TARTU RIIKLIK ÜLIKOOL

Matemaatilise statistika ja programmeerimise

kateeder

NOMINAALTUNNUSTE ANALÜÜSI STATISTILISEST METOODIKAST

Diplomitöö

Lugupeetud juhendajale.
autorilt
21. juunil 1972.a.

Teostaja: Matemaatikateaduskonna
V kursuse üliõpilane
Kalev P ä r n a

Juhendaja: dots. E. T i i t

TARTU 1972

SISSEJUHATUS

Matemaatilised meetodid on rakendamist leidnud väga paljudes teadusharudes. Küllalt kaua on kvantitatiivset analüüsi kasutatud ka sotsioloogias. Siin vaadeldakse teda omapärase loogilise "filtrina", mille abil on võimalik teadmiste hulgast välja eraldada põhjendatud teadmisi [29].

Vaatleme mõningaid sotsioloogiliste uurimiste probleeme, kus matemaatiline käsitlusviis on võimalik.

Tihti on teatud sotsiaalse objekti korral vaja prognoosida, millisesse etteantud objektide hulka (klassi) ta kuulub (näit.: kas antud tööline kuulub püsiva või potentsiaalselt voolava tööjõu hulka). Matemaatilises statistikas on taoline küsimus tuntud "klassifitseerimise", "identifitseerimise", "statistiliste otsustusfunktsioonide", "otsustuseeskirjade" jne. leidmise probleemina.

Objektide klassifitseerimisel on üheks oluliseks küsimuseks see, milliste karakteristikute e. tunnuste väärtusi objekti kohta on vajalik teada (st. mida mõõta), et klassifikatsioon vastaks etteantud nõuetele. Sellised nõuded fikseeritakse tavaliselt ebaõige klassifitseerimise lubatud protsendina või seejuures tekkivate "kadude" summana. Nimetame seda ülesannet informatiivse tunnuste süsteemi leidmise ülesandeks.

Mitmet laadi sotsioloogilistes uurimustes on kasulik etteantud objektide hulk jaotada teatud tunnuste alusel ho-

mogeensetesse rühmadesse (s.t. rühma sees on objektid üksteisele võimalikult "lähedal", objektid erinevatest rühmadest aga üksteisest võimalikult "kaugel"). Nii näiteks võib rühmitada elukutsed gruppidesse, nii et grupi sees on "ühetüübilised" elukutsed. Selline ülesanne kannab "taksonoomia", "stratifikatsiooni", "lähteklassifikatsiooni leidmise" jne. nimetust.

Seni nimetatud probleemide ringiga kui tervikuga tegeleb viimasel ajal kiiresti arenev tehnilise küberneetika haru kujundite eristamine.

On iseloomulik, et kujundite eristamise erinevad meetodid on põhiliselt välja töötatud kvantitatiivsete, mõningal määral ka dihhotoomiliste tunnuste jaoks. Sotsioloogilistes uurimustes aga, nagu märgivad mitmed autorid [4,12,27], kasvab kvalitatiivsete ehk nominaalsete tunnuste osatähtsus. Seda väidet põhjendatakse asjaoluga, et paljusid sotsiaalseid objekte pole võimalik adekvaatselt mõõta ilma nominaalsete tunnuste abita. Seepärast vaadeldakse käesolevas töös kujundite eristamise mõningaid ülesandeid kvalitatiivsete tunnuste kasutamise korral. Püütakse analüüsida olemasolevate meetodite rakendamisvõimalusi (I, II ptk.) ja leida uusi võimalikke lähenemisi (III, IV ptk.).

Peamiseks takistuseks, mis raskendab matemaatiliste meetodite kasutamist sotsiaalsetes teadustes, sealhulgas ka sotsioloogias, on mõõtmisprobleem [29] . Antud töös käsitletakse mõningaid sekundaarse mõõtmisega seotud küsimusi, kusjuures lähtutakse samuti kvalitatiivsete tunnuste kasutamisest. Sekundaarsel mõõtmisel ei leita mõõdetava karakteristiku väärtust objektil vahetult, vaid mõnede teiste tunnus-

te väärtuste põhjal sellel objektil. Esitatakse originaal - meetod ühe tihti esineva sekundaarse mõõtmise ülesande lahendamiseks (V ptk.). Osutub, et see ülesanne on teatud mõttes ekvivalentne objektide klassifitseerimise ülesandega.

Viidates tekstis käesolevale tööle, toimitakse järgmiselt: kui viidatav koht asub samas peatükis, märgitakse sulgudes ära ainult vastava punkti number, näiteks (2.); vastasel korral lisatakse ette ka peatüki number, näiteks (II, 1.).

Autor avaldab tänu dots. E. Tiidule väärtuslike nõuannete eest töö juhendamisel.

I KVALITATIIVSETE ANDMETE ISEÄRASUSED

1. Esmased kvalitatiivsed andmed

Eeldades, et kasutatakse ainult nominaalseid tunnuseid, võib sotsioloogilise uurimuse andmed esitada järgmisel kujul.

On antud lõplik objektide põhihulk (kas üldkogum või esinduslik[6] väljavõte sellest) $O=\{o_1, o_2, o_3, \dots, o_n\}$ ja nominaalsete tunnuste hulk $T=\{t_1, t_2, \dots, t_m\}$. Näiteks võivad objektideks olla indiviidid ja tunnusteks sugu, elukoht jne.. Tunnusel t_i, $(i=1,2,\dots,m)$, on v_i võimalikke väärtust, mis moodustavad hulga $V_i=\{t_{i1}, t_{i2}, \dots, t_{iv_i}\}$. Nominaaltunnus on iseloomustatud sellega, et tema väärtuste vahel pole muud suhet kui see, et iga kaks väärtust hulgast V_i on teineteisest erinevad. Eeldame (nagu tavaliselt sel juhul tehakse), et iga tunnus võib ühel suvalisel objektil omandada ühe ja ainult ühe oma võimalikest väärtustest. Siis saab igale objektile $o_j \in O$ vastavusse seada parajasti ühe korteeži $\langle o_j \rangle = \langle t_1(o_j), t_2(o_j), \dots, t_m(o_j) \rangle$, kus $t_i(o_j)$ on tunnuse t_i väärtus objektil o_j. Ühtlasi $t_i(o_j) \in V_i$. Esmased andmed on siis esitatavad $n \times m$ maatriksina mille ridadeks on hulga O elementide o_j korteežid $\langle o_j \rangle$. Maatriksi j -ndas reas ja i -ndas veerus asub seega suurus $t_i(o_j)$.

Järgnevas esitame esmaste andmete geomeetrilise interpretatsiooni.

2. Nominaaltunnuste meetriline ruum

Geomeetriliselt on võimalik esmaseid andmeid kujutada punktihulgana m -mõõtmelises ruumis. Nimetame m -mõõtmelist ruumi, mille koordinaattelgedeks on võetud tunnused $t_1, t_2, \ldots, t_m$, edaspidi m -mõõtmeliseks tunnuste ruumiks T ehk lihtsalt tunnuste ruumiks T (kui pole vaja rõhutada dimensioonide arvu). Igale objektile o_j hulgast O saab vastavusse seada punkti tunnuste ruumis. See punkt on üheselt määratud koordinaatidega $(t_1(o_j), t_2(o_j), \ldots, t_m(o_j))$. Kui hulga kahele objektile vastab tunnuste ruumis üks ja see sama punkt, siis nimetame neid objekte ekvivalentseteks. Samastades ekvivalentsed objektid hulgas O, saame isomorfse vastavuse hulga O ja vastava punktihulga vahel tunnuste ruumis. Seetõttu võime edaspidi termini "objekt o_j" asemel kasutada ka terminit "punkt o_j".

Järgmiste probleemide seisukohalt pakub huvi küsimus, kas tunnuste ruumi T on võimalik muuta meetriliseks ruumiks.

Kvantitatiivsete tunnuste korral saab meetrika sisse tuua näiteks eukleidilise kauguse abil. Selline lahendus ei ole võimalik aga nominaaltunnuste juures, kuna eukleidiline kaugus eeldab tunnuste reaalarvulisi väärtusi.

Lähtudes nominaaltunnuste iseärasustest, defineerime kahe punkti vahelise kauguse järgmiselt: vabalt võetud punktide o_j ja o_k vaheliseks kauguseks nimetame täisarvu $h(j,k)$, mis võrdub punktide o_j ja o_k erinevate koordinaatide arvuga. (⊛)

Erijuhul, kui kõik tunnused on dihhotoomilised (s.t.

$v_i = 2$, $i = 1,2,\dots, m)$, saame nn. Hamming'i kauguse [31].

Osutub, et nii defineeritud kaugus rahuldab kolme kauguse aksioomi.

1° $m \geqslant n(j,K) \geqslant 0$. Seejuures $n(j,K) = 0$ parajasti siis, kui $o_j = o_K$ s.t. $t_i(o_j) = t_i(o_K)$ iga $i = 1,2,\dots, m$ korral (o_j ja o_K ekvivalentsed) ja $n(j,K) = m$ parajasti siis, kui $t_i(o_j) \neq t_i(o_K)$ iga $i = 1,2,\dots, m$ korral.

Vajaduse korral võime kaugust normeerida. Normeeritud kaugust tähistame $n'(j,K)$ ja see on määratud nii: $n'(j,K) = \frac{n(j,K)}{m}$

Siis kehtib seos $1 \geqslant n'(j,K) \geqslant 0$.

2° $n(j,K) = n(K,j)$.

3° $n(j,K) \leqslant n(j,\ell) + n(\ell,K)$.

Omaduste 1° ja 2° kehtivus on ilmne.

Näitame võrratuse 3° kehtivuse.

Tähistame lihtsuse mõttes $o_j = (c_1, c_2, \dots, c_m)$, $o_\ell = (a_1, a_2, \dots, a_m)$ ja $o_K = (b_1, b_2, \dots, b_m)$. Olgu $n(j,K) = \gamma$, $n(j,\ell) = \beta$ ja $n(\ell,K) = \alpha$. On vaja näidata, et $\gamma \leq \alpha + \beta$.

Üldsust kitsendamata (vajaduse korral tunnuseid ümber nummerdades) võime siis eeldada, et

$a_1 \neq b_1, \dots, a_\alpha \neq b_\alpha, a_{\alpha+1} = b_{\alpha+1}, \dots, a_m = b_m$;

$a_1 = c_1, \dots, a_p = c_p, a_{p+1} \neq c_{p+1}, \dots a_{p+\beta} \neq c_{p+\beta}, a_{p+\beta+1} = c_{p+\beta+1}, \dots, a_m = c_m$,

kus p on mingi arv, mille korral $p \leq \alpha \leq p+\beta \leq m$.

Toodud kahest reast järeldub, et

$c_1 \neq b_1, \dots, c_p \neq b_p, c_{p+1} = b_{p+1}, \dots, c_{p+r} = b_{p+r}, c_{p+r+1} \neq b_{p+r+1}, \dots,$

$c_{p+\beta} \neq b_{p+\beta}, c_{p+\beta+1} = b_{p+\beta+1}, \dots, c_m = b_m$,

kus r on teatud arv, mis rahuldab seost $p \leq p+r \leq \alpha$. Leiame nüüd o_j ja o_K mittevõrdsete komponentide arvu γ.

$$\gamma = p + \beta - r \leq p + \beta \leqslant \alpha + \beta$$ m.o.t.t.

Järelikult võib nominaalsete tunnuste ruumi vaadelda

kui meetrilist ruumi seosega $(\otimes)$ defineeritud kauguse suhtes.

3. Kvantitatiivsete andmete töötlemise meetodite ülevõtmise võimalused

Nagu üldiselt teada, on enamik matemaatilise statistika meetodeid orienteeritud kvantitatiivsete tunnuste kasutamisele. See kehtib ka matemaatiliste meetodite kohta sotsioloogilise informatsiooni töötlemisel. Seega on mõtet vaadelda nende meetodite ülevõtmise võimalust kvalitatiivsete andmete töötlemisel.

