KAUNO TECHNOLOGIJOS UNIVERSITETAS

FUNDAMENTALIŲJŲ MOKSLŲ FAKULTETAS

TAIKOMOSIOS MATEMATIKOS KATEDRA

Darius Tamašauskas

KLASTERIZAVIMO METODŲ, TAIKOMŲ BINARINIAMS DUOMENIMS, TYRIMAS

Magistro darbas

Vadovas
dr. Tomas Ruzgas

KAUNAS, 2012

KAUNO TECHNOLOGIJOS UNIVERSITETAS

FUNDAMENTALIŲJŲ MOKSLŲ FAKULTETAS

TAIKOMOSIOS MATEMATIKOS KATEDRA

TVIRTINU
Katedros vedėjas
doc. dr. N. Listopadskis
2012 06 04

KLASTERIZAVIMO METODŲ, TAIKOMŲ BINARINIAMS DUOMENIMS, TYRIMAS

Taikomosios matematikos magistro baigiamasis darbas

Vadovas
dr. Tomas Ruzgas
2012 06 01

Recenzentas
doc. dr. Tomas Rekašius
2012 06 01

Atliko
FMMM-0 gr. stud.
Darius Tamašauskas
2012 06 01

KAUNAS, 2012

KVALIFIKACINĖ KOMISIJA

Pirmininkas: Rimantas Rudzkis, profesorius (VU MII)

Sekretorius: Eimutis Valakevičius, docentas (KTU)

Nariai: Jonas Valantinas, profesorius (KTU)

Vytautas Janilionis, docentas (KTU)

Vidmantas Povilas Pekarskas, profesorius (KTU)

Zenonas Navickas, profesorius (KTU)

Arūnas Barauskas, dr., vice-prezidentas projektams (UAB „BalticAmadeus“)

Tamašauskas D. Klasterizavimo metodų, taikomų binariniams duomenims, tyrimas: Taikomosios matematikos magistro darbas / vadovas dr. T. Ruzgas; Taikomosios matematikos katedra, Fundamentaliųjų mokslų fakultetas, Kauno technologijos universitetas. – Kaunas, 2012. – 94 p.

SANTRAUKA

Klasterinė analizė dažnai taikoma duomenims, kurie matuojami santykių skalėje. Šiame darbe klasterizavimo metodai taikomi binariniams duomenims ir tiriamas hierarchinių bei nehierarchinių metodų efektyvumas.

Binariniams duomenims gauti naudojamas Monte-Karlo metodas. Duomenys modeliuojami panaudojant binominį skirstinį. Yra sukuriami duomenų mišiniai su atsiskiriančiais klasteriais, vidutiniškai persidengiančiais klasteriais bei smarkiai persidengiančiais klasteriais. Binariniai duomenys gali būti matuojami tik dichotominėje skalėje, kurioje taikomi specifiniai atstumo matai transformuojant duomenis į atstumų matricas. Kuomet duomenys transformuojami į atstumų matricas jau galima taikyti klasterizavimo metodus.

Darbe tiriami hierarchiniai ir nehierarchiniai metodai. Hierarchinių metodų tyrimas apima 10 metodų paklaidų palyginimą esant 10 skirtingiems atstumo matams, keičiant klasterių skaičių, duomenų savybių vektoriaus ilgį bei skirtingus duomenų mišinius.

Nehierarchinių metodų tyrime analizuojamas vienas iš populiariausių k-vidurkių metodas. Tiriamas teisingas klasterių skaičiaus parinkimas bei paklaidų palyginimas naudojant skirtingus duomenų mišinius keičiant klasterių skaičių ir duomenų savybių vektoriaus ilgį.

Statistinės analizės paketo SAS pagalba sukurti algoritmai skirti analizuoti hierarchinius ir nehierarchinius metodus. Taip pat sukurta vartotojo sąsaja, kuri palengvina algoritmų naudojimą bei įgalina patogiau analizuoti metodus keičiant tam tikrus parametrus. Išnagrinėjus metodų paklaidas pastebėta, kad vieni metodai labai sėkmingai priskiria reikiamus klasterius, tačiau kiti metodai binariniams duomenims nėra pritaikyti.

