MTA SZTAKI

Clusteranalizis chemotaxonomiai alkalmazása

Csiszár Imréné és Bene Béla

A növényrendszertanban az utóbbi 15 évben eredményesen alkalmazták az objektumok csoportositásának olyan matematikai módszereit, melyek nagyszámu jellemző együttes figyelembevételén alapulnak. N. Jardine és R. Sibson (3) terminológiáját követve, a clusteranalizisnek ezt a fontos alkalmazási területét matematikai taxonomiának nevezhetjük.

Az előadás tárgyát képező feladatban a csoportositás kémiai jellemzők alapján történt. 400 növényegyedet kellett csoportositani 20 féle kémiai komponensre vonatkozó adat alapján, ezenkivül vizsgálandók voltak az illető komponensek közötti kapcsolatok is.

A komponensek abszolut mennyisége az egyes növényegyedekben nem volt ismeretes, csak százalékos arányuk. 8 komponens az egyedek több mint 80 %-ában egyáltalán nem volt kimutatható.

A legtöbb cluster-módszer alapját az objektumok közötti valamilyen alkalmasan választott távolság-mértékszám képezi. Ilyen mértékszámul a 20-dimenziós adatvektorok euklidesi távolságát vettük. Próbálkoztunk az adatok logaritmus- illetőleg gyöktranszformációjával, valamint szórással való standardizálással is, de ezek nem vezettek különböző strukturáju clusterekre.

A matematikai taxonómiában általában hierarchikus eljárásokat szoktak használni. Ezek azonban nagy memóriaigényük miatt csak kisebb méretü feladatokra alkalmazhatók (szükség van ugyanis az objektumok közötti távolságok mátrixának tárolására). Ezért a növényegyedek csoportositására a gyors és kis memóriaigényü Mac Queen-módszert alkalmaztuk, mégpedig ennek azt a változatát, amikor a létesitendő clusterek számát nem határozzuk meg előre, csak felső korlátot irunk elő rá. A módszer leirását ld. M.R. Anderberg könyvében (1).

Ismeretes, hogy a Mac Queen módszer által szolgáltatott cluster-rendszer függ egyrészt az eljárás paramétereitől, másrészt a csoportositandó objektumok sorrendjétől. Negativ példaként megemlitjük, hogy egyszer az objektumokat rokonsági sorrendben bevive a program - a folyamatos középpontmódositás miatt - majdnem mindegyiket egy clusterba sorolta. A módszer gyorsasága lehetővé tette, hogy a programot többször lefuttatva a paraméterek és a kiinduló cluster-középpontokul szolgáló objektumok megfelelő kiválasztásával jó csoportositást kapjunk. A "jóság" jellemzésére a csoportokon belüli varianciák összegét használtuk.

Az eredmény chemotaxonomiai kiértékelését azzal kivántuk elősegiteni, hogy a programba közelitő t-próbát épitettünk be annak tesztelésére, hogy az egyes clusterekben mely komponensek átlaga tér el szignifikánsan a teljes anyagra vonatkozó átlagtól. Igy hasznos információt nyertünk arra nézve, hogy az egyes clusterek kialakitásáért mely komponensek felelősek. Megjegyezzük, hogy csak kevés egyedben előforduló komponensekre a t-próba az eloszlás ferdesége miatt nem alkalmazható. Ezért ezeknél azt teszteltük, hogy az illető komponens előfordulásának csoporton belüli relativ gyakorisága - binomiális eloszlást feltételezve - szignifikánsan különbözik-e a teljes anyagbeli relativ gyakoriságtól.

A vizsgálatban szereplő komponensekre vonatkozólag az volt a biokémiai modell, hogy ezek egyrészt bizonyos más, a vizsgálatban nem szereplő anyagokból, másrészt egymásból épülnek fel az anyagcsere során, és ezt a fa-strukturát kellett volna az adatokból kikövetkeztetni. Ilyen tipusu kérdésfeltevéssel az irodalomban nem találkoztunk, és nekünk sem sikerült olyan matematikai modellt felállitani, melynek alapján a feladat ilyen általánosságban kezelhető lett volna. Ez valószinüleg nem is lehetséges, mégis szeretnénk felhivni a figyelmet egy leszármazási fa megtalálásának matematikai problémájára, legalább abban a speciális esetben, amikor minden csucs a vizsgálatba bevont komponenst reprezentál.

A komponensek közötti kapcsolatokra vonatkozó többirányu vizsgálatainkat az a cél motiválta, hogy támpontot nyerjünk egy biokémiai alapon felállitott - részben még hipotetikus és ugyanakkor nem teljes - fa-struktura igazolásához, illetőleg kiegészitéséhez. Első tájékoztatásul kiszámitottuk a komponensek közötti korrelációs együtthatókat, egyrészt a teljes anyagból, másrészt azoknak a növényegyedeknek az elhagyásával, melyekben mindkét komponens mennyisége 0.

Másodsorban megvizsgáltuk, hogy azokra a növényegyedekre, amelyekben az i-edik komponens mennyisége nem 0, a többi komponens átlagos mennyisége szignifikánsan különbözik-e a teljes anyagbeli átlagtól.

További hasznos információt kaptunk a komponensek közötti összefüggésről a komponensek (mint objektumok) hierarchikus clusterositásával a nearest neighbour módszer alapján. Itt a komponenseket megfelelően standardizált 400 dimenziós vektorokkal reprezentáltuk és az euklidesi távolságot használtuk. Megjegyzendő azonban, hogy a kis távolság leszármazási kapcsolatként való interpretálhatósága kérdéses.

Végül kissé részletesebben beszélünk a komponensek rokonságának clique-módszerrel való analiziséről. Ez a cluster-módszer átfedő clustereket szolgáltat. Lényege, hogy egy adott d korláthoz tekintjük azt a gráfot, melynek csucsai a vizsgált objektumok, és azok az élek vannak behuzva, melyek d-nél nem nagyobb távolságokat jelentenek. Ennek a gráfnak a clique-jei (maximális teljes részgráfjai) lesznek a clusterek. A d korlátot fokozatosan növelve, hierarchikus cluster-rendszert kapunk. Programunkban C. Bron és J. Kerbosch (2) clique-kereső algoritmusát használtuk, amely a jelenleg ismert eljárások közül a leggyorsabb. Megemlitjük, hogy az emlitett szerzők empirikus eredményei szerint a módszer egy clique-re vonatkoztatott időigénye lényegében független a gráf nagyságától.

Programunk hisztogramot rajzol az egyes komponensek cliqueben való szereplésének gyakoriságáról és statisztikát készit arról is, hogy egyes komponenspárok hány clique-ban fordulnak elő. Ez a statisztika a közös clique-képzés erősségére enged következtetni és a komponensek származási rokonságát valószinüleg jobban jellemzi, mint a közönséges távolság.

Irodalom

(1) M.R. Anderberg: Cluster Analysis for Applications. Academic Press, New York, 1973.

(2) C. Bron, J. Kerbosch: Algorithm 457 - Finding all cliques of an undirected graph. Comm. of ACM 16: 575-577, (1973)

(3) N. Jardine, R. Sibson: Mathematical Taxonomy. Wiley, New York, 1971.