Mitmesugused meetodid kvantitatiivsete tunnuste jaoks opereerivad selliste statistikutega nagu tsentraalset tendentsi näitavad suurused ehk paiknemise karakteristikud: aritmeetiline keskmine, mediaan; hajuvuse karakteristikud: dispersioon, standardhälve, kvantiilid jt.; seost iseloomustajad: korrelatsioonikordaja, korrelatsiooni suhe, astakkorrelatsiooni kordaja jms.. Siinkohal on tähtis, et sama tüüpi statistikud on olemas ka nominaaltunnuste juures. Need on vastavalt mood, entroopia $\mathcal{H}$ ja ülekanduv informatsioon $\mathcal{T}$. Paljudes kvantitatiivsete andmete meetoditeS kasutatakse ka punktidevahelise "kauguse" mõistet (tavaliselt eukleidiline kaugus, kasutatakse ka nn. divergentsi jms.). Et kauguse mõiste on defineeritud ka nominaaltunnuste jaoks $(\otimes)$, on ka kõik need meetodid rakendatavad nominaaltunnuste analüüsimisel. Olenemata mõõtmiskaala tüübist peegeldavad kõik nimetatud suurused mõõdetavate objektide tegelikku struktuuri, mis mõõtmisest ei sõltu.

Seega on olemas teatud alus kasutada kvantitatiivsete

andmete meetodite vastavaid analooge nominaaltunnuste juures. Ilmselt on otstarbekas seejuures arvestada järgmisi asjaolusid:

1) ülevõetava meetodi algoritmi operatsioonide keerukus ja arv,
2) esmastest andmetest arvutatavate statistikute keerukus ja arv meetodis.

Eesmärk on, et ülevõtmisel meetod kaotaks võimalikult vähe oma efektiivsusest (efektiivsus üldkeelelises, mitte statistilises tähenduses). Ei ole mõtet kasutada selliste algoritmide analooge, mille efektiivsuse kadu on suur. Kõik meetodid on loodud kasutatavate statistikute eripära silmas pidades ja on loomulik väita, et mida keerukam meetod on, seda "orgaanilisemalt" on ta seotud statistikutega, mida ta kasutab ja seda suurem on efektiivsuse kadu teisi samatüü - bilisi suurusi kasutades. Seda aluseks võttes, on mõtet üle võtta kvantitatiivsete andmete juurest lihtsad, kergelt interpreteeritavate tulemustega meetodid. Nendes kasutatavate statistikute arv ei tohiks olla suurem kahest, soovitav on üks. Kasutades rohkem kui ühte erinevat statistikut, mis arvutatakse esmaste andmete põhjal, peab meetodi algoritm olema lihtne ja ülevaatlik. Lõplikult lahendab ülevõtmise küsimuse vastava meetodi praktiline kasutamine normaaltun - nuste juures.

On ka teistsuguseid lähenemisi kvalitatiivsete andmete töötlemise küsimusele. Perspektiivikas näib olevat nominaaltunnuste "orienteerimise" meetod [12], mis võimaldab nominaaltunnuse vä ärtuste hulga muuta järjes-

tatud hulgaks. Järjestamine toimub baastunnuse suhtes, mis omab järjestatud skaalat. Sellel nn. orienteerimisel nõutakse, et oleks maksimiseeritud teatud seose näitaja orienteeritava tunnuse ja baastunnuse vahel (antud indiviidide hulgal).

Erinevad lähenemisviisid aitavad luua aluse spetsiaalsele statistilisele aparaadile nominaaltunnuste jaoks. Kvantitatiivsete andmete töötlemise metoodika ärakasutamine on seejuures üks võimalus. Põhirõhk tuleb asetada aga uutele, nominaaltunnustest lähtuvatele meetoditele.

II KUJUNDITE ERISTAMISE ÜLESANNETE PÕHITÜÜBID. OLEMASOLEVATE ALGORITMIDE SOBIVUS NOMINAALTUNNUSTE KORRAL

1. Põhimõisted ja ülesannete põhitüübid

Määratleme järgnevad mõisted töö [31] eeskujul.

Kujund on tunnuste ruumist T teatud eesmärgil välja eraldatud homogeenne piirkond. Ta on tõtud kuju ja suurusega punktihulk, mis ühendab endas "ühetüübilised", "sarnased" (ruumi tunnuste järgi) objektid. Olenevalt eesmärgist võime valida ühe või teistsuguse tunnuste hulga T, samuti mitmesuguse kujuga piirid homogeensetele piirkondadele. Selle tulemusena võib kahte objekti asetada kord samasse, kord erinevatesse kujunditesse. Tähistame kujundite hulga, mida on vaja omavahel eristada, tähega $S = \{S_1, S_2, \ldots, S_K\}$.

Seejuures kujundite eristamise ülesannetes on alati $1<K<\infty$.

Reeglid, mille järgi jaotatakse tunnuste ruum T homogeenseteks piirkondadeks, nimetatakse "otsustuseeskirjadeks" või "otsustusfunktsioonideks". Võib kasutada mitmesuguseid otsustuseeskirju. Tähistame nende hulga $D=\{\delta_1, \delta_2, \ldots, \delta_e\}$.

Pärast seda, kui ruum T on jaotatud piirkondadeks S (eristav seade on konstrueeritud), võetakse mingi objekt, mille kuuluvus ühte või teise klassi pole teada. See objekt mõõdetakse tunnuste T järgi ja olenevalt sellest, millisesse piirkonda hulgast S mõõtmistulemus langes, võetakse vastu otsus antud objekti kuuluvuse kohta sellesse kujundisse.

Konkreetsetes uurimustes pole tihti kuuluvus vastavasse kujundisse teada mitte kõikide vaadeldavate objektide korral vaid ainult osa puhul nendest, s.o. teatud väljavõttel, mida nimetatakse õpperühmaks. Tavaliselt ülesanne ongi selles, et kindlaks teha, millisesse õpperühma poolt määratud klassi (kujundisse) objekt kuulub. Selle tõttu, et väljavõte võib olla põhihulga O suhtes mitteesinduslik ja et mõõdetavate tunnuste hulk T on halvasti valitud, võivad tekkida vead objektide paigutamisel klassidesse e. objektide eristamisel. Mõnikord pakub huvi neid vigu P eelnevalt hinnata. Tähistame erinevaid hindamise meetodeid $R=\{r_1, \ldots, r_g\}$.

Need neli elementi S, D, T ja P esinevad ühel või teisel kujul igas kujundite eristamise ülesandes.

Lähtudes sellest võib jagada kujundite eristamise ülesanded 4 põhitüüpi

1. $A_1 = \min N(D) \mid S, T, P = \text{const.}$

See on "lahendavate funktsioonide" e. "reeglite" või "klas-

sifitseerimise", "identifitseerimise" jne. ülesanne.

2. $A_2 = \min N(R) \mid S, T, \mathcal{D} = const.$

Tuleb hinnata kõige ökonoomsemal viisil oodatavate kaotuste suurust.

3. $A_3 = \min N(T) \mid S, \mathcal{D}, P = const.$

See ülesanne kannab "informatiivse tunnuste süsteemi leidmise" nimetust.

4. $A_4 = \min N(S) \mid T, \mathcal{D}, P = const.$

Ülesanne on tuntud "lähteklassifikatsiooni leidmise", "taksonoomia", "iseõppimise" jne. nime all.

Igas ülesandes on kolm suurust ette antud, leida tuleb ülejäänud neljas, kuid nii, et minimiseeritakse hinna (keerukuse) funktsioon N.

Sotsioloogilistes uurimustes võivad esineda kõik need ülesanded, kuid sagedamini lahendatakse neljast tüübist kolme: klassifitseerimise (A_1), informatiivse tunnuste süsteemi leidmise (A_3) ja taksonoomia (A_4) ülesandeid. Peatume nendel lähemalt, arvestades seejuures, et kasutusel on ainult nominaaltunnused.

2. Klassifitseerimisülesanne

Klassifitseerimisülesandele on üldiselt olemas 2 võimalikku lähenemist: rangelt statistiline ja heuristiline. Selgitame nende vahekorda.

Statistilisel lähenemisel kasutatakse statistilist otsustuste teooria aparaati. Eeldatakse, et üldkogumi jao-

tus ruumis T, üldkogumi jaotused klassides (kujundites?) ja klasside aprioorsed tõenäosused on teada. Samuti on ette antud vigade hindade maatriks. Klassifitseerimiseeskirjana on sel juhul tuntud nn. "Bayes'i meetod", mille lähtemõte on see, et punkt asetatakse sellisesse kujundisse, mille aposterioorne tõenäosus selles punktis on suurim [9, 32]. Enamus otsustuste teooria meetoditest nõuab, et üldkogumi jaotus oleks normaalne. Praktikas seda alati eeldada ei saa. Teiseks takistuseks on see, et üldkogumi jaotuse (milline see ka ei oleks) parameetreid tuleb tavaliselt hinnata väljavõtte põhjal, mis võib aga olla mitteesinduslik. Seega on rangelt statistiliste meetodite kasutamine praktilistes uurimustes raskendatud. Viimasel ajal on hakanud levima mitteranged, heuristilised meetodid.

Heuristilise käsitluse puhul tehakse eeldus, et õpperühm kui väljavõte üldkogumist on esinduslik (tunnuste T suhtes). Seejuures nõutakse otsustusfunktsiooni, mis oleks küllalt lihtne ja annaks samal õpperühmal minimaalse või nulliga võrduva arvu vigu. Paljud heuristilised algoritmid lähenevad väljavõtte suurenedes optimaalsele, statistilisele lahendile. Seega praktikas kahe käsitluse vahe nivelleerub.

Heuristilistest algoritmidest on tuntumad "potensiaalsete funktsioonide" [8], "naabruse funktsiooni" [33], "kuuluvuse funktsiooni" [34], "läheduse funktsiooni" [37] jt. meetodid. Osutub, et ainus, mis nimetatud meetodite hulgas kasutab esmaseid andmeid ainult üht liiki suuruste (nimelt kauguste) leidmisel ja edasi opereerib ainult nendega, on potentsiaalsete funktsioonide meetod. Seega on seda mõtet

vaadelda "ülevõtmise" suhtes.

Meetodi olemus on selles, et õpperühma iga objekti o_i korral arvutatakse välja teatud funktsioon $K(o, o_i)$, mille argument o on määratud kogu tunnuste ruumis T ja sõltub o_i-st kui parameetrist. Funktsiooni $K(o, o_i)$ nimetatakse "potentsiaalseks funktsiooniks". Jada $K(o, o_i)$, kus i muutub õpperühmal, kasutatakse ära lahendava funktsiooni $\Psi(o, o_1, o_2, \ldots)$ konstrueerimisel. Potentsiaalsete funktsioonide kuju kohta kitsendusi pole, kuid tavaliselt kasutatakse sellist funktsiooni, mis kahaneb punktide o ja o_i vahelise kauguse d kasvamisega, näiteks e^{-ad^2}. Meetod on küllalt põhjalikult läbi töötatud, laialt levinud ning nähtavasti sobib ka nominaaltunnuste korral.

Mis puutub spetsiaalselt nominaaltunnuste jaoks väljatöötatud klassifitseerimise meetoditesse, siis kirjanduses on levinud ainultüks lähenemine [27, 39]. Seal lähtutakse kahe erineva tunnuse p ja q poolt tekitatud kahe erineva objektide jaotuse $\tilde{P}$ ja $\tilde{Q}$ vahelisest kaugusest $d(\tilde{P}, \tilde{Q})$, mis arvutatakse valemist [28]

$$d(\tilde{P}, \tilde{Q}) = \frac{1}{2} \sum_{i,j=1}^{n} |p_{ij} - q_{ij}|, \qquad (\oplus)$$

kus $p_{ij} = \begin{cases} 1, & \text{kui } o_i \text{ ja } o_j \text{ on tunnuse } p \text{ väärtused võrdsed;} \\ 0, & \text{vastasel juhul.} \end{cases}$

Suurus q_{ij} arvutatakse analoogselt (tunnuse q korral). Meetodit on küllalt edukalt rakendatud konkreetsetes uurimustes [13].

Mõned uued lähenemised klassifitseerimisprobleemile esitame peatükkides III ja IV.

3. Informatiivse tunnuste süsteemi leidmine

Informatiivse tunnuste süsteemi leidmise ülesanne kerkib praktikas üles klassifitseerimise juures. Antud on kujundid, lahendavad reeglid ja lubatud kaod P. Tekib küsimus, milliste tunnuste väärtusi objektidel on vaja teada, ehk milline on tunnuste informatiivne süsteem, et neid objekte klassifitseerida antud reeglite järgi lubatud kaotuste piires.

Probleemil on kaks külge: süsteemi informatiivsuse piisavus ja tarvilikkus. Piisavus tähendab seda, et nõuded S, D ja P on täidetud, tarvilik süsteem on lihtsaim (odavaim) piisav süsteem. Tarviliku süsteemi leidmise viis sõltub sellest, kas lähtetunnused on sõltumatud või sõltuvad. Vaatleme neid olukordi eraldi.