Tamašauskas D. Analysis of clustering methods for binary data: Master's work in applied mathematics / supervisor dr. T. Ruzgas; Department of Applied mathematics, Faculty of Fundamental Sciences, Kaunas University of Technology. – Kaunas, 2012. – 94 p.

SUMMARY

Clustering analysis is often applied to data, which can be measured in ratio scale. But in this work clustering methods are applied to binary data, and the research is made to compare hierarchical and partitive clustering methods and their efficiency.

Monte-Carlo simulation method is used for getting binary data. Data is being created using binomial distribution with given parameters for creating well separated clusters, average separated clusters and poorly separated clusters. Binary data can only be measured in nominal scale and there are used specific distance measures to transform data to distance matrices. When data transformation is finished, clustering methods can be used.

In this work we investigate hierarchical and partitive clustering methods. There are used 10 hierarchical methods with 10 different distance measures. We investigate how the error of certain method depends on these methods and distance measures, on cluster numbers, on different data distributions, on data property vector amount.

Partitive clustering methods investigations include popular partitive k-means method. Investigation is about determining the number of clusters and computing how the error of method depends on cluster numbers and different data distributions.

For analysis of clustering methods there were created algorithms with statistical analysis system SAS. Also program interface was created for easier way to analyse results. After investigation we saw that some methods perform well with binary data, but some are not very suitable.

TURINYS

LENTELIŲ SĄRAŠAS

PAVEIKSLŲ SĄRAŠAS

ĮVADAS

Klasterinės analizės pagrindinis tikslas yra suskaidyti tam tikrus objektus į grupes, vadinamas klasteriais, taip kad skirtumai klasterių viduje būtų kuo mažesni, o tarp klasterių – kuo didesni.

Daugelyje statistinės analizės darbų objektai yra klasifikuojami pagal daugiamačių vektorių reikšmes. O šiame darbe bus nagrinėjami klasterizavimo metodai, kai požymiai, pagal kuriuos atliekamas klasifikavimas yra dichotominiai. Dichotominėje matavimų skalėje požymių reikšmės yra priskiriamos prie vienos iš dviejų galimų kategorijų, t.y. gali įgyti reikšmę 1 arba 0.

Darbe nagrinėjami dviejų tipų metodai: hierarchiniai ir nehierarchiniai. Hierarchiniais metodais nustatoma bendra visų klasterių tarpusavio priklausomybių struktūra ir tik tada sprendžiama, koks klasterių skaičius optimalus. Metodai naudingi tada kai stebėjimų skaičius nėra labai didelis. Tyrėjas pats sprendžia, kuriuo etapu objektų paskirstymas į klasterius yra optimalus.

Nehierarchiniai metodai paprastai taikomi tada, kai iš anksto žinomas klasterių skaičius ir norima tiriamus objektus klasterizuoti. Vienas iš šio metodo privalumų yra tas, kad mums nereikia skaičiuoti atstumų tarp visų subjektų porų. Taip pat šis metodas yra efektyvesnis ir praktiškesnis kai turime labai didelius duomenų kiekius.

Kadangi realių binarinių duomenų praktikoje yra labai nedaug, binariniams duomenims gauti buvo naudojamas Monte-Karlo modeliavimo metodas, kurio pagalba modeliuojami binarinių duomenų mišiniai, panaudojant binominį skirstinį.

Modeliuotiems duomenims buvo skaičiuojamos atstumų matricos pagal dichotominei skalei tinkamus atstumo skaičiavimo būdus. Transformuotiems duomenims taikyti klasterizavimo metodai ir palyginamos paklaidos. Taip pat nagrinėta klasterių skaičiaus nustatymo problema nehierarchiniams metodams.

Darbe taikoma statistinė programinė įranga SAS, kurios pagalba sukurti algoritmai, kuriuos panaudojus išanalizuotos hierarchinių ir nehierarchinių klasterizavimo metodų paklaidos.

Darbo tikslas – ištirti klasterizavimo metodus taikomus binariniams duomenims ir nustatyti jų efektyvumą esant įvairioms duomenų aibėms.

