KLASSIFITSEERIMINE

X(N,M), i=1,...,N; j=1,...,M.
Rühmitamine; Tüpologiseerimine;
Taksonomeetria; Klasteriseerimine (KA)

<u>Klasteranalüüs</u> - arvutuslike protseduuride kogum, mida kasutatakse klassifikatsiooni loomisel.

***Klaster* - sarnaste objektide kogum.**

<u>Kasutatakse</u>:
- **tüpoloogiate ja klassifikatsioonide loomiseks,**
- **objektide rühmitusskeemide uurimiseks,**
- **hüpoteeside tekitamiseks uuritavate andmete kohta,**
- **hüpoteeside kontrollimiseks - kas saadud klastrid vastavad ootustele.**

<u>Etapid</u>.
1. Objektide väljavõtt
2. Tunnuste valik
3. Sobiva mõõdu valik
4. Klasteriseerimismeetodi valik, kasutamine
5. Klastrite sisulise sobivuse kontrollimine

<u>Ohud</u>.
1. Paljud KA meetodid on heuristilised, s.t. ei oma piisavat statistilist põhjendatust.
2. Paljud KA meetodid kannavad endas selle ainevaldkonna spetsiifikat, kus nad välja töötatud on.
3. Erinevad KA meetodid annavad tavaliselt samal andmehulgal erinevad lahendid.

4. KA eesmärgiks on andmehulga sisemise struktuuri leidmine. Samal ajal KA püüab andmehulgal kehtestada struktuuri. See aga ei pruugi kattuda reaalse struktuuriga.

<u>Lähedusmõõdud</u>.
1. Korrelatsioonikordaja
2. Kaugus
3. Assotsiatiivsuskordajad

1.Korrelatsioonikordaja

$r_{ik} = \{_j\sum(X_{ij} - M_i)(X_{kj} - M_k)\}/\sqrt{}_j\sum(X_{ij} - M_i)^2\ _j\sum(X_{kj} - M_k)^2$,
kus j=1,..., M, i,k=1,...,N,

X_{ij} on j-nda tunnuse väärtus i-ndal objektil,
M_i on i-nda objekti kõikide tunnuste väärtuste aritmeetiline keskmine.

Omadused (<u>+</u>).

<u>+</u> r_{jk} on tundlik objekti kuju suhtes üle tunnuste, s.t. $r_{jk} = 1{,}0 \neq> O_j = O_k$

- M_i arvutatakse eri tüüpi tunnustest üle i-nda objekti. Mis tähendus on siin keskväärtusel?

+ heaks mõõduks, kui tuleb eirata nihet eri objektide tunnuste väärtustes (mida suurem väärtus, seda lähedasema kujuga objektid).

NB! Mida suurem r_{jk}, seda lähedasemad objektid

2. *Kaugus*

Kaks objekti identsed, kui neid kirjeldavatel tunnustel ühesugused väärtused.

Eukleidiline kaugus $d_{ik} = \sqrt{}_j\sum(X_{ij} - X_{kj})^2$, j=1,...,M

Manhattani kaugus $d_{ik} = {}_j\sum|X_{ij} - X_{kj}|$

Minkowski meetrika $d_{ik} = \{{}_j\sum(X_{ij} - X_{kj})^p\}^{1/p}$

$d_{ik} = \max_j |X_{ij} - X_{kj}|$

Hammingu kaugus = kokkulangemiste arv

***Omadused* (+).**

- lähedushinnang sõltub nihetest tunnuste väärtustes eri objektidel (mida suurem väärtus, seda mõjusam tunnus)

- eri tunnused mõõdetud eri skaaladel

NB! Mida väiksem kaugus, seda lähedasemad

3. *Assotsiatiivsuskordajad*

Kasutatakse, kui tunnused dihhotoomilised

A / B	jah (1)	ei (2)
jah (1)	a	b
ei (2)	c	d

- $S_{ik} = (a + d)/(a + b + c + d)$, $0 \le S_{ik} \le 1$;
- Jaccard'i kordaja $J_{ik} = a / (a+b+c)$, $0 \le J_{ik} \le 1$;
- Gower'i kordaja

$$G_{ik} = {}_j\sum {}_jT_{ik} / {}_j\sum {}_jW_{ik}\ ,\ j=1,...,M,\ \ i,k=1,...,N.$$

Näide.