Tähistame lähtetunnuste hulga T ja otsitava süsteemi T'.

Lähtetunnuste sõltumatuse korral leitakse eelnevalt iga üksiku tunnuse suhteline tähtsus e. "informatsiooni kaal". Sel juhul on süsteemi informatiivsus võrdne süsteemi kuuluvate tunnuste informatsiooni kaalude summaga. Et süsteem saaks vajalikult informatiivne ja oleks seejuures minimaalne, tuleb sellesse võtta küllaldane hulk suurte kaaludega tunnuseid. Üksikute tunnuste informatsiooni kaalude hindamise moodused on toodud töödes [15, 31].

Neist ühe meetodi (lähemalt v.t. [21]) teatud modifikatsioon on kasutamiskõlblik ka kvalitatiivsete tunnuste juures. Informatsiooni kaal leitakse seal entroopia mõiste abil. Eeldatakse, et on teada:

1) kujundite $S_1, S_2, \ldots, S_K$ esinemise aprioorsed tõenäosused $q_1, q_2, \ldots, q_K$,

2) kujundite esinemise aposterioorsed tõenäosused $q_{1u}, q_{2u}, \ldots$ q_{Ku} tingimusel, et vaadeldav tunnus t_i omaks väärtust t_{iu},

3) tunnuse t_i väärtuste tõenäosuste tihedused $p_{i1}, p_{i2}, \ldots, p_{iv_i}$, kus v_i on võimalike väärtuste arv. Praktilistes ülesannetes leitakse need suurused õpperühma põhjal.

Oletame, et $t_i = t_{iu}$, $(u = 1,2,\ldots,v_i)$. Siis kujundite süsteemi S entroopia $\mathcal{H}_{iu}(S)$ on võrdne [40]

$$\mathcal{H}_{iu}(S) = -\sum_{l=1}^{K} q_{lu} \log q_{lu}.$$

Süsteemi S keskmine entroopia tingimusel, et t_i väärtus on teada, on

$$\mathcal{H}_i(S) = \sum_{u=1}^{v_i} p_{iu}\, \mathcal{H}_{iu}(S).$$

Süsteemi S aprioorne entroopia (t_i väärtus pole teada) võrdub

$$\mathcal{H}_0(S) = -\sum_{l=1}^{K} q_l \log q_l.$$

Tunnuse t_i informatsiooni kaaluks võib võtta suuruse

$$\gamma_i = \mathcal{H}_0(S) - \mathcal{H}_i(S)$$

s.o. kujundite süsteemi S entroopiate vahe enne ja pärast tunnuse t_i kasutamist.

Märgime, et tunnuse informatsiooni kaal on suhteline suurus, mis oleneb konkreetsest tunnuste ruumist T, kus kujundid S on välja eraldatud ja lahendavast reeglist δ, mille järgi seda tehakse.

Kui on vaja m tunnuse hulgas leida l tunnusest koosnev võimalikult informatiivne süsteem, siis pärast suuruste γ_i, $(i = 1,2,\ldots,m)$, leidmist valime välja l tunnust, mille informatsiooni kaalud on suurimad.

Sõltuvate tunnuste korral pole see moodus kasutatav.

On kindlaks tehtud [36] , et olenevalt sõltuvuse iseloomust võib süsteemi informatiivsus olla nii suurem kui ka väiksem temassekuuluvate tunnuste informatiivsuste summast. Kui m ja ℓ on väikesed arvud, siis võib ℓ tunnust m -st valida kõikide C_m^ℓ kombinatsioonide hulgast. Kuid tavaliselt esinevate m ja ℓ korral pole see mõeldav suure arvutuste mahu pärast. Seepärast on loodud mitmeid heuristilisi algoritme üleminekuks m tunnuselt ℓ tunnusele. Vaatleme neist kahte. Lühiduse mõttes tähistame esimese algoritmi [5] tähega A, teise [11] tähega B.

Algoritmi A korral võetakse lähtesüsteemist T ükshaaval kõik tunnused. Igale tunnusele leitakse teatud kriteeriumi J järgi tema informatiivsus ning valitakse välja kõige suurema informatiivsusega tunnus. Seejärel valitakse ülejäänute hulgast teine tunnus, mis koos esimesega moodustab kõige informatiivsema kahetunnuselise alamsüsteemi. Edasi leitakse ülejäänud $m-2$ tunnuse hulgast selline, mille abil saab moodustada kolmest tunnusest koosneva informatiivse süsteemi jne.. Niiviisi koostatakse lähtesüsteemi ℓ- tunnuseline alamsüsteem.

Teoreetiliselt on näidatud [25] , et algoritmid A ja B ei anna enamikul juhtudel rahuldavaid tulemusi. Optimaalsetele lähedased tulemused annab nn. "juhusliku otsimise meetod adaptatsiooniga" [24], mis on tuntud Monte-Carlo meetodi täiustus.

Konkreetsetes uurimustes on sageli vaja teada, millised küljed, karakteristikud on antud objektide või nähtuste tüübi jaoks kõige iseloomulikumad ja olulisemad. Siin formuleerime ülesande nii: antud on klass objekte (kujund),

leida tunnuste informatiivsused selle klassi suhtes. Dihhotoomiliste tunnuste jaoks on huvitav lahendus probleemile antud töödes [1, 15] . Hiljem anname töös [15] kasutatavale meetodile üldistuse nominaaltunnuste jaoks (IV, 2).

4. Taksonoomia

Sageli kerkib sotsioloogilistes uurimustes üles taksonoomia ülesanne ehk vajadus jaotada etteantud objektide või nähtuste hulk O antud tunnuste T alusel homogeenseteks (eelnevalt defineeritud "läheduse", "sarnasuse" jne. kriteeriumi δ järgi) rühmadeks. Teatavasti igasuguse rühmitamisega kaasneb alati individuaalsuste väiksem või suurem kaotus e. informatsiooni kadu. Taksonoomia ülesandes nõutakse, et see kadu ei tohi ületada etteantud suurust P. Säiliva informatsiooni hulk peab olema küllaldane konkreetseks eesmärgiks.

Mitmesugused taksonoomia algoritmid on esitatud töödes [1,10,16,17,18,19,31,41]. Vaatleme selliseid taksonoomia algoritme, mis on mõeldavad kasutamiseks nominaaltunnuste või dihhotoomiliste tunnuste korral.

Kõige lihtsam lahendusalgoritm on antud Bonneri poolt [1] . See nn. "maskide" meetod on algselt mõeldud dihhotoomiliste tunnuste jaoks. Esitame tema üldistuse nominaalsetele tunnustele. Lähteobjektide hulgast O võetakse vabalt üks element o_{i_1} ja tähistatakse ta "mask 1"-ga. Võetakse teine elemant o_{i_2} ja leitakse kaugus $h(i_1, i_2)$ (algvariandis Hamming'i

kaugus). Kui $h(i_1,i_2) \leq N$, kus N on etteantud täisarv e. "lävi", siis loetakse objektid o_{i_1} ja o_{i_2} "sarnasteks" ja kuuluvaks ühte kujundisse S_1. Kui $h(i_1,i_2) > N$, siis objekt o_{i_2} arvatakse kujundisse S_2 ja ta tähistatakse kui "mask 2". Kolmandat objekti o_{i_3} võrreldakse kõigepealt "mask 1"-ga. Kui $h(i_1,i_3) \leq N$, siis objekt o_{i_3} asetatakse kujundisse S_1, vastasel juhul võrreldakse teda "mask 2"-ga (kui see olemas on) ja asetatakse $h(i_2,i_3) \leq N$ korral kujundisse S_2 ning $h(i_2,i_3) > N$ korral uude kujundisse S_3. Kui "mask 2"-te veel pole, siis tähistatakse o_{i_3} "mask 2"-ga ja loetakse kuuluvaks kujundisse S_2. Kui o_{i_3} ei kuulu kujundisse S_1 ega S_2, tähistatakse ta "maskiga 3". Üldiselt, j-ndal sammul võetakse objekt o_{i_j} ja proovitakse, millisesse kujundisse, alates esimesest, ta kuulub. Kui ta ei kuulu neist mitte ühesse, tähistatakse ta uue "maskiga" ja arvatakse uude kujundisse. Algoritmi puuduseks on see, et tulemus sõltub oluliselt objektide võtmise järjekorrast.

Töös [17] on toodud taksonoomia algoritm, mis on mõeldud samuti dihhotoomiliste tunnuste jaoks.

Teatud mõttes optimaalse objektide jaotamise kujunditesse saab Šlezingeri meetodil [41]. Kujundite arv k on ette antud. Taksonoomia efektiivsust hinnatakse summaarsete kaotuste R järgi, mis arvutatakse valemist

$$R = \sum_{l=1}^{K} q_l \int_{S_l} \int_{S_l} S(o_i,o_j)\, p(o_i|l)\, p(o_j|l)\, d\lambda(o_i)\, d\lambda(o_j).$$

Tähistused on järgmised:

q_l - kujundi S_l aprioorne tõenäosus,

$p(o_i|l)$ - objekti o_i tõenäosuse tihedus tingimusel, et ta kuulub kujundisse S_l;

$S(o_i,o_j)$ - kaotus, kui o_i ja o_j langevad samasse klassi;

$\lambda(o_i)$ - teatav mõõt.

Ülesanne on leida kujundid $S_1, \ldots, S_K$, et $R \to \min$.

On võimalik anda analoogne mudel ka nominaaltunnuste jaoks. Sel juhul

$$R = \sum_{l=1}^{K} q_l \sum_{i=1}^{n} \sum_{j=1}^{n} S(o_i, o_j)\, p(o_i|l)\, p(o_j|l).$$

Võttes $S(o_i, o_j) = n(i,j)$ jäävad tundmatuteks suurused q_l, $(l = 1,2,\ldots,K)$ ja $p(o_i|l)$, $(i = 1,2,\ldots,n\ ;\ l = 1,2,\ldots,K)$, kokku $n \times K + K = K(n+1)$ tundmatut. Et R peab olema minimaalne, siis nõuame, et kõik tema osatuletised tundmatute järgi on võrdsed nulliga. Saame mittelineaarse võrrandsüsteemi

$$\begin{cases} \dfrac{\partial R}{\partial q_l} = 0 \quad , \quad (l = 1, \ldots, K), \\ \dfrac{\partial R}{\partial p(o_i|l)} = 0 \quad , \quad (i = 1, \ldots, n\ ;\ l = 1,2,\ldots,K), \end{cases}$$

mida saab lahendada tuntud meetoditega [7].

Nagu näha, saadakse toodud meetodite puhul kujundid küllalt formaalsete otsustuseeskirjade abil. On läbi viidud uurimusi [19], et selgitada, milliseid otsustusfunktsioone ("läheduse" kriteeriume) kasutab inimene, kui talle on esitatud objektide hulk, mis tuleb rühmitada "lähedastest" objektidest koosnevatesse gruppidesse. Eksperimentide tulemused näitavad, et inimene kasutab suhteliselt lihtsaid (lineaarseid) otsustusreegleid.

On loodud mitmed algoritmid, mis püüavad modelleerida "loomulikku" s.o. inimesele omast taksonoomiat (vt.näit. [31]). Nominaaltunnuste juures sobivad kasutamiseks järgmised kaks algoritmi, mis on toodud töös [16].

Esimese meetodi puhul on kujundite arv K ette antud. Defineeritakse potentsiaalne funktsioon

$$\mathcal{K}(C,D) = \frac{1}{N_C \cdot N_D} \sum_{o_i \in C} \sum_{o_j \in D} \mathcal{K}(o_i, o_j),$$

kus C ja D on hulga O mingid alamhulgad, N_C ja N_D on elementide arv vastavalt hulkades C ja D ning $\mathcal{K}(o_i, o_j)$ on potentsiaalne funktsioon, mis kahaneb punktide o_i ja o_j vahelise kauguse $r(i,j)$ kasvades (2.). Suurus $\mathcal{K}(C,D)$ iseloomustab kahe lõpliku hulga C ja D "lähedust".

Taksonoomia seisneb hulga O elementide järk-järgulises ühendamises rühmadesse (klassidesse). Esmalt loetakse iga element $o_1, o_2, \ldots, o_n$ omaette klassiks vastavalt $A_1, A_2, \ldots, A_n$. Seejärel ühendatakse klassid A_i ja A_j, mille korral

$$\mathcal{K}(A_i, A_j) = \max_{\substack{p,q \\ p \neq q}} \mathcal{K}(A_p, A_q).$$

Ühendamist korratakse $n-K$ korda, mille tulemusena saame K klassist (kujundist) koosneva süsteemi. Meetodit saab kiirendada objektide lähteklassifikatsiooni teatud parandamise abil. Kirjeldatud algoritmil on hea omadus: resultaat ei sõltu objektide vaatlemise järjekorrast.

