ALGORITMUS A LEGJOBB PREDIKTORHALMAZ KIVÁLASZTÁSÁRA

Balás Éltes András
Semmelweis Orvostudományi Egyetem Számitóközpont

A klinikai gyakorlatban ma már egyre több diagnosztikus feladat megoldásához létezik szokásos laborvizsgálati rend, un. protokoll. Igy például magasvérnyomás, diabetes pontosabb diagnózisához, vagy chronikus vesebajos, müvese kezelt betegek állapotának periodikus ellenőrzéséhez. A jó protokoll korszerü sémát ad a szükséges és elégséges információtartalmu laboratóriumi vizsgálatok megrendeléséhez. A használatos vizsgálati protokollok, szokásos rendek jelentős része "házi" szabály, az adott kórház orvosainak tapasztalataira, véleményére épül, igazodik a helyi lehetőségekhez. Protokolljavaslatokat ugyanakkor nagy számban közöl a hazai és nemzetközi szakirodalom is.

Diagnosztikai protokollokkal szerzett tapasztalatok statisztikai elemzésekor gyakori észrevétel, hogy a vizsgálatok eredményei láthatóan összefüggnek, mégis két vizsgálat között olyan szoros kapcsolat ritkán bizonyitható, amely az egyik elhagyását a protokollból lehetővé teszi. Többváltozós lineáris kapcsolat is gyakran sejthető, de a célszerüen elhagyandó vizsgálatok kiválasztását ez még nem oldja meg.

Orvosi értelemben a feladat egy protokollban szereplő, szokásosan egyszerre megrendelt laboratóriumi vizsgálati halmazból a legköltségesebbek elhagyása, ugy, hogy az elhagyott vizsgálatok eredményét egy alkalmas matematikai formulával, kellő pontossággal kiszámithassuk a protokollban meghagyott laboratóriumi vizsgálatok eredményeiből.

A feladat matematikai megfogalmazása: $\underline{X} \in \mathbb{R}^n$ valószinüségi vektorváltozó, $\underline{c}$ a hozzárendelt költségvektor és a D mátrix kijelöli a prediktorváltozókat:

$$D_{n\times k} = \begin{cases} d_{ij} = 1 & \text{ha az i-edik prediktor a j-edik változó} \\ d_{ij} = 0 & \text{egyébként} \end{cases}$$

ahol $k = \Sigma\Sigma d_{ij}$. A cél D megválasztása ugy, hogy

a/ $M(X_j) = \Theta_j \cdot D \cdot M(\underline{X})$ ha $\exists$ i, melyre $d_{ij} = 1$

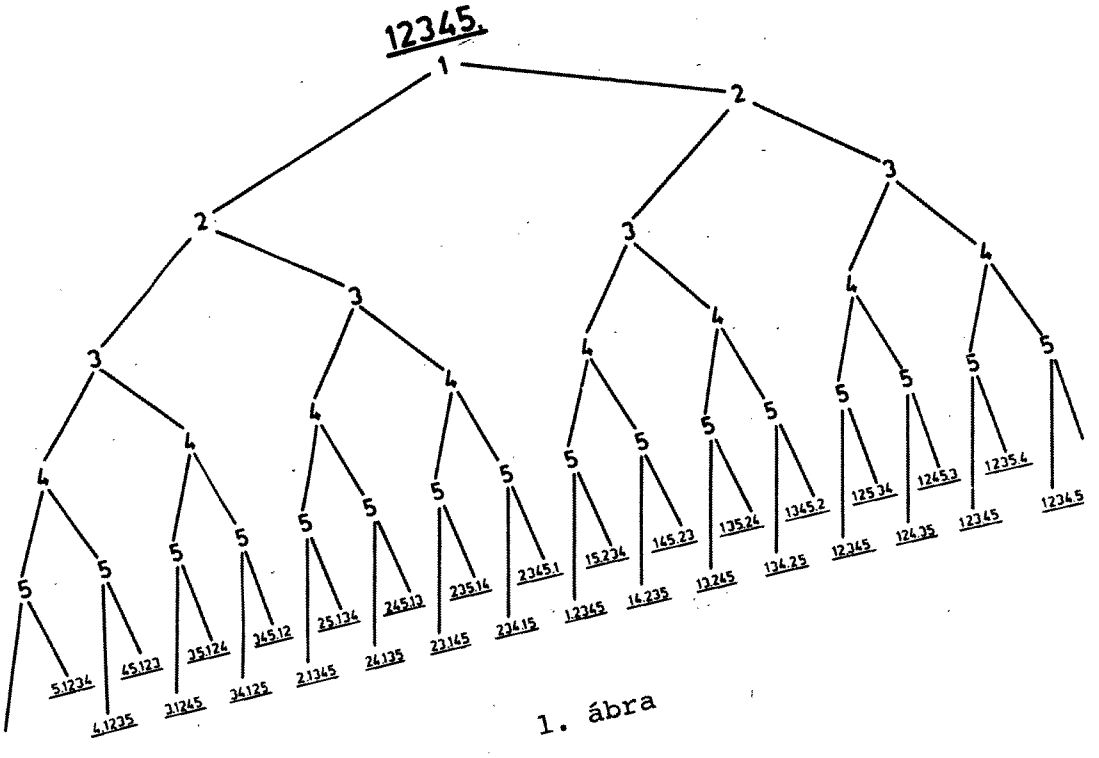
b/ $\min_i R_{X_j} \cdot D \cdot X \geq R_{min}$
$d_{ij} = 0$