Darbo uždaviniai:

- Išanalizuoti hierarchinius klasterizavimo metodus, taikomus binariniams duomenims, nustatyti kurie metodai yra tikslesni naudojant įvairius atstumo matus remiantis Monte-Karlo metodu.
- Išanalizuoti nehierarchinius klasterizavimo metodus, taikomus binariniams duomenims, išanalizuoti klasterių skaičiaus nustatymo principus.

- Sukurti programinės įrangos sąsają su vartotoju, kurioje galima būtų nagrinėti hierarchinius ir nehierarchinius klasterizavimo metodus binariniams duomenims.

Magistrinio darbo tematika yra išspausdintas 1 straipsnis 2012 m. studentų konferencijos „Taikomoji matematika“ medžiagoje. Konferencijoje skaitytas pranešimas.

Darbą sudaro įvadas, trys skyriai, išvados, literatūros sąrašas ir priedai. Pirmas skyrius skirtas klasterizavimo metodų ir algoritmų skirtų binariniams duomenims apžvalgai, antras skyrius – darbo rezultatams bei jų interpretacijai, o trečias skyrius sukurtos programinės įrangos sąsajai su vartotoju.

1.1. TEORINĖ DALIS

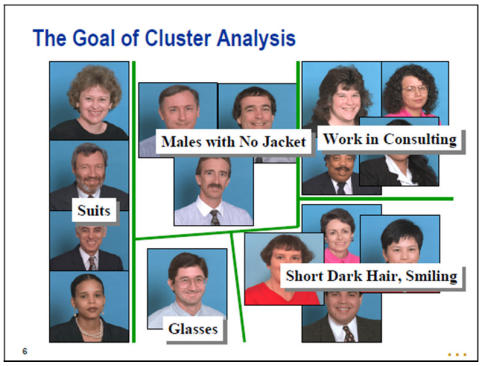
1.1. KLASTERIZAVIMAS

Klasterinė analizė – tai statistinis metodas, skirtas identifikuoti homogenines objektų arba stebėjimų grupes (klasterius). Objektai suskirstomi taip, kad skirtumai klasterių viduje būtų kuo mažesni, o tarp klasterių – kuo didesni.

Pagrindinės klasterinės analizės sąvokos yra panašumas ir skirtingumas (atstumas). Atstumas nurodo, kiek objektai nutolę vienas nuo kito (skirtingi), o panašumas rodo objektų artimumą. Panašūs objektai priklauso tam pačiam klasteriui, nutolę objektai – skirtingiems klasteriams.

Pats paprasčiausias būdas suvokti dviejų kintamųjų, kurių duomenys sudaro keletą homogeninių grupių, skirstymo į klasterius reiškinį, yra nubraižyti šių dviejų kintamųjų sklaidos (*angl. scatter*) diagramą. Realiai tokia galimybė pasitaiko labai retai – paprastai klasterių struktūros nėra aiškiai atskirtos ir persidengia, o kintamųjų skaičius žymiai viršija du, t. y. analizė atliekama *n*-matėje erdvėje.

Nepaisant akivaizdaus klasterinės analizės tikslo paprastumo, kaip matome iš 1.1 pav., klasterinė analizė yra labai sudėtingų veiksmų visuma, apimanti panašumo nustatymo problemas, kaip pateikti norimus rezultatus ir kita.



1.1 pav. Klasterinės analizės tikslas

Taikant klasterinės analizės metodus reikia turėti omenyje, kad:

- klasterinėje analizėje yra daug euristinių, neturinčių teorinio pagrindimo, metodų;
- klasterinės analizės metodai turi nemažai specifiškumo;
- tiems patiems duomenims taikant skirtingus klasterinės analizės metodus, galima gauti skirtingus rezultatus.

Skiriami du pagrindiniai klasterinės analizės metodų tipai:

- hierarchiniai metodai;
- nehierarchiniai metodai.