Teise meetodi korral pole klasside arv K varem teada. Teatud mõttes optimaalne K väärtus leitakse algoritmi realiseerimisel. Defineeritakse mõõt

$$\mathcal{K}(D,D) = \frac{2}{N_D(N_D-1)} \sum_{i=1}^{N_D-1} \sum_{j>i} \mathcal{K}(o_i, o_j),$$

(tähistused analoogsed varasemaga), mis iseloomustab lõpliku hulga D "kompaktsust". Seejärel leitakse suurused

$$J_1(m) = \frac{1}{m} \sum_{i=1}^{m} \mathcal{K}(A_i, A_i)$$

ja

$$J_2(m) = \frac{2}{m(m-1)} \sum_{i=1}^{m-1} \sum_{j>1} \mathcal{K}(A_i, A_j)$$

, kus hulgad A_i, $(i = 1,2,\ldots,m)$,

on saadud $k-m$ sammu järel eelmist algoritmi kasutades. Need suurused iseloomustavad vastavalt hulkade keskmist "kompaktsust" ja erinevate hulkade keskmist "lähedust". Optimaalseks loetakse sellist klasside arvu m, mille korral vahe

$$J(m) = J_1(m) - J_2(m)$$

on maksimaalne.

Taksonoomia algoritme on välja töötatud ka spetsiaalselt nominaaltunnuste jaoks. Töös [35] toodud meetod seisneb samuti lähtehulga elementide järk-järgulises ühendamises gruppidesse. Selleks kasutatakse mitmesuguseid informatsioonistatistikuid ning diferentsiaal- ja integraalarvutuse aparaati. Meetod on küllalt komplitseeritud.

Teisel lähenemisel [27] kasutatakse tunnuste poolt määratud jaotuste vahelise kauguse mõistet ⟨⊛⟩.

Lõpuks märgime, et iga taksonoomia on kokkuvõttes subjektiivne. Lähtudes uurimise eesmärgist määrab spetsialist tunnuste ruumi T, mille suhtes rühmitamine toimub. Lõpptulemus sõltub oluliselt ka "läheduse" kriteeriumist ja lubatud kadudest P, mis on samuti uurimise eesmärki arvestades eelnevalt fikseeritud.

III TUNNUSTEVAHELISE "KAUGUSE" KASUTAMISE KUJUNDITE ERISTAMISEL

1. Tunnustevahelise kauguse definitsioon ja omadused

Eeldame, et esmased andmed on esitatud $n \times m$ maatriksina M, kus n on objektide arv põhihulgas O ja m on tunnuste arv hulgas T (I, 1.).

Vaatleme suvalist tunnust t_i, $(i = 1,2,\dots,m)$. Tähistame selliste objektide $o_j \in O$, mille korral $t_i(o_j) = t_{iu}$, arvu sümboliga f_{iu}. Siis tõenäosus p_{iu}, et mingi objekt hulgast O omaks tunnuse t_i väärtust t_{iu}, on võrdne

$$p_{iu} = \frac{f_{iu}}{n}, \quad (u = 1,2,\dots,v_i), \tag{1}$$

kus v_i on tunnuse t_i väärtuste arv. Seejuures $\sum_{u=1}^{v_i} p_{iu} = 1$. (Kui hulga O põhjal tehakse järeldusi üldisema kogumi kohta, siis annab valem (1) tõenäosuse nihutamata hinnangu, kuna eeldame O esinduslikkust).

Järgnevas toetume mõningatele mõistetele ja tulemustele informatsiooniteooriast [40].

Tunnuse t_i entroopiaks $\mathcal{H}(i)$ nimetatakse suurust

$$\mathcal{H}(i) = -\sum_{u=1}^{v_i} p_{iu} \log p_{iu} . \tag{2}$$

Logaritmi alus on siin suvaline, see määrab ära mõõtmise mastaabi. Seega tunnuse entroopia on tema väärtuste esinemise tõenäosuste logaritmide keskväärtus. Valemit (2) tuntakse Shannoni valemi nime all.

Analoogiliselt on tunnuse t_j, $(j \neq i)$, entroopia $\mathcal{H}(j)$ võrdne

$$\mathcal{H}(j) = -\sum_{s=1}^{v_j} p_{js} \log p_{js} , \tag{3}$$

kus p_{js} on tõenäosus, et hulgast θ vabalt võetud objekt omaks tunnuse t_j väärtust t_{js}.

Tähistame tõenäosuse, et hulga θ suvaline objekt omab tunnuse t_i väärtust t_{iu} ja tunnuse t_j väärtust t_{js}, sümboliga $p_{iu,js}$.

Tunnuste t_i ja t_j ühiseks entroopiaks $\mathcal{H}(i,j)$ nimetatakse suurust

$$\mathcal{H}(i,j) = -\sum_{u=1}^{v_i}\sum_{s=1}^{v_j} p_{iu,js} \log p_{iu,js}. \qquad (4)$$

On teada, et alati kehtivad võrratused

$$\mathcal{H}(i,j) \geq \mathcal{H}(i), \quad \mathcal{H}(i,j) \geq \mathcal{H}(j), \qquad (5)$$

kusjuures võrdused saavutatakse juhul, kui tunnused t_i ja t_j on üksüheses sõltuvuses (s.o. funktsionaalse sõltuvuse tüüp, kus tunnuste väärtuste vahel on üksühene vastavus). Samuti kehtib alati võrratus

$$\mathcal{H}(i,j) \leq \mathcal{H}(i) + \mathcal{H}(j), \qquad (6)$$

kus võrdus on õige juhul, kui tunnused t_i ja t_j on sõltumatud (seda on lihtne näidata, tehes võrduses (4) asenduse $p_{iu,js} = p_{iu} \cdot p_{js}$).

Vahet $$\mathcal{T}(i,j) = \mathcal{H}(i) + \mathcal{H}(j) - \mathcal{H}(i,j) \qquad (7)$$

nimetatakse tunnuste t_i ja t_j ülekantavaks informatsiooniks (või seesmiste kitsenduste mõõduks). Nimetus tuleb asjaolust, et $\mathcal{T}(i,j)$ näitub informatsiooni hulka, mille saame ühe tunnuse kohta juurde, kui on teada teise tunnuse väärtus. Viimase asjaolu tõttu võib ülekantavat informatsiooni vaadelda kui tunnuste t_i ja t_j vahelise seose teatavat mõõtu. Nagu kõik statistilised seose näitajad, nii ka $\mathcal{T}(i,j)$ iseloomustab ainult teatud tüüpi seose tugevust. Informatsiooniteooriast on teada, et selleks seose tüübiks $\mathcal{T}(i,j)$ puhul on üksüheline vastavus tunnuste t_i ja t_j väärtuste vahel [23].

Seosest (7) on vahetult näha, et $\mathcal{T}(i,j)$ on sümmeetri -

line tunnuste t_i ja t_j suhtes, s.t.

$$T(i,j) = T(j,i) \quad . \tag{8}$$

Võrratuse (6) põhjal

$$T(i,j) \geq 0 \quad . \tag{9}$$

ja võrdus saavutatakse siis, kui tunnused t_i ja t_j on sõltumatud. Võrratustest (5) järeldub, et

$$T(i,j) \leq H(i) \;,\; T(i,j) \leq H(j) \;, \tag{10}$$

kusjuures võrdus kehtib juhul, kui tunnused t_i ja t_j on üksüheselt sõltuvad. Seoste (9) ja (10) põhjal võib kirjutada, et

$$0 \leq T(i,j) \leq \max\,(H(i), H(j)) \quad . \tag{11}$$

Seejuures

$$T(i,j) = \max\,(H(i), H(j)) = H(i) = H(j) = H(i,j), \tag{12}$$

kui t_i ja t_j on üksüheselt sõltuvad. Nagu seosest (11) on näha, sõltub ülekantava informatsiooni $T(i,j)$ muutumise piirkond tunnuste t_i ja t_j entroopiatest. Sellepärast pole võimalik ülekantavat informatsiooni kasutada tunnustevaheliste seoste omavaheliseks võrdlemiseks. Et seda teha, peab seose näitaja olema normeeritud.

Käesoleval juhul normeerime ülekantava informatsiooni järgmiselt:

$$T'(i,j) = \frac{2\,T(i,j)}{H(i) + H(j)} \quad . \tag{13}$$

Suurus $T'(i,j)$ näitab, kui palju informatsiooni, võrreldes tunnuste entroopiate keskmisega, saame ühe tunnuse kohta juurde, kui on teada teise tunnuse väärtus. Kuna $T(i,j)$ on üksühese vastavuse mõõt tunnuste t_i ja t_j väärtuste vahel, siis $T'(i,j)$ on selle normeeritud mõõt.

Võrd ustst (8) järeldub, et ka normeeritud ülekantav informatsioon $T'(i,j)$ on sümmeetriline t_i ja t_j suhtes, st.

$$\tilde{T}'(i,j) = T'(j,i) \quad . \tag{14}$$

Võrratuste (9) ja (10) abil saame, et

$$T'(i,j)\begin{cases} = 1 & \text{, kus tunnused on funktsionaalselt sõltuvad,} \\ 0 < T'(i,j) < 1 & \text{, kui tunnused on statistiliselt sõl-} \\ = 0 & \text{tuvad,} \\ & \text{, kui tunnused on sõltumatud.} \end{cases} \tag{15}$$

On loomulik defineerida tunnuste vaheline kaugus nii, et ta oleks maksimaalne tunnuste sõltumatuse korral ja minimaalne, kui tunnused on funktsionaalselt sõltuvad.

Tunnuste t_i ja t_j vaheliseks kauguseks nimetame suurust $d(i,j)$, mis võrdub

$$d(i,j) = 1 - T'(i,j) \quad . \tag{16}$$

Osutub, et seosega (16) määratud kaugus rahuldab kolme meetrika aksiooni.

1^0 $0 \leq d(i,j) \leq 1$. See järeldub seostest (15) ja (16). Seejuures $d(i,j) = 0$ parajasti siis, kui tunnused t_i ja t_j on funktsionaalselt sõltuvad ja $d(i,j) = 1$ parajasti siis, kui tunnused t_i ja t_j on sõltumatud.

2^0 $d(i,j) = d(j,i)$. Kuna $T'(i,j) = T'(j,i)$, siis ka $d(i,j) = 1 - T'(i,j) = 1 - T'(j,i) = d(j,i)$.

3^0 $d(i,j) \leq d(i,k) + d(k,j)$. Kolmnurga aksioomi täidetuse tõestamine on küllalt keeruline ning äratoomiseks ülearu pikk.

Kasutame tunnustevahelise kauguse mõistet mitmet tüüpi kujundite eristamise ülesannete lahendamiseks.

2. Klassifitseerimine.

Olgu antud objektide õpperühm $O^* = \{o^*_1, o^*_2, \ldots, o^*_k\}$, mille elementidel on teada tunnuste $T = \{t_1, t_2, \ldots, t_m\}$ väärtused ja on teada ka elementide klassifikatsioon kujundite (klasside) $S = \{S_1, S_2, \ldots, S_K\}$ vahel. Ülesanne on leida otsustuseeskiri, mille järgi klassifitseerida kujunditesse $S_1, S_2, \ldots, S_K$ hulga $O \setminus O^*$ elemendid.

Toome mugavuse mõttes sisse identifitseeriva tunnuse λ mõiste. Identifitseerivaks tunnuseks nimetame tunnust λ, mille väärtus mingil objektil $o_j \in O$ on võrdne kujundi S_ℓ, kuhu see objekt kuulub, indeksiga ℓ. Seega $\lambda(o_j) = \ell$, kui $o_j \in S_\ell$. Identifitseeriva tunnuse sisuline tähendus ei ole üldiselt teada.

Meie ülesandes on teada identifitseeriva tunnuse λ väärtused õpperühma O^* elementidel ja on vaja leida λ väärtused hulga $O \setminus O^*$ elementidel.

