Budapesti Müszaki Egyetem Matematikai Tanszék

Egy "cluster"-eljárás

Fenyő István, Sima Dezső és Siminszky Márió

Adva N számu egyed, gyakori feladat ezeket bizonyos szempontok szerint k osztályba besorolni. Egy ilyen osztályt cluster-nek nevezünk ujabb irodalmi elnevezési szokás szerint. Ilyen cluster beosztás lehet diszjunkt, de nem feltétlenül kell annak lennie. Ilyen beosztási feladatot kombinatórikus uton elvégezni gyakorlatileg nem lehet, hiszen N egyedet k osztályba

$$\sum_{n_1+n_2+\ldots+n_k=N} \left[\binom{N}{n_1} + \binom{N-n_1}{n_2} + \ldots + 1 \right]$$

féleképpen lehet. Ha k = 2, már ebben az egyszerü esetben is

$$2^{N-1} - 1$$

elrendezés lehetséges.

Feladatkitüzés

Adott tehát N egyed, ezeket jelöljük $x_1, x_2, \ldots, x_N$-el. Mindegyik egyedet p számmal jellemezzük, ezeket az egyedek attributumainak nevezzük. Az x_i egyedet az $x_{i1}, x_{i2}, \ldots, x_{ip}$ attributumai definiálják, tehát minden x_i egyedhez egy p dimenziós vektort lehet hozzárendelni. Mi a jövőben az x_i egyedet az

$$x_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \qquad (i = 1,2, \ldots, N)$$

vektorral fogjuk azonositani.

Egy clusterba azokat az egyedeket - vektorokat - fogjuk sorolni, melyek "közel" vannak egymáshoz. Mi két vektort akkor fogunk közelinek tekinteni, ha euklideszi távolságuk

$$\| x_i - x_j \|^2 = \sum_{r=1}^{p} (x_{ir} - x_{jr})^2$$

bizonyos értelemben kicsi. Tegyük fel, hogy az i-ik cluster az $x_{i_1}, x_{i_2}, \ldots, x_{i_{n_i}}$ vektorokat tartalmazza.

Nyilvánvaló, hogy az i-ik cluster-t az $(i_1, i_2, \ldots, i_{n_i})$ multiindex jellemzi. Az $(1,2,3,\ldots,N)$ számok $(i_1, i_2, \ldots, i_{n_i})$ részhalmazát jelöljük a_i-vel, egyuttal ezzel jelöljük a fenti multiindexet is. Lehet természetesen $a_i = \emptyset$ (azaz $a_i = 0$), ez azt jelenti, hogy az i-ik cluster üres. n_i jelenti az i-ik cluster elemeinek a számát. Az i-ik cluster vektoraiból képezzük az

$$\bar{x}_{ij} = \frac{1}{n_i} \sum_{r \in a_i} x_{rj} \qquad i = 1,2,\ldots,k, \quad j = 1,2,\ldots,p$$

számokat. Ezekkel képezett

$$\bar{x}_i = (\bar{x}_{i1}, \bar{x}_{i2}, \ldots, \bar{x}_{ip})$$

vektort az i-ik cluster centrumának nevezzük.

Legyen most a clusterek száma, k adott. Azt a cluster beosztást fogjuk optimálisnak nevezni, melynél az

$$N_k = \sum_{i=1}^{k} \sum_{r \in a_i} \| x_r - \bar{x}_i \|^2$$

a lehető legkisebb.

A feladatot másképpen is meg lehet fogalmazni. Tekintsük a clustercentrumok sulyozott négyzettávolságát:

$$R_k = \sum_{i=1}^{k} \frac{n_i}{N} \| \bar{x}_i - \bar{\bar{x}} \|^2 ,$$

ahol

$$\bar{\bar{x}} = \frac{\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_k}{k}$$

és $M_k = N R_k$. Legyen továbbá

$$S = \sum_{i=1}^{N} \| x_i - \bar{\bar{x}} \|^2$$

Akkor

$$N_k + M_k = S \qquad (k = 1,2, \ldots, N)$$

minden clusterbeosztásnál. Jegyezzük meg, hogy S minden clusterbeosztásnál ugyanaz a szám, tehát a clusterbeosztásokkal szemben állandó. Ez azt jelenti, hogy ha valamely clusterbeosztásnál N_k minimális, akkor ugyanannál a clusterbeosztásnál M_k a legynagyobb. Tehát a feladatot ugy is meg lehet fogalmazni, az N vektort ugy kell k clusterbe csoportositani, hogy M_k a lehető legnagyobb legyen.