1.2. HIERARCHINIAI KLASTERIZAVIMO METODAI

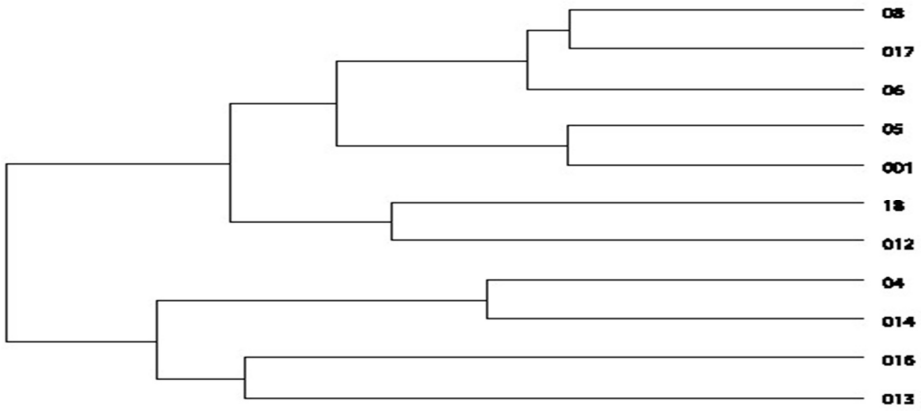
1.2.1. HIERARCHINIŲ METODŲ APIBENDRINIMAS

Hierarchiniai klasterizavimo metodai – tai tokie klasterizavimo metodai, kuriais nustatoma bendra visų klasterių tarpusavio priklausomybių struktūra ir tik tada sprendžiama, koks klasterių skaičius optimalus. Hierarchiniai metodai toliau skirstomi į jungimo ir skaidymo metodus.

Taikant jungimo metodus, iš pradžių visi stebėjimai traktuojami kaip atskiri klasteriai. Pirmuoju žingsniu du stebėjimai sujungiami į klasterį, kiekvienu kitu žingsniu naujas stebėjimas jungiamas prie esamo klasterio arba sujungiami du klasteriai. Suformuotas klasteris vėliau jau negali būti skaldomas – jis gali būti tik jungiamas su kitais klasteriais. Vyksmas kartojamas tol, kol lieka vienas klasteris. Tyrėjas pats sprendžia, kuriuo etapu objektų paskirstymas į klasterius yra optimalus.

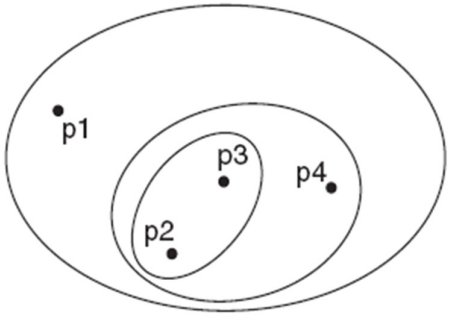
Skaidymo metodai yra loginė jungimo metodų priešingybė – vienintelis klasteris nuosekliai skaidomas į dalis.

Hierarchinis klasterizavimas dažnai pateikiamas grafiškai naudojant tam tikrą diagramą, vadinamą dendrograma, kuri parodo kiekvieno iš klasterių sąryšį bei tvarką kuria klasteriai buvo sujungti. 1.2 pav. matome dendrogramos pavyzdį, kuriame pavaizduota kaip buvo sujungti tam tikri stebėjimai į klasterinę struktūrą (iš pradžių jungtas 8 su 17 stebėjimu, tuomet iš jų sudarytas klasteris buvo sujungtas su 6 stebėjimu, kuris toliau buvo sujungtas su klasteriu, gautu sujungus 5 ir 1 stebėjimus ir t.t.)



1.2 pav. Dendrogramos pavyzdys

Taip pat klasterizavimas gali būti pateikiamas lizdo tipo klasterine diagrama, kurią matome 1.3 pav. Joje klasteriai sujungiami analogiškai kaip ir pavaizduota dendrogramoje (p2 stebėjimas sujungtas su p3, iš jų sudarytas klasteris sujungtas su p4 stebėjimu, kuris vėliau sujungtas su p1 stebėjimu).



1.3 pav. Lizdo tipo klasterinės diagramos pavyzdys

Hierarchiniai metodai nėra labai praktiški labai dideliems duomenų masyvams klasterizuoti, tačiau jie naudingi dėl to, kad nereikia daryti pradinio spėjimo apie klasterių skaičių, taip pat hierarchinių metodų neįtakoja išdėstymo tvarka.