Arvutame õpperühma O^* baasil kõikide tunnuste t_i kaugused tunnusest λ. Need kaugused iseloomustavad tunnuste t_i struktuuri õpperühmas O^*, mille elementide klassifikatsioon on antud. Kui paigutame katseliselt suvalise objekti o_j hulgast $O \setminus O^*$ mingisse kujundisse S_k, $(k = 1,2,\ldots,K)$, siis üldiselt nimetatud kaugused muutuvad. Seejuures on kauguste muutused seda suuremad, mida vähem on objekt o_j "sarnane" kujundi S_k objektidega. Võtamegi objekti o_j klassifitseerimise aluseks tunnuste t_i, $(i = 1,2,\ldots,K)$ ja λ vaheliste kauguste muutused (e. nihked), mis tekivad selle objekti katselisel paigutamisel mingisse kujundisse S_k. Seejuures klassifitseerime objekti

$o_j \in O \setminus O^*$ niisugusesse klassi S_ℓ, mille korral need nihked on minimaalsed.

Objekti o_j katselise paigutamise all klassi S_k mõistame selle klassi mingi objekti o_p^* asendamist objektiga o_j. Selline paigutamise moodus on parem kui lihtne objekti o_j lisamine klassi S_k teistele objektidele. Nimelt säilib sel juhul elementide üldarv hulgas G^* ja tunnuste nihked tekivad ainult objektide o_j ja o_p^* vahelise erinevuse tõttu, mitte aga hulga G^* objektide üldarvu muutumisest.

Klassifitseerimise algoritm on järgmine:

1. Võrdleme eristatavat objekti (s.o. objekti, mille klassilist kuuluvust tahame teada) järgemööda õpperühma kõikide objektidega ja leiame arvud ν_k, $(k = 1,2,\dots,K)$, mis näitavad objektiga o_j ekvivalentsete (I, 2.) objektide (neid võib õpperühmas esineda) arvu klassis S_k.

2. Kui arvude ν_k hulgas leidub suurim, mille tähistame ν_ℓ, $\nu_\ell = \max\limits_{1 \le k \le K} \nu_k$, siis klassifitseerime objekti o_j kujundisse S_ℓ. Lühidalt märgime seda järgmise sümboolika abil (sellist sümboolikat kasutame ka edaspidi):

$$o_j \in S_\ell = \arg \max_{S_k \in S} \nu_k$$

3. Kui arvude ν_k seas ei leidu suurimat, siis asendame järgemööda kõik õpperühma G^* objektid o_p^*, $(p = 1,2,\dots,n)$, objektiga o_j, kusjuures $s(o_j) = s(o_p^*)$.

4. Igal asendusel arvutame suurused d_{pi}, $(i = 1,2,\dots,m)$, mis võrduvad tunnuse t_i kaugusega tunnusest s hulga $G^* \cup \{o_j\} \setminus \{o_p^*\}$ baasil.

5. Arvutame tunnuste t_i individuaalsed nihked δ_{pi} valemi

$$\delta_{pi} = d_{pi} - d_i \quad , (i = 1,2,\dots,m), \qquad (17)$$

järgi, kus d_i on tunnuse t_i ja tunnuse s vaheline kaugus õp-

perühma G^* baasil.

6. Arvutatakse klassi S_k, $(k=1,2,\dots,K)$, nihe δ_k valemi

$$\delta_k = \frac{1}{n_k} \sum_{o_p^* \in S_k} \sum_{i=1}^{m} |\delta_{pi}| \qquad (18)$$

abil, kus n_k on klassi S_k kuuluvate objektide arv hulgast G^*.

7. Leitakse klass S_l, mille nihe on minimaalne, s.t.

$$S_l = \arg \min_{S_k \in S} \delta_k \qquad (19)$$

ja klassifitseeritakse objekt o_j klassi S_l.

Lühidalt võib algoritmi esitada otsustuseeskirjana $\delta = (1.,2.)$:

1. Kui leidub $\nu_l = \max_{1 \le k \le K} \nu_k$, siis $o_j \in S_l$, vastasel juhul
2. $o_j \in S_l = \arg \min_{S_k \in S} \delta_k$.

3. "Keskmine" tunnus ja taksonoomia ülesanne

Kasutades identifitseeriva tunnuse mõistet (2.), seisneb taksonoomia ülesanne identifitseeriva tunnuse väärtuse leidmises kõigi põhihulga G objektidel. Vaatleme sellist taksonoomia juhtu, kus kujundite arv K pole ette antud.

Defineerime mõned vajalikud mõisted.

Olgu g mingi tunnus, mis võib, aga ei pea kuuluma tunnuste lähtehulka T. Samuti ei pea tal olema konkreetset sisulist tähendust, ta tekib vaid hulga G mingi jaotuse kategooriate vahel, millede arv võrdub g erinevate väärtuste arvuga ja ühte kategooriasse kuuluvad üht ja sama g väärtust omavad objektid.

Tunnuse g ja tunnuste hulga T vaheliseks kauguseks ni-

meteme suurust $d(g,T)$, mis võrdub

$$d(g,T) = \frac{1}{m} \sum_{t_i \in T} d(g,i) \quad , \tag{20}$$

kus $d(g,i)$ on tunnuste g ja t_i vaheline kaugus (16).

Seega g ja T vaheline kaugus on tunnuse g ja tunnuste $t_i \in T$ vaheliste kauguste aritmeetiline keskmine. Vastavalt tunnustevahelise kauguse tähendusele, näitab suurus $1-d(g,T)$ tunnuste g ja $t_i \in T$ väärtuste üksühese vastavuse tugevust.

Tunnust q_T, mille korral

$$d(q_T,T) = \min_g d(g,T) \tag{21}$$

nimetame tunnuste süsteemi keskmiseks tunnuseks.

Seega süsteemi keskmine tunnus on selle süsteemi "lähim" tunnus (kauguse (20) mõttes). Selgitame seda mõistet lähemalt.

Keskmise tunnuse q_T konkreetne sisuline tähendus pole, nagu g sisuline tähenduski, üldiselt teada. Tunnus q_T on iseloomustatud sellega, et tema väärtuste ja süsteemi T tunnuste väärtuste vahelise üksühese seose keskmine tugevus on maksimaalne (üle kõik võimalike tunnuste g). See tähendab, et kui kaks objekti omavad üht ja sama q_T väärtust, siis on tõenäoline, et nad omavad üht ja sama väärtust ka tunnuste $t_i \in T$ juures ning vastupidi. Järelikult tunnuse q_T järgi lähedased objektid on lähedased ka tunnuste $t_i \in T$ suhtes (s. o. "tegelikult lähedased"). Kuna taksonoomia eesmärk ongi jagada objektide hulk etteantud tunnuste T järgi omavahel "lähedastest" objektidest koosnevatesse kujunditesse, siis on otstarbekas võtta identifitseerivaks tunnuseks süsteemi T keskmine tunnus q_T. Taksonoomia ülesande lahend on seega:

$$1(o_j) = q_T(o_j) \quad , \quad (j = 1,2,\ldots,m), \tag{22}$$

kus q_T on süsteemi T keskmine tunnus. Lahend on optimaalne selles mõttes, et informatsiooni kadu rühmitamisel on võrdne informatsiooni kaoga tunnuste $t_i \in T$ väärtuste "taastamisel" rühmituste põhjal.

Kui identifitseerivale tunnusele $\wedge$ tahetakse anda sisulist tähendust, siis tuleb võtta selleks tunnus q'_T, mis leitakse seosest

$$q'_T = \arg \min_{t_i \in T} d(i,T) . \qquad (23)$$

Nimetame selles punktis esitatud taksonoomia algoritmi keskmise tunnuse meetodiks. Meetod erineb olemasolevatest taksonoomia meetoditest selle poolest, et pole vaja teada kaugusi hulga O punktide vahel. Punktidevaheliste kauguste leidmisel tekkivad põhimõttelised või tehnilist laadi raskused aga ongi põhiliseks takistuseks nende meetodite kasutamisel.

4. Varjatud tunnuste mõõtmine (latentanalüüs)

Kuigi probleem ei kuulu kujundite eristamise ülesannete hulka, käsitleme küsimust siinkohal sellepärast, et varjatud tunnuste mõõtmisel lähtutakse nagu selles peatükis varemgi käsitletud ülesannete (klassifitseerimine, taksonoomia) lahendamisel tunnustevahelise kauguse mõistest.

Vaatleme mõõtmisülesannet antud sõltuvate tunnuste süsteemi T põhjal saadud sõltumatute tunnuste suhtes.

Kvantitatiivsete tunnuste juhul lahendatakse see üles-

anne faktoranalüüsi meetoditega, kus antud tunnuste alusel moodustatakse nende lineaarsed kombinatsioonid (faktorid), mille vahel korrelatsioon on võrdne nulliga [26], (mõnikord kasutatakse ka nn. kaldtunnuseid, millede vahel korrelatsioon ei võrdu nulliga [2]).

Nominaalsete tunnuste puhul toimime järgmiselt. Koostame maatriksi $D = \| d(i,j) \|$, $(i,j = 1,2,\ldots,m)$, kus i -ndas reas ja j -ndas veerus asub tunnuste t_i ja t_j vaheline kaugus $d(i,j)$ (16).

Jaotame nüüd süsteemi T rühmadeks nii, et rühma sees oleksid tunnused maksimaalselt "lähedased" ja erinevates rühmades maksimaalselt "kauged". Seda võib teha olemasolevate taksonoomia meetodite abil (II, 4.). Kui tunnuste rühmitamine on teostatud, leiame igale saadud rühmale keskmise tunnuse. Neid keskmisi tunnuseid vaatlemegi faktoritena.

Sellise meetodiga leitud faktorid ei pruugi olla omavahel sõltumatud. Seost faktorite vahel võib seejuures mõõta nendevahelise kauguse (16) abil. Üksiku tunnuse kaalu faktoris saab hinnata selle tunnuse ja faktori vahelise kauguse abil. Mõõta võib ka faktori kaalu kogu süsteemis .

Käesolevas peatükis toodud käsitluse iseärasuseks on see, et ta võimaldab ühtsest vaatepunktist lähtudes kujundite eristamise meetoditele taandades lahendada mitmeid sisuliselt erinevaid ülesandeid.

IV TEST-MEETODI ÜLDISTUS KUJUNDITE ERISTAMISEL

1. Üldmärkused

Üksiku tunnuse suhtelise tähtsuse e. informatsiooni kaalu hindamine on objektide klassifitseerimisel olulise tähtsusega küsimus (II, 3.). Sõltumatute tunnuste korral on tunnuste informatsiooni kaalude alusel võimalik lähtesüsteemist T vahetult leida piisavalt informatiivne minimaalne tunnuste alamsüsteem T', mille alusel objekte klassifitseerida. Kui lähtesüsteemis T on tunnused sõltuvad, siis leitakse informatiivne alamsüsteem T' teistel meetoditel (II, 3.). Sel korral pakub huvi üksiku tunnuse informatsiooni kaalu hindamine süsteemis T'. Nii ühel kui teisel juhul seisneb ülesanne selles, kuidas leida tunnuse informatsiooni kaal etteantud süsteemis.

Üks võimalikest meetoditest, kuidas hinnata tunnuse suhtelist tähtsust, on esitatud töödes [14, 15] . Seal vaadeldakse informatsiooni kaalu kahes erinevas tähenduses: 1) kogu lähteklassifikatsiooni (õpperühma) suhtes s.o. tavalises tähenduses (II,3.), 2) mingi kindla kujundi suhtes. Esimesel juhul räägitakse tunnuse "eristamise kaalust", teisel juhul tunnuse "informatsiooni kaalust". Tunnus on teatud kujundi suhtes suure "informatsiooni kaaluga", kui ta peegeldab sellesse kujundisse kuuluvate objektide olulisi, tähtsaid külgi. Nii "eristamise kaalu" kui "informatsiooni kaalu" hindamise idee on ühesugune. Tunnust loetakse seda informatiivsemaks (ühes või teises tähenduses), mida rohkem kordi ta si-

saldub teatud kõrge informatiivsusega alamsüsteemides.

Meetodit on edukalt rakendatud sotsioloogilistes ja samuti geoloogilistes uurimustes [14,22] .

Mainitud töödes on tunnustele esitatud oluliselt kitsendav nõue - tunnuste dihhotoomilisus. Järgnevas püüame käsitlust üldistada suvaliste nominaaltunnuste jaoks.

2. Tunnuse eristamise kaal

Eeldame, et esmased andmed on esitatud $n \times m$ maatriksina M (I, 1.). Olgu antud õpperühm $O^* = \{o_1^*, o_2^*, \ldots, o_n^*\}$, mis on klassifitseeritud kujunditesse $S = \{S_1, S_2, \ldots, S_K\}$. Eeldame, et hulgas O^* ei ole ekvivalentseid objekte.

