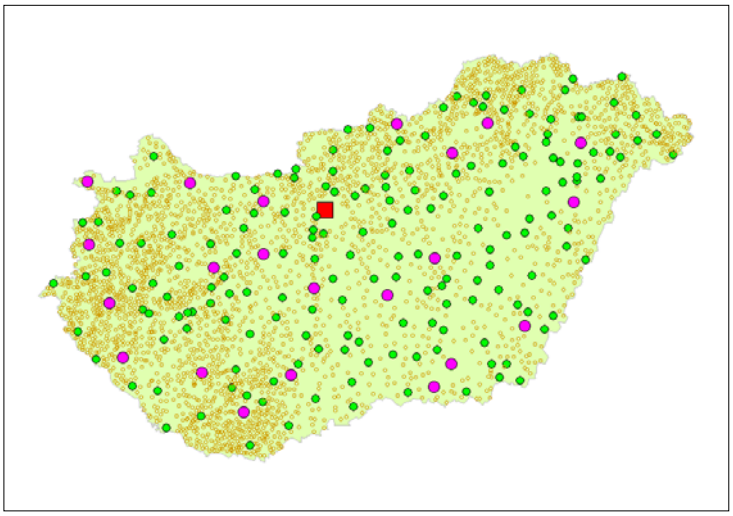
egy „szintetikus” adatsort, és ez alapján hozza létre a hagyományos tematikus térképet.

Hasonlítsuk össze a csak egyetlen megfigyelési egység adataiból származó különböző tematikus csoportosítások eredményét (*3.* és *4. ábrá*k) egy az első főkomponensen alapulóval (*5. ábra*).

Konklúzió

Minden olyan esetben, amikor sokdimenziós adatrendszerrel dolgozunk (bármelyik térinformatikai szoftverplatformon), és olyan osztályozásra van szükségünk, amely lehetőleg minden adatot figyelembe vesz, olyankor alkalmazzuk a főkomponens analízist. Ennek menete a következő:

– jelöljük ki a főkomponens analízisbe bevonni kívánt adatok körét (oszlopokat);
– exportáljuk a kijelölt adatbázis/tábla részt olyan formátumba, amit a rendelkezésünkre álló statisztikai szoftverünk képes beolvasni;
– hajtsuk végre a számításokat, tároljuk új oszlopként az első főkomponenst (vagy, ha úgy gondoljuk, többet is);
– importáljuk az első főkomponenssel bővített adatbázist;
– készítsünk az adott GIS szoftverre jellemző tematikus térképet az első főkomponens alapján.



5. ábra
A települések statisztikai adataiból számított első főkomponens szerinti tematikus térkép. Érdekes, hogy az így kapott tematikus térkép megegyezik a települések jogállását mutató térképpel

Amint a főkomponens analízist ismertető részben látható volt, a főkomponensek dimenziótlan számok, emiatt fizikai jelentésük sem nyilvánvaló. Egyes esetekben hipotézist állíthatunk fel arra vonatkozóan, hogy az első főkomponens az adatok hátterében meghúzódó valamilyen fizikailag is azonosítható, közös ok változó. Máskor meg kell elégednünk az első főkomponens variancia tartalmára épülő előnyökkel anélkül, hogy azonosítani tudnánk a háttérben lévő hatókat.

Data mining methods in vector based GIS

Elek, I.

Summary

The satellite image processing often uses data clustering methods to make images segmented. The result of the segmentation is a supervised or unsupervised classification on the certain image. In the vector based GIS there is a well known simple segmentation technique that is the thematic mapping. Unfortunately the vector GIS uses clustering methods very rarely, although the thematic mapping technique takes only one kind of data into consideration. The classical thematic mapping neglects the most of data except for only one, but the classification would be much better probably if you take all data into account.

The method introduced in this article suggests a new approach for thematic mapping based on the built-in software solutions in the existing GIS softwares. The main concept of this approach is the application of principal component analysis which produces the first principal component beeing the target of thematic mapping.

Irodalom

– *Iványi A.* (szerk) „Informatikai algoritmusok 1–2.”, ELTE Eötvös Kiadó, 2005
– *A. Stein–F. Meer–B. Gorte* (edited): „Spatial Statistics for Remote Sensing”, Kluwer Academic Publishers, 1999
– *I. Elek*: „Fast Porosity Estimation by Principal Component Analysis”, GEOBYTE, june 1990, Tulsa, Oklahoma
– *I. Elek*: „Some Applications of Principal Component Analysis: Well-to-Well Correlation, Zonation”, GEOBYTE, may 1988, Tulsa, Oklahoma
– *J. F. Richards*: „Remote sensing Digital image analysis”, Springer-Verlag, 1986, Australia
– *Vincze I.*: „Matematikai statisztika”, Tankönyvkiadó, Budapest, 1980
– *J. Davis*: „Statistics and Data Analysis in Geology”, John Wiley & Sons, Inc., 1973