Hierarchiniai klasterizavimo metodai jungia stebėjimus ir/arba klasterius, kurie yra labiausiai panašūs, t.y. jungia mažiausiai nutolusius klasterius.

Yra daugybė būdų apskaičiuoti atstumą tarp dviejų klasterių. Darbe yra naudojami 10 hierarchinių metodų, kiekvienas iš jų skiriasi savo skaičiavimo būdais.

1.2.2. HIERARCHINIŲ METODŲ PALYGINIMAS

Hierarchiniai klasterizavimo metodai vienas nuo kito skiriasi pagal jų skaičiavimo metodiką, pagal sudėtingumą, pagal tai kokius duomenis galima klasterizuoti. Kaip mes matome iš 1.1 lentelės tik su keliais iš jų galima tiesiogiai naudoti daugiamačius duomenis, tačiau daugeliui metodų reikia, kad duomenims būtų apskaičiuojama atstumo matrica [1].

1.1 lentelė

Hierarchinių klasterizavimo metodų palyginimas

Hierarchiniai klasterizavimo metodai	Daugiamačiai duomenys	Atstumo matrica
Average Linkage (vidutinio atstumo)	+	+
Centroid Linkage (atstumo tarp centrų)	+	+
Complete Linkage (tolimiausio kaimyno)		+
Density Linkage (tankių)		+
EML	+	
Flexible-Beta (beta-jautrus)		+
McQuitty‘s		+
Median		+
Single-Linkage (artimiausio kaimyno)		+
Two-Stage Density Linkage (dviejų pakopų tankių)		+
Ward‘s	+	+

Vidutinio atstumo metodas (*angl. average linkage*) – tai metodas, kuriame atstumas tarp dviejų klasterių yra apibrėžiamas kaip vidutinis atstumas tarp stebėtų porų kiekviename klasteryje. Vidutinio atstumo metodas apibrėžia atstumą tarp klasterių C_K ir C_L kaip:

$$D_{KL} = \frac{1}{n_K n_L} \sum_{i \in C_K} \sum_{j \in C_L} d(x_i, x_j) \quad \textbf{(1.1)}$$

kur n_K yra stebėjimų skaičius klasteriuose C_K, o n_L – stebėjimų skaičius klasteriuose C_L. Atstumas tarp stebėjimų i ir j yra $d(x_i, x_j)$.

Kadangi vidutinio atstumo metodas linkęs jungti klasterius su mažais standartiniais nuokrypiais, jis yra nežymiai šališkas atskiriant klasterius su tuo pačiu standartiniu nuokrypiu. Todėl kad jis naudoja visus klasterio narius, o ne tiesiog atskirą tašką. Tačiau vidutinio atstumo metodas efektyviau susidoroja su kraštutinėmis reikšmėmis. Taip pat vidutinio atstumo metodui nėra būtina atstumo matrica, todėl jis gali būti vykdomas greičiau negu kiti metodai.

Atstumo tarp centrų metodas (*angl. centroid linkage*) – tai metodas, kuriame atstumas tarp dviejų klasterių apibrėžiamas kaip kvadratinis Euklido atstumas tarp klasterio vidurkių ar centrų.

$$D_{KL} = \|\bar{x}_K - \bar{x}_L\|^2 \quad \textbf{(1.2)}$$

Kadangi atstumo tarp centrų metodas lygina klasterių vidurkius, tai šį metodą mažiausiai iš kitų hierarchinių metodų įtakoja išskirtys. Iš dviejų nelygiomis dalimis sugrupuotų klasterių didesnė dalis linkusi dominuoti mažesnę dalį. Taip pat stebėjimai gali būti priskirti vienam klasteriui, o vėliau priskirti jau kitam klasteriui. Taip pat kaip ir vidutinio atstumo metodas, šis metodas nereikalauja atstumo matricos, todėl gali būti vykdomas greičiau negu kiti hierarchiniai metodai.