Toome sisse vajalikud tähistused. Maatriksi M sellist $n \times m$ alammaatriksit, mille read on hulga O^* elementide korteežid, tähistame sümboliga M^*. Vastavalt sellele, kuidas hulk O^* jaguneb kujunditesse S, jaguneb maatriks M^* omakorda alammaatriksiteks $M_1^*, M_2^*, \ldots, M_K^*$, kus alammaatriksi M_ℓ^*, $(\ell = 1,2,\ldots,K)$ read on hulga $S_\ell \cap O^*$ elemantide korteežid. Tähistame ridade arvu maatriksis M_ℓ^* tähega n_ℓ, $(\ell = 1,2,\ldots,K)$. Seega $\sum_{\ell=1}^{K} n_\ell = n$.

Defineerime testori mõiste, mis algsel kujul on esitatud töös [38] .

Tunnuste hulga T alamhulka $\tau_\lambda = \{t_{\lambda_1}, t_{\lambda_2}, \ldots, t_{\lambda_{m_\lambda}}\}$ ja vastavat maatriksi M^* veergude komplekti $M^*_{\tau\lambda} = \{\lambda_1, \lambda_2, \ldots, \lambda_{m_\lambda}\}$ nimetatakse testoriks kujundite S suhtes, kui pärast kõiki-

de veergude, mis ei kuulu komplekti M_{τ}^{*}, väljaeraldamist maatriksist M^{*} iga maatriksi $M_{\ell_1}^{*}$, $(\ell_1 = 1, 2, \ldots, K)$ kõik read jäävad erinevateks iga maatriksi $M_{\ell_2}^{*}$, $(\ell_2 = 1, 2, \ldots, K)$ kõikidest ridadest.

Antud definitsiooni kohaselt testor kujundite S suhtes on selliste tunnuste hulk, mille väärtused objektil määravad üheselt ära kujundi, kuhu see objekt kuulub. Kuna eeldasime, et hulgas 6^{*} pole ekvivalentseid objekte, siis leidub vähemalt üks testor S suhtes - see on hulk T. Seejuures võib esineda palju testoreid kujundite S suhtes. Näiteks, kui $\tau_1 \subset T$ on testor, siis on testor ka iga hulk $\tau_1 \cup \{t_i\}$, kus $t_i \in T$. Enamat huvi pakuvad ainult teatud tüüpi testorid.

Testorit kujundite S suhtes nimetatakse tupiktestoriks S suhtes, kui ükski tema alamhulk ei ole enam testor S suhtes.

Tupiktestor on seega teatud mõttes minimaalne testor, mis määrab üheselt ära klassi, kuhu objekt peab kuuluma.

Tähistame tähega K kõikvõimalike tupiktestorite arvu ja tähega K_i kõikide selliste tupiktestorite arvu, mis sisaldavad tunnust t_i.

Suurust

$$\varepsilon_i = \frac{K_i}{K} \qquad (1)$$

nimetame tunnuse t_i eristamise kaaluks.

Ilmselt kehtib seos $0 \leq \varepsilon_i \leq 1$.

Selgitame täpsemalt tunnuse eristamise kaalu sisulist tähendust.

Õpperühma 6^{*} kirjeldamine kõikide T tunnuste abil on tavaliselt üleliigne. Pärast mõne tunnuse eraldamist jääb

alles kirjelduse põhiomadus - eristab objektid erinevatest kujunditest. Eemaldades järk-järgult tunnuseid jõuame kirjelduseni, mis edasisel kokkusurumisel ei erista enam kõiki objekte erinevatest kujunditest. Sellised mittekokkusurutavad kirjeldused on nagu aluseks kõikidele teistele kirjeldustele.

On loomulik, et mida enam kordi antud tunnus sisaldub sellistes põhikirjeldustes, seda olulisem ja tähtsam on ta objektide klassifitseerimisel s.t. seda suurem on tema eristamise kaal.

Seega on võimalik hinnata õpperühma põhjal üksikute tunnuste eristamise kaalu. Seda kaalu kasutame ülejäänud objektide klassifitseerimisel.

3. Tunnuse informatsiooni kaal

Vaatleme $k_\ell \times m$ -mõõtmelist maatriksit M^*_ℓ, $(\ell = 1,2,\dots,K)$, mille read on kujundisse S_ℓ kuuluvate hulga G^* elementide korteežid.

Tunnuste hulga T alamhulka $\tau_\ell = \{t_{\ell_1}, t_{\ell_2}, \dots, t_{\ell_{n_\ell}}\}$ ja vastavat maatriksi M^*_ℓ veergude komplekti $M^*_{\tau_\ell} = \{\ell_1, \ell_2, \dots, \ell_{n_\ell}\}$ nimetatakse testiks kujundi S_ℓ suhtes, kui pärast kõikide veergude, mis ei kuulu komplekti $M^*_{\tau_\ell}$, eemaldamist maatriksi M^*_ℓ kõik read jäävad teineteisest erinevateks.

Seega test antud kujundi suhtes on selline tunnuste hulk, mille väärtusi teades võib identifitseerida igat objek-

ti antud kujundis. Eelduse tõttu, et hulgas G^* (järelikult ka hulgas $G^* \cap S_\ell$) ei ole ekvivalentseid objekte, võime väita, et iga kujundi S_ℓ suhtes leidub vähemalt üks test - see on hulk T. Kuid ühe kujundi suhtes võib esineda ka mitu testi. Näiteks, testi iga laiend mingi tunnusega on samuti test. Rohkem huvi pakuvad teatud omadusega testid.

Testi kujundi S_ℓ suhtes nimetatakse tupiktestiks S_ℓ suhtes, kui ükski tema alamhulk ei ole enam test S_ℓ suhtes.

Tupiktest S_ℓ suhtes on seega teatud mõttes minimaalne hulk tunnuseid, mille väärtuste põhjal võib identifitseerida igat objekti kujundist S_ℓ.

Tähistame arvuga N^ℓ kõikvõimalike tupiktestide arvu S_ℓ suhtes ja arvuga N_i^ℓ selliste tupiktestide arvu, mis sisaldavad teatud tunnust t_i.

Suurust

$$\eta_i^\ell = \frac{N_i^\ell}{N^\ell} \quad , (i=1,2,\dots,m\,;\ \ell=1,2,\dots,k), \qquad (2)$$

nimetatakse tunnuse t_i informatsiooni kaaluks kujundi S_ℓ suhtes.

Tunnuse t_i informatsiooni kaal η_i^ℓ, näitab tunnuse t_i esinemise suhtelist sagedust kujundi S_ℓ teatud põhikirjeldustes (tupiktestides). On loomulik, et mida enam kordi tunnus nendesse põhikirjeldustesse kuulub, seda olulisem ta on kujundisse S_ℓ kuuluvate objektide tundmaõppimisel. Seega on suure informatsiooni kaaluga tunnused selle kujundi tähtsad karakteristikud.

Definitsioonist (2) on näha, et tunnus võib olla ühe kujundi suhtes suure, teise kujundi suhtes aga väikese informatsiooni kaaluga. Seda asjaolu kasutame ära objektide klassifitseerimisel.

4. Klassifitseerimisalgoritmid

Olles defineerinud tunnuse eristamise kaalu (1) ja tunnuse informatsiooni kaalu (2), mida saab arvutada õpperühma põhjal, rakendame neid suurusi objektide klassifitseerimisel.

Olgu antud eespoolkirjeldatud maatriksid $M^*_1, \ldots, M^*_K$ ja leitud valemi (1) põhjal tunnuste eristamise kaalud $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_m$ ning valemi (2) põhjal tunnuste informatsiooni kaalud $\eta^\ell_1, \eta^\ell_2, \ldots, \eta^\ell_m$, $(\ell = 1,2,\ldots,K)$.

Edaspidi kasutatakse operatsiooni $A \sim B$, mis on määratud mingi fikseeritud tunnuse t_i, $(i=1,2,\ldots,m)$ väärtuste hulgal $V_i = \{t_{i1}, t_{i2}, \ldots, t_{iv_i}\}$ ning avaldub järgmiselt: iga $A,B \in V_i$ korral

$$A \sim B = \begin{cases} 1, & \text{kui} \quad A = B, \\ 0, & \text{kui} \quad A \neq B. \end{cases} \tag{3}$$

Oletame, et maatriks M^*_ℓ on antud kujul

$$M^*_\ell = \begin{pmatrix} \alpha^{(\ell)}_{11} & \alpha^{(\ell)}_{12} & \ldots & \alpha^{(\ell)}_{1m} \\ \alpha^{(\ell)}_{21} & \alpha^{(\ell)}_{22} & \ldots & \alpha^{(\ell)}_{2m} \\ \vdots & \vdots & & \vdots \\ \alpha^{(\ell)}_{n_\ell 1} & \alpha^{(\ell)}_{n_\ell 2} & \ldots & \alpha^{(\ell)}_{n_\ell m} \end{pmatrix}, \quad (\ell = 1,2,\ldots,K) \tag{4}$$

Tarvis on klassifitseerida objekt o, millele vastab korteež

$$\langle o \rangle = \langle \beta_1, \beta_2, \ldots, \beta_m \rangle .$$

Järgnevas esitame lühidalt 6 klassifitseerimisalgoritmi. Neist 3 algoritmi kasutab otsustuseeskirja δ määramisel tunnuste eristamise kaale ja 3 algoritmi tunnuste informatsiooni kaale.

1° Leitakse suurused g_ℓ valemi

$$g_\ell = \frac{\sum_{p=1}^{n_\ell} \sum_{i=1}^{m} (\alpha^{(\ell)}_{pi} \sim \beta_i) \cdot \varepsilon_i}{n_\ell}, \quad (\ell = 1,2,\ldots,K) \tag{5}$$

järgi. Suurusi g_ℓ võib vaadelda "sarnasustena" objekti o ja

kujundi S_ℓ vahel. Leitakse maksimaalne "sarnasus" ρ_k,

$$\rho_k = \max_{1 \le \ell \le K} \rho_\ell$$

ja objekt o klassifitseeritakse kujundisse S_k. Otsustuseeskiri δ_1 on seega järgmine:

$$\delta_1 : \; o \in S_k = \arg \max_{S_\ell \in S} \rho_\ell \; . \tag{6}$$

Osutub, et algoritm on stabiilne (väikesed muudatused õpperühmas 6^* ei too kaasa klassifitseerimise olulist muutumist). Stabiilsus tuleb sellest, et iga ρ_ℓ leidmisel arvestatakse kõiki elemente maatriksis M_ℓ^* (4). Maatriksi M_ℓ^* mõne üksiku elemendi muutmine ei avalda ρ_ℓ väärtusele olulist mõju. Kuid algoritmil on ka puudus: lihtne proovimine näitab, et klassifitseerimisvead võivad tekkida õppegrühmal. Võttes mingist kujundist ühe objekti ja rakendades sellele eeskirja δ_1, võime klassifitseerida selle objekti mingisse teise kujundisse.

2^0 Leiame suurused $\bar{\rho}_\ell$ valemi

$$\bar{\rho}_\ell = \frac{\sum_{p=1}^{r_\ell} \sum_{i=1}^{m} (\alpha_{p_i}^{(\ell)} \sim \beta_i) \cdot \eta_i^\ell}{r_\ell} \tag{7}$$

järgi. Suurus $\bar{\rho}_\ell$ on objekti o "sarnasus" kujundiga S_ℓ. Seejuures iga kujundi S_ℓ korral leitakse "sarnasus" $\bar{\rho}_\ell$ selle kujundi S_ℓ suhtes võetud informatsiooni kaalude η_i^ℓ abil.

Otsustuskiri δ_2 on järgmine:

$$\delta_2 : \; o \in S_k = \arg \max_{S_\ell \in S} \bar{\rho}_\ell \; . \tag{8}$$

Algoritmil 2^0 on samad omadused, mis algoritmil 1^0.

3^0 Arvutatakse suurused

$$(\rho_{max})_\ell = \max_{1 \le p \le r_\ell} \sum_{i=1}^{m} (\alpha_{p_i}^{(\ell)} \sim \beta_i) \cdot \varepsilon_i \; , \; (\ell = 1,2,\dots,K) \tag{9}$$

Suurus $(\rho_{max})_\ell$ näitab objekti o "sarnasust" kõige "sarnasema" objektiga klassist S_ℓ. Otsustuseeskiri δ_3 on:

$$\delta_3 : = o \in S_k = \arg \max_{S_\ell \in S} (\rho_{min})_\ell \; . \tag{10}$$

Seega klassifitseerimine toimub klassifitseeritavale objektile kõige "sarnasema" objekti järgi. Lihtne on veenduda, et algoritm 3^0 on ebastabiilne, kuid ei lase läbi vigu õppe - rühmal.