c/ $\ell = \underline{1} \cdot D \cdot \underline{c} \rightarrow \min$

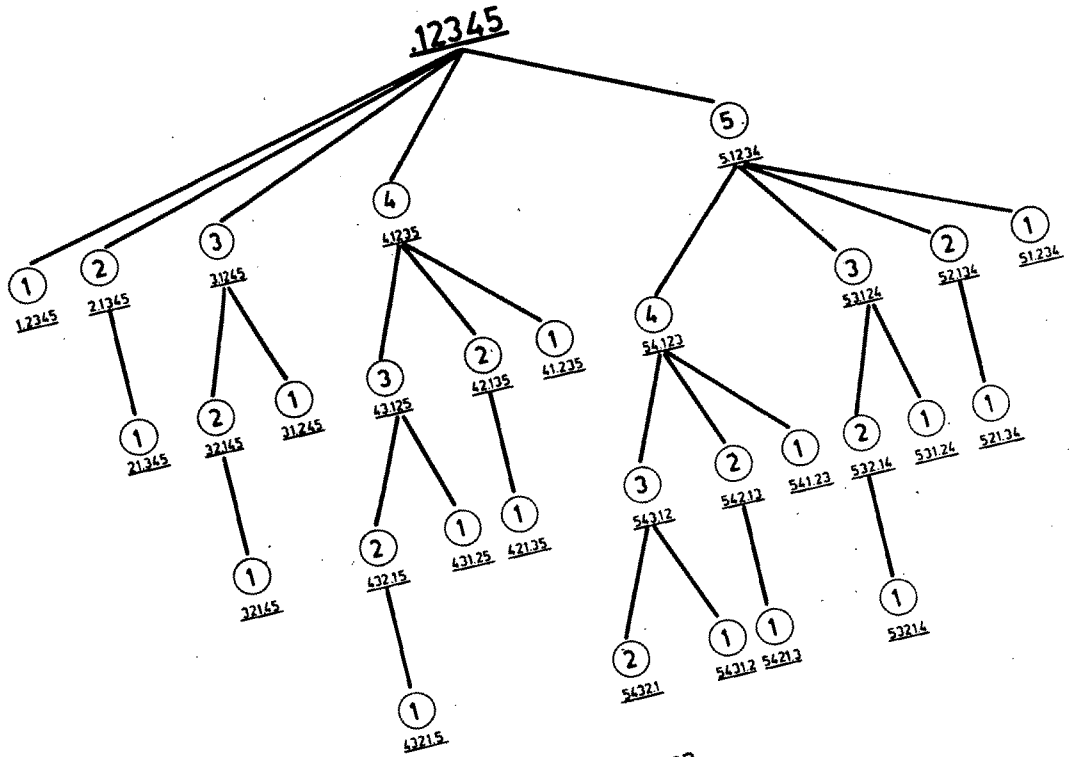
A Θ_j regressziós paramétervektor meghatározása, az R többváltozós korreláció számitása a prediktor változók kovarianciamátrixának invertálását jelenti.

Az egyszerünek látszó megközelités, minden lehetséges prediktorhalmaz kipróbálása és igy a legjobb kiválasztása rendkivül számításigényes, p változó esetén 2^p invertálást jelent.

Az 1. ábra módszert, fastrukturát ad a prediktorhalmazok megha-


12345.
1
2
3
4
5
5.1234
45.123
35.124
34.125
25.134
24.135
23.145
23.415
1.2345
14.235
13.245
134.25
12345
124.35
12345
12345

1. ábra


12345
1
2
3
4
5
12345
21345
31245
41235
51234
54123
53124
52134
51234
21345
32145
31245
43125
42135
41235
54312
54213
54212
54321
53214
43215

2. ábra

tározásához, a többváltozós analizisben használatos "pont" jelöléssel /a pont mögött a prediktorok, előtte a becsült változók, a prediktandusok/.