Tolimiausio kaimyno metodas (*angl. complete linkage*) – tai metodas, kuriame atstumas tarp dviejų klasterių apibrėžiamas kaip didžiausias atstumas tarp stebėjimo viename klasteryje ir tarp stebėjimo kitame klasteryje.

$$D_{KL} = \max\ i \in C_K \max\ j \in C_L d(x_i, x_j) \quad \textbf{(1.3)}$$

Metodas rodo blogus rezultatus atskiriant klasterius su apytikriai lygiais diametrais ir gali būti iškreiptas vidutiniškų išskirčių.

Tankių metodas (*angl. density linkage*) – tai metodas, kuris naudoja vieną iš trijų neparametrinių tikimybės tankio įverčių tam kad paskaičiuoti naują atstumo metriką:

$$d^*(x_i, x_j) = \frac{1}{2}\left(\frac{1}{f(x_i)} + \frac{1}{f(x_j)}\right) \quad \textbf{(1.4)}$$

Galutinis klasterizavimo sprendimas yra paremtas nauja atstumo metrika.

Tankių metodas netaiko apribojimų klasterių formoms ir gali aptikti klasterius su ištęstomis arba nevienodomis formomis. Tačiau jis yra mažiau efektyvus negu kiti metodai aptinkant kompaktiškus klasterius.

Beta-jautrus metodas (*angl. flexible-beta*) – tai metodas, kuriame atstumas skaičiuojamas nuo naujai suformuoto klasterio *M* iki kito klasterio *J*:

$$D_{JM} = \left(D_{JK} + D_{JL}\right)\frac{(1-b)}{2} + D_{KL}b \quad \textbf{(1.5)}$$

kur b yra speciali reikšmė, dažnai rekomenduojama naudoti *-0,25*.

McQuitty‘s metodas (*angl. McQuitty‘s*) – tai metodas, kuriame atstumas tarp dviejų klasterių yra apskaičiuojamas taip:

$$D_{JM} = \frac{(D_{JK}+D_{JL})}{2} \quad \textbf{(1.6)}$$

kur D_{KL} yra bet koks nepanašumo ar atstumo matas tarp klasterių C_K ir C_L, todėl gaunasi, kad McQuitty metodas skaičiuoja svertinį vidurkį tarp klasterių.

Medianinis metodas (*angl. median*) – tai metodas, kurio atstumas, kai naudojama Euklido kvadratinė metrika, yra apskaičiuojamas taip:

$$D_{JM} = \frac{(D_{JK}+D_{JL})}{2} - \frac{D_{KL}}{4} \quad \textbf{(1.7)}$$

kur D_{ij} yra atstumas tarp klasterių i ir j. Atstumai D_{JK}, D_{JL} ir D_{KL} gali būti skaičiuojami naudojant bet kurią metriką. Medianinis metodas yra vienas iš seniausių klasterizavimo metodų.

Artimiausio kaimyno metodas (*angl. single linkage*) – tai metodas, kuriame atstumas tarp klasterių yra apskaičiuojamas kaip minimalus atstumas tarp stebėjimo viename klasteryje ir stebėjimo kitame klasteryje.

$$D_{KL} = \min\ i \in C_K \min\ j \in C_L d(x_i, x_j) \quad \textbf{(1.8)}$$

Šis metodas turi galimybę aptikti ištęstus ir netvarkingus klasterius, bet rodo gan blogus rezultatus su modeliuotais duomenimis Monte Karlo būdu.

Dviejų pakopų tankių metodas (*angl. two-stage density linkage*) – tai metodas, kuris yra labai panašus į artimiausio kaimyno metodą, paprastai naudojamas su tankio ryšiu. Su išimtim, kad kai 2 klasteriai yra jungiami kiekvienas turi turėti mažiausiai *n* narių.

Ward‘o metodas (*angl. Ward‘s*) – tai metodas, kuriame atstumas tarp klasterių yra apskaičiuojamas taip:

$$D_{KL} = \frac{\|\bar{x}_K - \bar{x}_L\|^2}{(\frac{1}{n_K} + \frac{1}{n_L})} \quad \textbf{(1.9)}$$

kur $\bar{x}$ yra klasterio C vektoriaus vidurkis, o n yra stebėjimų skaičius.