4^0 Leitakse suurused

$$(\bar{\rho}_{max})_\ell = \max_{1 \le p \le \ell} \sum_{i=1}^{m} (\alpha_{pi}^{(\ell)} \sim \beta_i) \cdot \eta_i^{\ell}, \quad (\ell = 1,2,\dots,k). \tag{11}$$

Otsustuseeskiri δ_4 on järgmine:

$$\delta_4 : o \in S_k = \arg \max_{S_\ell \in S} (\bar{\rho}_{min})_\ell . \tag{12}$$

Algoritmil 4^0 on samad omadused kui algoritmil 3^0.

Algoritmides 5^0 ja 6^0 püüame ühendada eelnevate algoritmide 1^0 - 4^0 head omadused.

5^0 Leiame arvud

$$N_\ell = \rho_\ell \cdot (\rho_{max})_\ell , \quad (\ell = 1,2,\dots,k). \tag{13}$$

Otsustuseeskiri δ_5 on:

$$\delta_5 : o \in S_k = \arg \max_{S_\ell \in S} N_\ell . \tag{14}$$

6^0 Leiame arvud

$$\bar{N}_\ell = \bar{\rho}_\ell \cdot (\bar{\rho}_{max})_\ell , \quad (\ell = 1,2,\dots,k). \tag{15}$$

Otsustuseeskiri δ_6 on:

$$\delta_6 : o \in S_k = \arg \max_{S_\ell \in S} \bar{N}_\ell . \tag{16}$$

Algoritmid 5^0 ja 6^0 ei ole oluliselt töömahukamad kui algoritmid 1^0 - 4^0, kuid arvestavad nii "püsivust" (1^0, 2^0) kui ka "o-viga õpperühmal" (3^0, 4^0).

Toodud eeskirjadest on praktikas kasutatud algoritme 1^0 ja 2^0 [14,22] . Ülejäänute sobivus tegelikuks rakendamiseks tuleb konkreetsete näidete põhjal veel tõestada.

V USALDUSKORDAJATE MEETOD SEKUNDAARSEL MÕÕTMISEL

1. Sekundaarse mõõtmise ülesanne

Sotsioloogilistes uurimustes saadakse esmased andmed (või vähemalt osa nendest) tavaliselt hinnangute, arvamuste, otsustuste jms. kujul. Seepärast kerkib tihti üles nn. sekundaarsete mõõtmiste probleem, mis seisneb selles, et nende hinnangute, arvamuste jne. põhjal tuleb leida mõõdetavatele objektidele teatud karakteristiku väärtused, mida pole võimalik leida otsesel meetodil.

Vaatleme edaspidi lähemalt sekundaarse mõõtmise üht lihtsamat tüüpi ülesannet, mida saab kirjeldada järgmise näite varal.

Olgu antud objektide hulk $O=\{o_1, o_2, \dots, o_n\}$, mille igal elemendil on tarvis teada mingi antud karakteristiku t väärtus $a_1, a_2, \dots, a_v$. Seejuures pole võimalik t väärtust leida otsese mõõtmise abil. Võtame appi eksperdid $t_1, t_2, \dots, t_m$, kes peavad igaüks eraldi hindama karakteristiku t väärtust igal objektil $o_j \in O$. Olgu eksperdi hinnang väärtuse kohta objektil o_j võrdne $t_i(o_j)$. Saame hinnangute maatriksi $M=\|t_i(o_j)\|$, $(i=1,2,\dots,m;\ j=1,2,\dots,n)$. Ülesanne on leida selle maatriksi põhjal igale objektile $o_j \in O$ teatud mõttes optimaalne karakteristiku t väärtus.

Kõige lihtsam, aga mitte kõige õnnestunum, lahendus ülesandele on järgmine: omistada objektile o_j karakteristiku t selline väärtus, mis sisaldub hinnangutes $t_i(o_j)$ kõige sagedamini.

Selle lahenduse teatud täiustus seisneb selles, et arvestatakse lisaks ka iga eksperdi "usaldusväärsust". Kuidas seda "usaldusväärsust" hinnata ja kuidas hinnangut kasutada objektile karakteristiku "optimaalse" väärtuse omistamisel, vaatleme edaspidi (2. - 4.).

Praktikas esineb näites toodud ülesanne tavaliselt mõnel teisel kujul. Kui ekspertide $t_1, t_2, \ldots, t_m$ all mõista paralleelseid tunnuseid, millest igaüks mõõdab (hindab) väärtuste $a_1, a_2, \ldots, a_v$ abil karakteristiku t väärtust, või kui mõista ekspertide all üht ja sama tunnust erinevatel ajamomentidel ja erinevates mõõtmistingimustes, siis ülesande mõte jääb samaks. Nimetatud olukorrad (paralleelsed mõõtmised ja korduvad mõõtmised) esinevad konkreetsetes uurimustes väga sageli. Käsitlemegi järgnevas eksperte $t_1, t_2, \ldots, t_m$ kui tunnuseid ning formuleerime ülesande üldisel kujul.

Olgu antud esmased andmed hinnangute maatriksina

$M = \| t_i(o_j) \|$, $(i = 1,2,\ldots,m;\ j = 1,2,\ldots,n)$, kus t_i on tunnus hulgast $T = \{t_1, t_2, \ldots, t_m\}$ ja o_j on objekt hulgast $O = \{o_1, o_2, \ldots, o_n\}$ ning $t_i(o_j)$ on tunnuse t_i poolt antud hinnang karakteristiku t väärtuse kohta objektil o_j (s.o. tunnuse t_i väärtus objektil o_j). Seejuures tunnuste $t_1, \ldots, t_m$ ja karakteristiku t väärtused kuuluvad hulka $A = \{a_1, a_2, \ldots, a_v\}$.

Ülesandeks on omistada hulga O igale objektile o_j karakteristiku t selline väärtus $t^*(o_j)$, mis oleks teatud mõttes optimaalne selle objekti jaoks.

2. Tunnuse väärtuse usalduskordaja

Fikseerime suvalise tunnuse t_i $(1 \le i \le m)$ ja suvalise väärtuse a_σ $(1 \le \sigma \le \nu)$. Tähistame niisuguste objektide o_j, mille korral $t_i(o_j) = a_\sigma$, arvu sümboliga f_i^{σ}. Kehtivad seosed

$$f_i^{\sigma} \ge 0, \qquad \sum_{\sigma} f_i^{\sigma} = N. \tag{1}$$ (2)

Tähistame sümboliga f_{ik}^{σ} selliste objektide $o_j \in \mathcal{O}$ arvu, mille korral $t_i(o_j) = t_k(o_j) = a_\sigma$. Ilmselt on õiged võrratused

$$0 \le f_{ik}^{\sigma} \le f_i^{\sigma}, \quad (k = 1, 2, \ldots, m). \tag{3}$$

Seejuures kehtib samasus

$$f_{ik}^{\sigma} = f_{ki}^{\sigma}, \quad (k = 1, 2, \ldots, m), \tag{4}$$

kusjuures

$$f_{ii}^{\sigma} = f_i^{\sigma}.$$

Juhul, kui $f_i^{\sigma} = 0$ (s.o. kui hulgas $\mathcal{O}$ ei leidu ainsatki objekti o_j, mille korral $t_i(o_j) = a_\sigma$), siis võrratuse (2) põhjal ka $f_{ik}^{\sigma} = 0$, $(k = 1, 2, \ldots, m)$ ja $\sum_k f_{ik}^{\sigma} = 0$.

Suurust

$$w_i^{\sigma} = \begin{cases} \dfrac{\sum_{k=1}^{m} f_{ik}^{\sigma}}{f_i^{\sigma} \cdot m}, & \text{kui } f_i^{\sigma} \neq 0, \\ 0, & \text{kui } f_i^{\sigma} = 0, \end{cases} \tag{5}$$

$(i = 1, 2, \ldots, m)$
$(\sigma = 1, 2, \ldots, \nu)$

nimetame väärtuse a_σ $(\sigma = 1, 2, \ldots, \nu)$ usalduskordajaks tunnuse t_i juures.

Väärtuse a_σ usalduskordaja w_i^{σ} tunnuse t_i juures on kahelt poolt tõkestatud suurus ja kehtib võrratus

$$0 \le w_i^{\sigma} \le 1. \tag{6}$$

See järeldub seosest

$$0 \le \sum_k f_{ik}^{\sigma} \le f_i^{\sigma} \cdot m,$$

mis on õige võrratuse (3) tõttu.

Defineeritud suurusele $w_i^{\wedge}$ võib anda tõenäosusteoreetilise tõlgenduse. Summa $\sum_k f_{ik}^{s}$ murru (5) lugejas näitab, mitu hinnangut $t_k(o_j)=a_s$ üle kõikide tunnuste $t_k \in T$ kokku said sellised objektid o_j hulgast O, mille korral hinnang tunnuse t_i poolt on $t_i(o_j)=a_s$. Suurus $f_i^{\wedge s} \cdot m$ murru (5) nimetajas näitab hinnangute $t_k(o_j)=a_s$, $(k=1,\dots,m)$, maksimaalset üldarvu objektide $o_j \in O$ korral, millel $t_i(o_j)=a_s$ (viimaseid on $f_i^{\wedge}$ tükki). Seega suurus $w_i^{\wedge}$ on tõenäosus, et objekt o_j, mis tunnuse t_i poolt saab hinnangu a_s, saab hinnangu a_s ka juhuslikult valitud tunnuse t_k, $(k=1,\dots,m)$ poolt. On loomulik eeldada, et mida suurem see tõenäosus on, seda usaldusväärsem on tunnuse t_i poolt antud hinnang a_s. Usaldusväärsuse mõõduks ongi usalduskordaja $w_i^{\wedge}$.

3. Tunnuse usalduskordaja

Suuruste $w_i^{\wedge}$ $(i=1,2,\dots,m;\ s=1,2,\dots,r)$ leidmisega saame kätte tunnuste $t_1,\dots,t_m$ poolt hulga O objektide kohta antavate kõikvõimalike hinnangute usalduskordajad. Seejuures ühe kindla tunnuse t_i poolt tehtavad erinevad hinnangud võivad olla erineva usalduskordajaga. Kuna igat tunnust tuleb vaadelda kui karakteristiku t väärtuse halvemat või paremat hindajat, siis ei tohiks tema poolt antava hinnangu usaldusväärsus sõltuda sellest hinnangust endast vaid ainult tunnusest. Seepärast defineerime tunnuse usalduskordaja, mis arvutatakse tunnuse poolt antavate üksikute hin-

nangute usalduskordajate alusel.

Suurust

$$w_i = \frac{\sum_{j=1}^{r} f_i^j \cdot w_i^j}{N} \tag{7}$$

nimetame i-nda tunnuse usalduskordajaks. Seega tunnuse usalduskordaja on tema väärtuste usalduskordajate kaalutud keskmine. Arvestades väärtuste usalduskordajate tähendust, võime öelda, et tunnuse t_i usalduskordaja näitab tõenäosust w_i, et suvalise tunnuse t_k $(k=1,2,\dots,m)$ hinnang mingil objektil $o_j \in O$ langeb kokku tunnuse hinnanguga sellel objektil.

Vaatleme tunnuste usalduskordajate mõningaid omadusi.

a) Tunnuse usalduskordaja on kahelt poolt tõkestatud suurus. Täpsemalt,

$$0 < w_i \leq 1 , \quad (i=1,2,\dots,m). \tag{8}$$

See tuleneb võrratustest

$$0 < \sum_{j=1}^{r} f_i^j \cdot w_i^j \leq \sum_{j=1}^{r} f_i^j = N,$$

mis kehtivad võrratuse (6) tõttu.

b) Ühe ja sama hinnangute maatriksi M alusel leitud tunnuste usalduskordajad on ilmselt omavahel seotud. Osutub, et see seos on seda tugevam, mida väiksem on tunnuste arv m. Juhul $m=2$ on mõlema tunnuse usalduskordajad w_1 ja w_2 võrdsed s.t. $w_1=w_2=w$, kusjuures suurust w võib vaadelda korrelatsioonina tunnuste t_1 ja t_2 poolt antud hinnangute vahel. Kui m muutub suureks, siis tunnuste usalduskordajate vaheline seos väheneb.