Pontok átrendezése

Tekintsük az x_s vektort, mely az i-ik clusterben van, azaz legyen $s \in a_i$. Azt fogjuk megvizsgálni, mi történik akkor az N_k--val, ha az x_s vektort az i-ik clusterből az j-edikbe csoportositjuk át.

Egy ilyen átrendezésnél az $\bar{x}_i$ és $\bar{x}_j$ centrumok megváltoznak, de megváltoznak a clusterre jellemző centrumoktól való távolságok négyzetösszegei is.

Legyen

$$\sigma_i^2 = \sum_{r \in a_i} \| x_r - \bar{x}_i \|^2 \; ; \quad \sigma_j^2 = \sum_{r \in a_j} \| x_r - \bar{x}_j \|^2 \; .$$

Átrendezés után az i-ik és j-ik clusterben a centrumok legyenek $\bar{x}_i'$ és $\bar{x}_j'$, akkor

$$\sigma_i'^2 = \sum_{r \in a_i - \{s\}} \| x_r - \bar{x}_i' \|^2 \; ; \quad \sigma_j'^2 = \sum_{r \in a_j \cup \{s\}} \| x_r - \bar{x}_j' \|^2 \; .$$

Ha az uj tagszámokat n_i' és n_j'-vel jelöljük, akkor nyilván

$$n_i' = n_i - 1 \; , \quad n_j' = n_j + 1 \; .$$

Átrendezés révén az N_k is megváltozik N_k'-ra. Ez a megváltozás az i-ik és j-ik clusterek megváltozásai következtében jött létre azáltal, hogy az x_s vektort az egyik clusterből a másikba rendeztük át. Ezért célszerü a

$$D_{ij}^{(s)} = N_k' - N_k$$

jelölés bevezetése.

Ha figyelembe vesszük, hogy

$$N_k = \sum_{i=1}^{k} \sigma_i^2 \; ,$$

akkor nyilvánvaló, hogy

$$D_{ij}^{(s)} = \sigma_i'^2 + \sigma_j'^2 - (\sigma_i^2 + \sigma_j^2) \; .$$

Tegyük fel, hogy $n_i > 1$. Az uj centrumokra nézve érvényes

$$\bar{x}'_i = (\bar{x}'_{i1}, \bar{x}'_{i2}, \ldots, \bar{x}'_{ip})$$

$$\bar{x}'_i = (\bar{x}'_{i1}, \bar{x}'_{i2}, \ldots, \bar{x}'_{ip}) ,$$

ahol most

$$\bar{x}'_{ir} = \frac{n_i \bar{x}_{ir} - x_{sr}}{n_i - 1} \qquad (r = 1,2,\ldots,p)$$

$$x'_{ir} = \frac{n_i x_{ir} + x_{sr}}{n_i + 1} .$$

Igy tehát

$$\sigma'^2_i = \sum_{h \in a_i - \{s\}} \| x_h - \bar{x}'_i \|^2 = \sum_{h \in a_i - \{s\}} \sum_{r=1}^{p} (x_{hr} - \bar{x}'_{ir})^2 =$$

$$= \sum_{h \in a_i - \{s\}} \sum_{r=1}^{p} \left(x_{hr} - \frac{n_i \bar{x}_{ir} - x_{sr}}{n_i - 1} \right)^2$$

és

$$\sigma'^2_i - \sigma^2_i \sum_{h \in a_i - \{s\}} \sum_{r=1}^{p} \left(x_{hr} - \frac{n_i \bar{x}_{ir} - x_{sr}}{n_i - 1} \right)^2 -$$

$$- \sum_{h \in a_i - \{s\}} \sum_{r=1}^{p} (x_{hr} - \bar{x}_{ir})^2 = - \frac{n_i}{n_i - 1} \| x_s - \bar{x}_i \|^2$$

Hasonlóan kapjuk meg $\sigma'^2_i - \sigma^2_i$ értékét is, ennek alapján

$$D_{ij}^{(s)} = \frac{n_j}{n_j + 1} \|x_s - \bar{x}_j\|^2 - \frac{n_i}{n_i - 1} \|x_s - \bar{x}_i\|^2 .$$

Ebből adódik, hogy az x_s vektort akkor érdemes átrendezni az a_i clusterből az a_j clusterba, ha

$$D_{ij}^{(s)} < 0,$$

egyébként az x_s marad az a_i clusterban.

Ha $n_i = 1$ lenne, akkor az x_s az a_i egyetlen eleme, ezért $\bar{x} = x_s$. Definiáljuk $\bar{x}_i' = 0$ és $n_i = 0$-nak, ezért

$$D_{ij}^{(s)} = \frac{n_j}{n_j + 1} \|x_s - \bar{x}_j\|^2 .$$

Ez pedig mindenképpen nemnegativ, amiből következtetünk arra, hogy egyelemü clusterből nem veszünk el elemet.