Ward metodas linkęs jungti klasterius su mažais stebėjimų skaičiais. Metodas yra labai jautrus išskirtims. Taip pat susiduria su problemomis atskiriant sferinius klasterius. Wardo metodas palyginus su kitais metodais yra vykdomas greičiau. Kaip ir vidutinio atstumo metodas bei atstumo tarp centrų metodas nereikalauja atstumo matricos.

1.2.3. ATSTUMŲ MATAI

Turėdami daugiamačius duomenis juos turime transformuoti į atstumų matricą. Tik keletui hierarchinių metodų tinka daugiamačiai duomenys, todėl norint ištirti visus hierarchinius metodus reikia modeliuotiems duomenims sudaryti atstumų matricas.

Matematikoje, kompiuterių moksle ar grafų teorijoje atstumų matrica – tai dvimatė matrica, sudaryta iš reikšmių vaizduojančių atstumus tarp dviejų skirtingų porų. Jos dydis yra *NxN*, kur *N* yra skirtingų objektų skaičius.

Jeigu turime duomenų masyvą sudarytą iš *N* eilučių ir *M* stulpelių, tai norėdami sudaryti šio duomenų masyvo atstumų matricą turime apskaičiuoti atstumą tarp kiekvienos poros iš *N* elementų. Jos dydis bus *NxN*, ir kiekviena atstumų matricos reikšmė vaizduos atstumą tarp 2 skirtingų porų.

Tarkime, kad *N*=*p*, o *M*=*q*. 1.2 lentelėje turime tokį daugiamatį duomenų masyvą, kuriame reikšmės yra binarinės: 1 arba 0.

1.2 lentelė

Binarinio duomenų masyvo pavyzdys

	1	2	3	...	q
1	1	1	1	…	1
2	1	0	0	…	1
3	1	0	1	…	1
...	…	…	…	…	…
p	1	0	1	…	1

Norėdami šiuos duomenis klasterizuoti ir suskaidyti į tam tikrus klasterius, reikia sudaryti atstumų matricą, kurios dydis bus *pxp*. Atstumų matricos pavyzdį matome 1.3 lentelėje, kurioje klaustukai žymi tam tikrais atstumų matais apskaičiuotus atstumus tarp skirtingų stebėjimo porų. Tarp tų pačių stebėjimų porų (*1* su *1*, *2* su *2* ir t.t.) atstumas yra 0.

1.3 lentelė

Atstumų matricos pavyzdys

	1	2	3	...	p
1	0	?	?	…	?
2	?	0	?	…	?
3	?	?	0	…	?
...	…	…	…	…	…
p	?	?	?	…	0

Atstumo sąvoka iš esmės atrodo labai paprasta, tačiau yra skirtingų atstumo matų, kurie apibrėžiami skirtingose matavimų skalėse ir apskaičiuojami skirtingais būdais.

Matavimų skalė – tai požymių reikšmių matavimo būdas. Matavimų skalės yra įvairios [13]:

- Nominalioji matavimų skalė – tai požymio reikšmių matavimo būdas, kuriuo tiriamieji objektai ar individai, atsižvelgiant į nagrinėjamo požymio kategorijas, skirstomi į grupes. Grupės numeruojamos laisvai, o jų numeriai turi ne skaičių, bet tam tikrų grupių (vardų) prasmę. Šia skale užrašoma žmogaus lytis, šeiminė padėtis, profesija, tautybė, diagnozė ir kiti požymiai.
    - Dichotominė matavimų skalė – taikant šią skalę, požymių reikšmės yra priskiriamos prie vienos iš dviejų galimų kategorijų. Pavyzdžiui, gyvas arba miręs, rūko – nerūko ir kt.
- Ranginė matavimų skalė – tai požymio reikšmių matavimo būdas, kuriuo skirtingai nei nominaliosios matavimų skalės, tiriamieji objektai ar individai gali būti sudėlioti matuojamo požymio didėjimo arba mažėjimo tvarka. Vartojant šią skalę galima teigti, kad vieno objekto požymio reikšmė didesnė, lygi ar mažesnė už kito objekto, nors neįmanoma nustatyti, kiek ji didesnė ar mažesnė.
- Intervalinė matavimų skalė – tai požymių reikšmių matavimo būdas, taikomas kiekybiniam dviejų požymių reikšmių skirtumui nustatyti. Šios skalės absoliutusis nulis nežinomas, bet nulinė reikšmė gali būti nustatoma bendru mokslininkų susitarimu. Tipinis šios skalės vartojimo pavyzdys – temperatūros matavimas vartojant Celsijaus ar Farenheito temperatūros skalę.
- Santykių matavimų skalė – tai požymio reikšmių matavimo būdas, kuriuo nulinė požymio reikšmė yra nustatoma ne susitarimu (kaip kad yra intervalo matavimų skalės), o yra absoliuti, susijusi su reiškinių esme. Santykių skale matuojamas žmogaus amžius, ūgis, kūno svoris, kraujospūdis, cholesterolio kiekis kraujyje ir kt.