4. Sekundaarne mõõtmine

Oletame, et on leitud tunnuste $t_1, t_2, \ldots, t_m$ usaldus - kordajad $w_1, w_2, \ldots, w_m$ hinnangute maatriksi $M = \|t_i(o_j)\|$ alusel. Kasutame usalduskordajaid karakteristiku t "optimaalse" väärtuse $t^*(o_j)$ omistamisel hulga O objektidele o_j (sekundaarsel mõõtmisel).

Nimetame vektorit $\bar{o}_j = (t_1(o_j), t_2(o_j), \ldots, t_m(o_j))$ objekti o_j hinnangute vektoriks.

Vektorit $\Omega = (w_1, w_2, \ldots, w_m)$ nimetame usaldusvektoriks.

Moodustame lineaarse kombinatsiooni

$$\tau_j = \Omega \cdot \bar{o}_j = w_1 \cdot t_1(o_j) + \ldots + w_m \cdot t_m(o_j) \qquad (9)$$

Kuna iga $i = 1,2,\ldots,m$ ja iga $j = 1,2,\ldots,n$ korral leidub arv s nii, et $t_i(o_j) = a_s$, siis võime kirjutada, et

$$\tau_j = \alpha_1^j \cdot a_1 + \ldots + \alpha_v^j \cdot a_v \quad , \qquad (10)$$

kus $\alpha_s^j = \sum_{t_i(o_j)=a_s} w_i$. (11)

Leiame kordajate α_s^j $(s=1,2,\ldots,v)$ hulgast maksimaalse. Olgu see α_p^j ,

$$\alpha_p^j = \max_s \alpha_s^j \quad . \qquad (12)$$

Anname karakteristiku t "optimaalse" hinnangu $t^*(o_j)$ objektil o_j järgmiselt:

$$t^*(o_j) = a_p . \qquad (13)$$

Kokkuvõttes on sekundaarse mõõtmise algoritm järgmine:

$$t^*(o_j) = a_p = \arg\max_{a_s \in A} \alpha_s^j \quad ,$$

kus $\alpha_s^j = \sum_{t_i(o_j)=a_s} w_i$ ja w_i on leitud valemist (7) .

Võib täheldada sarnasust vaadeldud sekundaarse mõõtmise ülesande ja klassifitseerimisülesande (II,2.) vahel. Mõlemal juhul lähtutakse algandmete maatriksist M ja ülesan -

deks on selle maatriksi alusel omistada igale objektile teatud tunnuse väärtus, kusjuures selle tunnuse väärtuste arv on ette antud. Sekundaarse mõõtmise ülesande korral on nimetatud tunnusel kindel sisuline tähendus (s.o. karakteristik, mida me mõõdame), klassifitseerimisülesandes on selleks tunnuseks nn. identifitseeriv tunnus (III, 2.), mille konkreetne sisuline tähendus ei ole üldiselt teada.

Seega võib käsitletud sekundaarse mõõtmise ülesannet vaadelda spetsiaalse algandmete kujuga ($V_i = V$, $i = 1, 2, \ldots, n$) klassifitseerimisülesandena.

KOKKUVÕTE

Töös on käsitletud kvalitatiivsete andmete töötlemise mõningaid küsimusi, mis on seotud kujundite eristamise ja sekundaarse mõõtmise ülesannetega.

Esmalt vaadeldi olemasolevate, põhiliselt kvantitatiivsete tunnuste jaoks välja töötatud kujundite eristamise algoritmide rakendamisvõimalusi nominaaltunnuste juures. Põhimõtteliselt on kvantitatiivsete andmete töötlemise meetodite ärakasutamine võimalik selle tõttu, et enamikule nendes meetodites kasutatavatele statistikutele on olemas samatüübilised statistikud ka nominaaltunnuste korral. Lähem analüüs näitas, et mõtet on üle võtta lihtsamaid, kergelt interpreteeritavate tulemustega meetodeid. Mõnedele sellistest esitati ka vastav modifikatsioon nominaaltunnuste jaoks.

Mitmete kujundite eristamise ülesannete lahendamiseks toodi informatsioonistatistikute abil sisse tunnustevahelise kauguse mõiste. Seda kaugust kasutades on võimalik lahendada klassifitseerimise, informatiivse tunnuste süsteemi leidmise ja taksonoomia ülesandeid ning läbi viia ka latentanalüüsi. "Kauguste meetodi" sobivus tegelikuks rakendamiseks on vaja tõestada konkreetsete näidete varal.

Töös on üldistatud suvaliste nominaaltunnuste jaoks nn. test-meetodit, mis algselt oli mõeldud kasutamiseks ainult dihhotoomiliste tunnuste korral. Samas on esitatud ka mõned uued klassifitseerimisalgoritmid, mis on küllalt stabiilsed ja ei lase läbi vigu õpperühmal.

Klassifitseerimise ülesandega on küllalt sarnane sekundaarse mõõtmise ülesanne. Töös esitatud "usalduskordaja-

te meetod" sekundaarseks mõõtmiseks on kergesti tõenäosusteoreetiliselt interpreteeritav. Suhteliselt väikese arvutuste mahu tõttu sobib teda kasutada ka praktikas.

Kõik nominaaltunnuste jaoks loodavad statistilised meetodid on universaalsed selles mõttes, et nad on rakendatavad ka kvantitatiivsete ja järjestatud tunnuste juures. Seejuures tuleb ainult arvestada tekkiva informatsiooni kaoga.

Kuigi probleemide püstitamisel on antud töös lähtutud sotsioloogiliste uurimuste vajadustest, võib saadud tulemusi kasutada ka teistel aladel.

KASUTATUD KIRJANDUS

1. B o n n e r, R. E. A "Logical Pattern" Recogniton Program. IBM J. Res. and Dev., 1962, v. 6, VII, 3. Ref. 31. järgi.

2. H a r m a n, H. H. Modern Factor Analysis. Chicago, 1960.

3. K a n g r o, G. Matemaatiline analüüs II. Tallinn, 1968.

4. L a z a r s f e l d, P. F. H e n r y, N. W. Latent structure analysis. N.-Y., 1968.

5. M a r i l l, T., G r e e n, D.M. On the effectiveness of receptors on recognition systems. IEEE Trans., 1963, IT-9, 1. Ref. 31. järgi.

6. T i i t, E. Matemaatiline statistika I. Tartu, 1971.

7. V õ h a n d u, L., T a m m e, E., L u h t, L. Arvutusmeetodid I. Tallinn, 1971.

8. А й з е р м а н М.А., Б р а в е р м а н Э.М., Р о з о н о э р Л.И. Теоретические основы метода потенциальных функций в задаче об обучении автоматов разделению входных ситуаций на классы. - Авт. и телемех., 1964, № 6.

9. А н д е р с о н Т. Введение в многомерный статистический анализ. Москва, 1963.

10. Б р а в е р м а н Э.М. Метод потенциальных функций в задаче обучения машины распознаванию образов без учителя. - Авт. и телемех., 1966, № 10.

11. Вопросы статистической теории распознавания. Под ред. Б.В.Варского. Москва, 1967.

12. Г е р ч и к о в В.И. Взаимное ориентирование социологических шкал. - В сб.: Измерение и моделирование в социологии. Новосибирск, 1969.

I3. Гершензон Г.И., Черный Л.Б. Сравнительная оценка значимости важнейших факторов приживаемости трудовых ресурсов. - В сб.: Социальная мобильность и проблемы формирования и использования трудовых ресурсов. Новосибирск-Иркутск, 1970.

I4. Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П. Об одном принципе классификации и прогноза геологических объектов и явления. - Геол. и геоф., 1968, № 5.

I5. Дмитриев А.Н., Журавлев Ю.И., Кренделев Ф.П. О математических принципах классификаций предметов и явлений. - В сб.: Дискретный анализ. Вып. 7, 1966.

I6. Дорофеюк А.А. Алгоритмы обучения машины распознаванию образов без учителя, основанные на методе потенциальных функций. - Авт. и телемех., 1966, № 10.

I7. Елкина В.Н., Елкин Е.А., Загоруйко Н.Г. О возможности применения методов распознавания образов в палеонтологии. - Геол. и геоф., 1967, № 9.

I8. Елкина В.Н., Загоруйко Н.Г. Об алфавите объектов распознавания. - В сб.: Вычислительные системы. Вып. 22, 1966.

I9. Загоруйко Н.Г. Какими решающими функциями пользуется человек? - В сб.: Вычислительные системы. Вып. 28, 1967.

20. Загоруйко Н.Г. Классификация задач распознавания образов. - В сб.: Вычислительные системы. Вып. 22, 1966.

2I. Загоруйко Н.Г. Методика оценки информационной эффективности независимых параметров речевого сигнала. - В сб.: Вычислительные системы. Вып. I0, 1964.

22. К а н д ы б а В.Н. Применение логико дискретного метода для исследования потенциальной текучести кадров. - В сб.: Социальная мобильность и проблемы формирования и использования трудовых ресурсов. Новосибирск - Иркутск, 1970.

23. К а с т л е р Г. Азбука теории информации. - В сб.: Теория информации в биологии. Москва, 1960.

24. Л б о в Г.С. Выбор эффективной системы зависимых признаков. - В сб.: Вычислительные системы. Вып. 19, 1965.

25. Л б о в Г.С. Некоторые вопросы минимизации исходной системы признаков в распознавании образов. Диссертация, Новосибирск, 1967.

26. Л о у л и Д., М а к с в е л л А. Факторный анализ как статистический метод. Москва, 1967.

27. М и р к и н Б.Г. Новый подход к обработке социологической информации. - В сб.: Измерение и моделирование в социологии. Новосибирск, 1969.

28. М и р к и н Б.Г., Ч е р н ы й Л.Б. Измерение близости между различными разбиениями конечного множества объектов. - Авт. и телемех., 1970, № 5.

29. О с и п о в Г.В., А н д р е е в Э.П. Проблемы формирования точного знания в процессе конкретных социальных исследований. - В сб.: Социология и математика. Новосибирск, 1970.

30. Распознавание образов в социальных исследованиях. Отв.ред. Н.Г. Загоруйко и Т.И.Заславская. Новосибирск, 1968.

31. Распознавание слуховых образов. Под ред. Н.Г.Загоруйко и Г.Я.Волошина. Новосибирск, 1970.

32. Р а о С.Р. Линейные статистические методы и их применения. Москва, 1968.

33. Себестиан Г.С. Процессы принятия решений при распознавании образов. Киев, 1965.

34. Турбович И.Т. Об оптимальном методе опознания образов при взаимнокоррелированных признаках. Опознание образов. Теория передачи информации. Москва, 1965.

35. Устюжанинов В.Л. Проблема классификации в социологии и теория информации. - В сб.: Измерение и моделирование в социологии. Новосибирск, 1969.

36. Француз А.Г. Некоторые вопросы статистической теории распознавания образов. - В сб.: Бионика. Москва, 1965.

37. Француз А.Г. Распознавание образов с использованием "функции близости". Докл. на Всес. симпозиуме по распознаванию образов. Москва, июнь, 1965.

38. Чегис И.А., Яблонский С.В. Логические способы контроля работы электрических схем. - Труды Мат. Ин-та им.Стеклова. т. 51, 1958.

39. Черный Л.Б. Обобщение метода последовательных расчетов в одной задаче классификации. - В сб.: Математические модели и методы в социально-экономических исследованиях. Новосибирск, 1968.

40. Шеннон К. Работы по теории информации и кибернетике. Москва, 1968.

41. Шлезингер М.И. О самопроизвольном различении образов. - В сб.: Читающие автоматы. Киев, 1965.

О СТАТИСТИЧЕСКОЙ МЕТОДИКЕ АНАЛИЗА НОМИНАЛЬНЫХ ПРИЗНАКОВ

Резюме

В данной работе рассматриваются некоторые вопросы статистической обработки качественной информации, связанные с задачами распознавания образов и секундарных измерений.

Существующие алгоритмы распознавания образов выработаны в главном образом для количественных признаков.

Дискутируется проблема использования существующих алгоритмов распознавания образов у номинальных (качественных) признаков.

Даны некоторые новые методы для решения задач классификации, выбора информативной системы и таксономии. При этом используются понятие "расстояние" номинальных признаков, которое вычисляется по формуле $d(x,y) = 1 - \frac{2T(x,y)}{H(x)+H(y)}$, где x и y - признаки, $H(x)$ и $H(y)$ - энтропии признаков x и y, $T(x,y)$ - переданная информация признаков x и y.

Для решения одной задачи секундарных измерений предъявлен метод "коэффициентов надежности" признаков.

SISUKORD