Most rögzitsük i-t és futtassuk j-t az $1,2,\ldots,(i-1),(i+1),\ldots,k$ számokon át. Minden j-hez kiszámitjuk a $D_{ij}^{(s)}$ számokat. Azokat a j--ket elhagyjuk, melyekre $D_{ij}^{(s)}$ nem negativok és az a_i cluster azon pontjait, melyekre a fenti kifejezés negativ az a_j clusterba helyezzük át.

Ha több mint egy j-re teljesül, hogy $D_{ij}^{(s)} < 0$, akkor azt a j_o-t vesszük, melyre $D_{ij_o}^{(s)}$ maximális és az x_s-et az a_{j_o}-ba csoportositjuk át. (Ha több ilyen j_o volna, akkor ezek közül bármelyiket veszszük.) Átrendezés után

$$N_k' = N_k + D_{ij_o}^{(s)} .$$

Ezt az eljárást minden ponttal megismételjük. Konvergensnek tekintjük clusterezési eljárásunkat, ha átfutva a pontokon, ujabb átrendezés nem szükséges többé. Ekkor N_k-nak legalábbis lokális minimumát értük el.

A leirt eljárásnak a hátránya, hogy több vektor szimultán átrendezése során az N_k mennyiség még jobban csökkenhet. E hátrány korrigálása érdekében egyik alkalmas ideiglenes clustert kettéosztunk, miáltal ideiglenesen az eredetileg k cluster helyett k+1 cluster keletkezik, ezekkel a fenti átrendezést megismételjük és aztán két clustert ismét egyesitünk. Ezt az eljárást is akkor fejezzük be, ha N_k már tovább nem csökken.

Egy cluster kettéosztása

A clusterek homoganitását varianciájukkal fogjuk mérni. A variancia

$$v_i = \frac{\sigma_i^2}{n_i} \quad (i = 1,2,\ldots,k) \quad (n_i > 1)$$

Azt a clustert tekintjük, mely legkevésbé homogén, azaz melynél a variancia a legnagyobb. Ha ilyen több van, akkor bármelyiket tekintjük. Legyen

$$z = \bar{x}_i + at \ ,$$

ahol z és a p-dimenziós vektorok, t pedig egy valós változó, mely minden valós értéket felvesz. Definiáljuk az i-ik clusterben a

$$s_{ij} = \frac{1}{n_i} \sum_{r \in a_i} x_{rj}^2 - \bar{x}_{ij}^2 \qquad (j = 1,2,\ldots,p)$$

$$s_{sj} = \frac{1}{n_i} \sum_{r \in a_i} x_{rs} x_{rj} - \bar{x}_{is} \bar{x}_{ij}$$

mennyiségeket. Legyen

$$\varrho_{i_o i_o} = \max_{1 \leqq i \leqq p} \varrho_{ii} > 0$$

és tekintsük az alábbi regressziós együtthatókat:

$$b_{i_o,i} = \frac{\varrho_{i_o,i}}{\varrho_{i_o,i_o}} \quad (i \neq i_o), \quad b_{i_o,i_o} = 1 .$$

A z vektor definiciójában szereplő a vektor j-ik komponense legyen $a_i = b_{i_o,i}$. Vegyük most az a irányába mutató egységvektort:

$$e = \frac{a}{\|a\|} ,$$

ennek komponenseit e_r-el fogjuk jelölni (r=1,2, ..., p). Legyen továbbá

$$\varrho_{i_o M} = \frac{1}{n_i} \sum_{r \in a_i} x_{r,i_o} (e, x_r) - x_{i,i_o}(e, \bar{x}_i) .$$

Akkor, ha azt akarjuk, hogy

$$\varrho_{i_o,M} = \frac{\sum_{j=1}^{p} \varrho_{i_o,i} a_i}{(\sum a_i^2)^{1/2}} = \max$$

legyen, akkor

$$\frac{\partial \varrho_{i_o,M}}{\partial a_i} = 0 .$$

Ezekből az egyenletekből adódik, hogy

$$\frac{a_j}{\varrho_{i_o,i}} = \text{konstans} \quad (j = 1,2, \ldots, p.)$$

Helyettesitsük t helyébe $\pm \sqrt{\varrho_{i_o,i_o}}$ -t és

$$\bar{x}'_{ij} = \bar{x}_{ij} + \frac{\varrho_{i_o,i}}{\sqrt{\varrho_{i_o,i_o}}} \qquad (j = 1,2, \ldots, p)$$