A számitást célszerüen lépésenként, azaz változónként sweepeléssel, vagy másnéven pivotálással hajthatjuk végre. A sweep operátor a sweepelt változók kovarianciamátrixának inverzét, a többi változókra pedig a reziduális négyzetösszeget adja. /Ez utóbbit azzal a feltevéssel számitja, hogy a sweepelt változók a prediktorok./ Az operátor invertálható, ezért a továbbiakban be-, illetve kisweepelésről beszélünk, aszerint, hogy az eredeti vagy az inverz leképezésről van szó. A sweepelés azonban különösen nagyobb mátrixok esetén rendkivül számitásigényes.

Az ábrán látható fán a szükséges besweepelések száma:

$$n = p \cdot (2^{p-1}-1)$$

A szükséges sweepelések száma majdnem a felére csökkenthető, ha a fastrukturát átrendezzük, ugy hogy minden sweepelés értékelhető prediktor részhalmazhoz vezessen. A 2. ábra példája mutatja, hogy a számitás ekkor valamennyi változó besweepelésével indul, majd az ábrán vázolt sorrendben kisweepelünk. A szükséges sweepelések száma:

$$n = (p+1) \cdot 2^{p-2} + p - 1$$

A sweepelések száma tovább csökkenthető, a fa alkalmas átrendezésével. Ennek szabályai:

- a változók költség szerint monoton csökkenő sorrendben vannak /1-es számu a legdrágább/,

- az egyes változók első kisweepelését jelentő főtörzsek alá a tőle balra eső főtörzsek forditott sorrendben kerülnek.

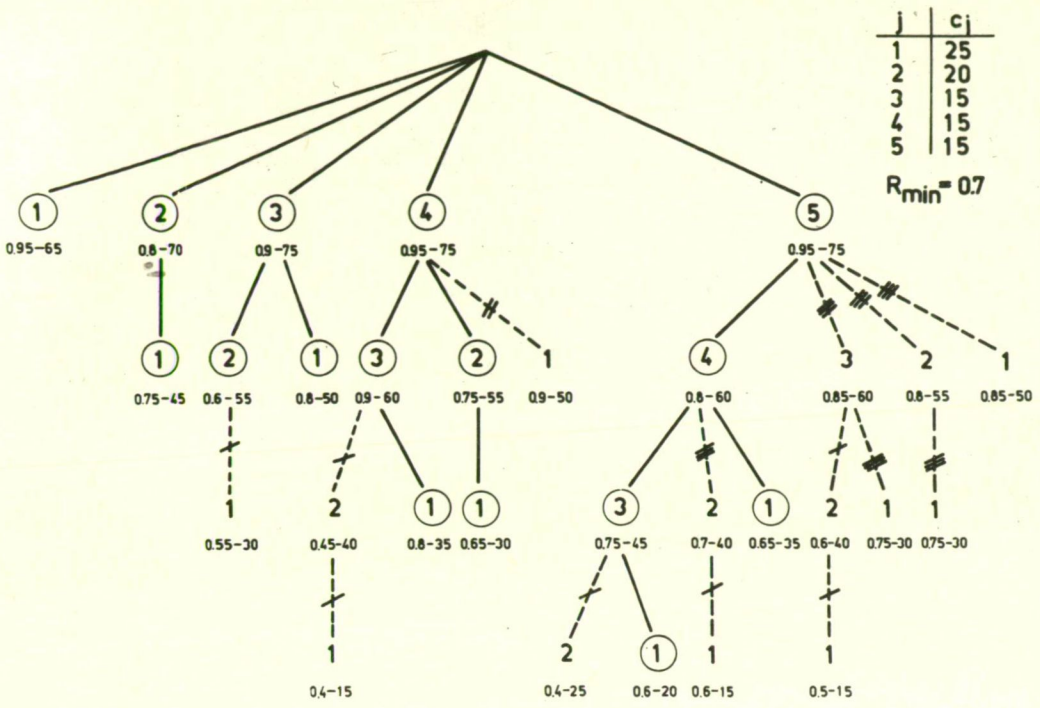
A 3-as ábra 5-változós számpéldával illusztrálja, hogyan hagyhatók el további sweepelések ebben az elrendezésben. /Sweepelni csak a bekarikázott csomópontokban kell./ A csomópontok alatt az adott prediktorhalmazokhoz tartozó minimális többváltozós korreláció és a költség látható. Felülről lefelé haladva az ágakon a vizsgálati költség és a minimális korreláció monoton csökken. Kisweepelés elhagyható:

- ha a minimális többváltozós korreláció az adott limit alá esik; a prediktorváltozók számának csökkentésével ez nem javulhat, igy az ág hátralévő része elhagyható a továbbiakban / $\not\downarrow$ jellel/,