$Objekt_i$	1	1	0	0		
$Objekt_k$	1	0	1	0		
					$\sum$	
${}_jT_{ik}$	1	0	0	0	1	(AND)
${}_jW_{ik}$	1	1	1	0	3	(OR)

$G_{ik} = 1/3$

***Probleemid*.**

a) korrelatsioon

Tegelikel andmetel

r_{ik}	A	B	C	D
A	*	0,77	0,70	0,74
B	(3)	*	0,73	0,78
C	(6)	(5)	*	0,94
D	(4)	(2)	(1)	*

Standartiseeritud andmetel

r_{ik}	A	B	C	D
A	*	0,60	0,28	0,43
B	(2)	*	0,37	0,58
C	(6)	(5)	*	0,80
D	(4)	(3)	(1)	*

b) Eukleidiline kaugus

Tegelikel andmetel

r_{ik}	A	B	C	D
A	*	266	732	736
B	(2)	*	532	465
C	(5)	(4)	*	144
D	(6)	(3)	(1)	*

Standartiseeritud andmed

r_{ik}	A	B	C	D
A	*	0,70	2,57	2,07
B	(1)	*	2,14	1,31
C	(6)	(5)	*	0,87
D	(4)	(3)	(2)	*

Klastrite omadused

- **tihedus**
- **dispersioon**
- **mõõtmed**
- **kuju**
- **eristatavus**

KA meetodid

Rühmitatakse:

1. Hierarhilised ühendavad (HÜ) meetodid
2. “ jagavad “
3. Iteratiivsed meetodid
4. Faktormeetodid
5. Varia

NB! Erinevad meetodid võivad anda väga erinevaid tulemusi.

1.Hierarhilised ühendavad meetodid

Algseis - iga objekti vaadeldakse iseseisva klastrina.

- **Otsitakse kõige lähemad objektid;**
- **Uus kandidaat ühendatakse klastriga, kui talle kõige lähem objekt sisaldub selles klastris.**

HÜ meetodid jagatakse:

- **üksikseose meetod - objekt ühendatakse klastriga, kui klastris leidub vähemalt üks objekt sama kaugusega;**
- **täisseoste meetod - objekt ühendatakse klastriga, kui ta lähedus selle klastri kõigi objektidega $\geq$ LÄVI;**
- **keskmiste seoste meetod - arvutatakse objekti keskmine kaugus klastri objektidest. Kui see $\geq$ LÄVI, siis ühendatakse.**

Omadused (+):
- kauguste maatriks (N*N),
- väga tundlik objektide järjekorra suhtes,
+ iga objekti vaadeldakse 1 kord (N-1 sammu).

2. Hierarhilised jagavad meetodid

Algseis - kõik objektid kuuluvad ühte klastrisse.

3. Iteratiivsed meetodid

S1. Objektid jagatakse etteantud arvuga klastriks. Igale klastrile arvutatakse nn kese;
S2. Objektide ümberpaigutamine: objekt ühendatakse klastriga, mille suhtes tal vähim kaugus (keskme suhtes);
S3. Saadud klastritele arvutatakse kese. Mine S2
Samme S2, S3 korratakse seni, kuni tekivad püsivad klastrid.

Omadused (+):
+ töötavad lähteandmetel, s.t. ei vaja kauguste maatriksit,
+ objektide ümberpaiknemine,
- ei pruugi moodustuda stabiilseid klastreid.

4. ***Faktormeetodid***

1) leitakse korrelatsioonid objektide vahel (N*N);
2) teostatakse faktoranalüüs.

Probleemid:
- mida teha objektidega, milledel tugev seos mitme faktoriga;
- kuidas määrata oluliste faktorite (klastrite) arvu

5. *Varia*

“Objekti kuhjumite” otsimine
S1. Antakse ette raadius R keskmest;
S2. Objektid jagatakse mingiks hulgaks klastriteks;
S3. Klastritele arvutatakse kese. Kui klastrite keskmete kaugus $\leq$ R, nad ühendatakse;
S4. Objektide ümberpaiknemine: kui objekti kaugus igast keskmest >R, siis moodustab ta uue klastri, vastasel korral ühendatakse ta nende klastritega, millede korral kaugus $\leq$ R. Mine S3.

Klastrite arvu määramine

- **Kui dendrogramm formeeritud, siis kust teha läbilõige (ehk mitu klastrit)?**

Subjektiivne, ei osata püstitada H_0, kuna statistika mõttes määratlemata mõiste “struktuur”.
Pole, mille suhtes võrrelda.
1) Ühendamise kauguste kumeruspunktid.
2) Järsk hüpe ühendamise kaugustes.
3) Erimõõdud:
- Mojena-Wishart
$Z_j < K_j$, kus
$K_j = (Z_{j+1} - Z)/S_Z$, kus
K_j - standardhälve,
Z_{j+1} - kauguse väärtus (j+1)-sel sammul klastrite ühendamisel,
Z - keskmine ühenduskaugus,
S_Z - ühenduskauguse standardhälve.

***Näide*.**

Klastrid	Z_j	K_j
4	0,46	0,47
3	0,97	1,07
2	1,93	1,71
1	3,68	2,87

Tulemuste õigsuse kontroll

Proovida teise, analoogilise valimi klasterdamist. Kui saadud

- tulemused erinevad, pole lahend õige,

- ei erine, siis ei saa väita, et lahend õige (üldkogumi suhtes).