$$\bar{x}'_{k+1,j} = \bar{x}_{ij} - \frac{\varrho_{i_o,i}}{\sqrt{\varrho_{i_o,i_o}}}$$

kifejezéseket képezzük. Ha most az i-ik clusterből kettőt képeztünk, akkor az egyikhez az $\bar{x}'_i$ centrumot, másik feléhez az $\bar{x}''_i = \bar{x}'_{k+1}$ centrumot rendeljük hozzá. Ez utóbbi lesz a k+1-ik cluster. Az i-ik cluster n_i számu pontja közül azokat fogjuk az egyik vagy másik "csonka" clusterbe sorolni, melyek az uj centrumokhoz közelebb vannak. Most a keletkezett k+1 clusterre számitjuk ki az N_{k+1} számot és a pontokat az előbb leirt módszer szerint ugy csoportositjuk át, hogy az N_{k+1} a lehető legkisebb legyen. Ezekután a varianciák alapján esetleg ujabb clustereket osztunk szét és az eljárást többször megismételjük.

Clusterek egyesitése

Miután feladatunk szerint a clusterek száma adott, egyesiteni kell egyes clustereket, ha az előbbi módon kényszeritve voltunk egyes clustereket szétválasztani. Mindenesetre ha a clusterek számát csökkentjük, az N_k nem csökken:

$$N_k \geqq N_{k+1}$$

Ezért tehát azokat a clustereket kell egyesiteni, melyek összevonásakor N_k a legkevésbé növekszik.

A két cluster legyen az i-ik és a j-edik, ezekre legyen

$$\sigma_i^2 = \sum_{r \in a_i} \| x_r - \bar{x}_i \|^2, \quad \sigma_j^2 = \sum_{r \in a_j} \| x_r - \bar{x}_j \|^2 .$$

Ha egyesitjük e két clustert, akkor a szórásnégyzet

$$\sigma_{ij}^2 = \sum_{r = a_i \cup a_j} \| x_r - \bar{x}_{ij} \|^2 ,$$

ahol $\bar{x}_{ij}$ koordinátái

$$\bar{x}_{ij}^{(r)} = \frac{n_i \bar{x}_{ir} + n_j \bar{x}_{jr}}{n_i + n_j}$$

$\bar{x}_{ij}$ lesz a két cluster egyesitésekor keletkezett uj cluster centruma. Az egyesitéssel az N_k növekedése legyen $D_{ij} N_k$, ez pedig

$$D_{ij} N_k = \sigma_{ij}^2 - (\sigma_i^2 + \sigma_j^2) = N_k - N_{k+1}$$

Ebbe az előző kifejezéseket behelyettesitve az adódik, hogy

$$D_{ij} N_k = \frac{n_i n_j}{n_i + n_j} \| \bar{x}_i - \bar{x}_j \|^2 \qquad (i \neq j).$$

Természetesen, mint ahogyan azt az előbb már mondtuk az i és j számokat ugy határozzuk meg, hogy $D_{ij} N_k$ a lehető legkisebb legyen. Ha ezt a minimumot $D'_{ij} N_k$-val jelöljük, akkor

$$N_k = N_{k+1} + D'_{ij} N_k .$$

Iterative ezt az eljárást addig folytatjuk, mig az optimális N_k számot el nem érjük.

A clusterek száma nem adott

Ha a clusterek számát is meg kell határozni, akkor a következőképpen lehet eljárni. Adva kell, hogy legyen a clusterek számának legkisebb és legnagyobb értéke. Konkrét feladatnál ezt a fizikai adottságokból adódó becsléssel kell megállapitani. A clusterek számának legkisebb értékét jelöljük k_m, legnagyobb értékét k_M-el. ($k_m \geqq 2$). A keresett k a k_m és k_M között van.

Kiindulunk cikkünk elején emlitett

$$N_k + M_k = S$$

összefüggésből. Legyen

$$Z_k = \frac{M_k}{N_k} \frac{N-k}{k-1} \quad . \qquad (k = 1,2, \ldots, N-1) \; .$$

Látható, ha N_k minimális, akkor Z_k maximumát veszi fel. De

$$Z_k = \frac{S-N_k}{N_k} \frac{N-k}{k-1} \quad ,$$

Ugy járunk tehát el, hogy minden számbajövő k-nál megkeressük az optimális N_k-t a leirt módszerrel és egyuttal a Z_k számokat is kiszámitjuk. Azt a k-t tekintjük, melynél Z_k a legnagyobb.