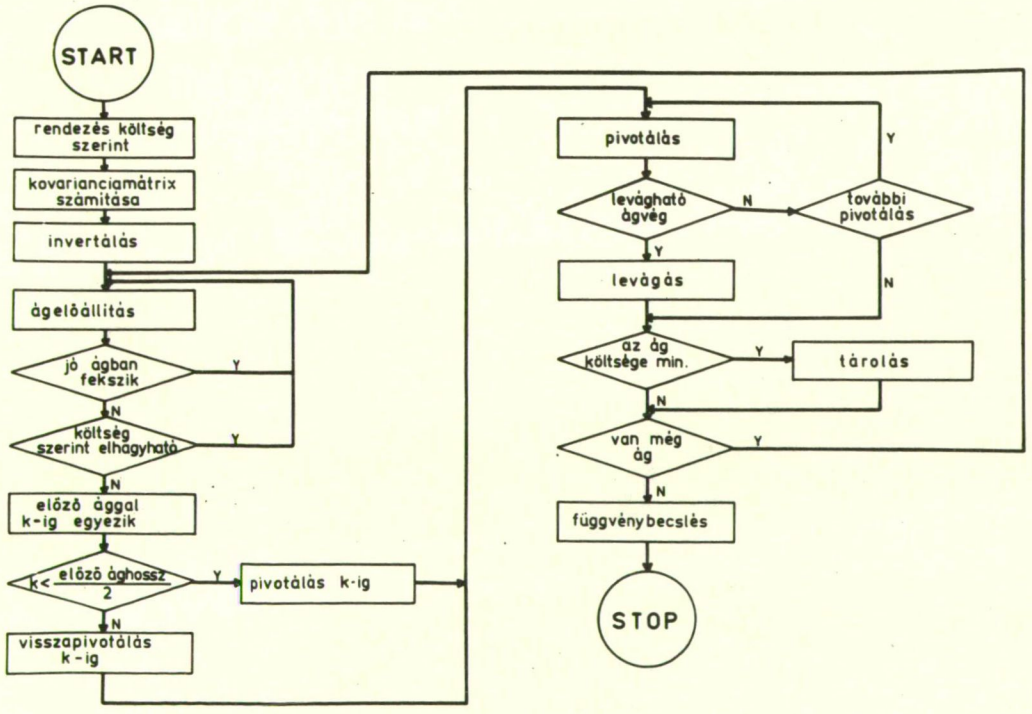
- ha egy ág sweepelési csomópontjai már mind szerepeltek egy korábbi jó ágban, akkor számitása az adott főtörzs alatt elhagyható, mert a pontossági kritériumnak biztosan megfelel, de költsége nagyobb, vagy egyenlő / $\not\leftarrow$ jellel/,

- ha egy ágon végighaladva a költség sem az adott főtörzs alatt, sem később nem lesz jobb, mint az addigi minimum, akkor számitása elhagyható / $\not\leftleftarrows$ jellel/.

A 4. ábra az algoritmust foglalja össze. Látható, hogy további megtakaritást tesz lehetővé a sweepelések számában, ha figyelembe vesszük, hogy az uj ág a réginek oldalága-e. Ha igen, a csomóponti elágazásig, ahol az uj ág indul, az előző ágon besweepelve mehetünk vissza és állithatjuk elő az oldalág számitásának alapjául szolgáló mátrixot.


j | cj
1 | 25
2 | 20
3 | 15
4 | 15
5 | 15
Rmin= 0.7
0.95-65
0.8-70
0.9-75
0.95-75
0.95-75
0.75-45
0.6-55
0.8-50
0.9-60
0.75-55
0.9-50
0.8-60
0.85-60
0.8-55
0.85-50
0.55-30
0.45-40
0.8-35
0.65-30
0.75-45
0.7-40
0.65-35
0.6-40
0.75-30
0.75-30
0.4-15
0.4-25
0.6-20
0.6-15
0.5-15

3. ábra


START
rendezés költség szerint
kovarianciamátrix számítása
invertálás
ágelőállítás
jó ágban fekszik
költség szerint elhagyható
előző ággal k-ig egyezik
k < előző ághossz / 2
pivotálás k-ig
visszapivotálás k-ig
pivotálás
levágható ágvég
további pivotálás
levágás
az ág költsége min.
tárolás
van még ág
függvénybecslés
STOP
Y
N

4. ábra

Az algoritmus alapján FORTRAN nyelvü interaktiv programot készitettem HwB 66/20-as gépre, amely outputján megadja az igy kiválasztott prediktor subsetet, költségét, a százalékos megtakaritást és az elhagyható változók /betegvizsgálatok/ becslésére szolgáló többváltozós lineáris regressziós egyenleteket.

Megjegyzendő, hogy a feladat megoldására a Furnival és Wilson által 1974-ben leirt "leaps and bounds" algoritmus, illetve a stepwise regressziós módszer /amelyek több statisztikai programcsomagban megtalálhatók, igy többek között az BMDP-ben is/ nem alkalmasak, mert a jóslandó változókat előre rögziteni kell és a prediktor változók kiválasztásában a költség nem játszik szerepet.