Darbe atstumai yra skaičiuojami nominalioje matavimų skalėje, kurioje binarinis kintamasis gali įgyti 2 galimas reikšmes: 1 arba 0.

Atstumai nominalioje matavimų skalėje yra skirstomi į simetrinius ir nesimetrinius [6].

- Simetriniai atstumai. Jeigu nėra skirtumo kuri reikšmė yra vaizduojama 1, o kuri 0, tai binarinis kintamasis vadinamas simetriniu.

Pvz. Tarkime turime kintamąjį "Ar medis numeta lapus žiemai", o galimos reikšmės yra "Taip" (1) ir "Ne" (2). Abi reikšmės yra tarpusavyje lygiaverčios, todėl nėra skirtumo kuri reikšmė bus vaizduojama kaip 1, o kuri kaip 0.

- Nesimetriniai atstumai. Jeigu yra skirtumas kuri reikšmė bus 1, o kuri 0, tai toks kintamasis vadinamas nesimetriniu.

Pvz. Tarkime turime kintamąjį „Ar žmogus yra aklas", o galimos reikšmės yra „Taip" (1) ir "Ne" (2). Abi reikšmės nėra tarpusavyje lygiaverčios, nes jeigu mes galime sakyti, kad 2 akli žmonės turi kažką bendro, tai tikrai negalime sakyti kad 2 matantys žmonės turi kažką bendro.

Darbe atstumai yra taikomi 5 simetriniams ir 5 nesimetriniams matams nominalioje matavimų skalėje [6].

Simetriniams matams atstumas randamas tokiu būdu:

"Hamming" atstumas:

$$d_1(x,y) = X \tag{1.10}$$

"Dmatch" atstumas:

$$d_2(x,y) = \sqrt{X/N} \tag{1.11}$$

"Dsqmatch" atstumas:

$$d_3(x,y) = X/N \tag{1.12}$$

"Roger and Tanimoto" atstumas:

$$d_4(x,y) = M/(M + 2X) \tag{1.13}$$

"Sokal and Sneath 1" atstumas:

$$d_5(x,y) = 2M/(2M + X) \tag{1.14}$$

kur:

$M = \sum_{j=1}^{v} w_j \partial_{x,y}^j$, $\partial_{x,y}^j = 1$, jeigu $x_j = y_j$, $\partial_{x,y}^j = 0$, kitais atvejais.

$X = \sum_{j=1}^{v} w_j \partial_{x,y}^j$, $\partial_{x,y}^j = 1$, jeigu $x_j \neq y_j$, $\partial_{x,y}^j = 0$, kitais atvejais.

$N = \sum_{j=1}^{v} w_j$ (visų porų skaičius)

Nesimetriniams matams atstumas randamas taip:

"Djaccard" atstumas:

$$d_6(x,y) = 1 - \frac{\sum_j^v w_j(x_j y_j)}{\sum_{j=1}^v w_j(x_j y_j) + \sum_j^v w_j(x_j - y_j)^2} + X/P \quad \textbf{(1.15)}$$

"Dice" atstumas:

$$d_7(x,y) = 2PM/(P + PM) \quad \textbf{(1.16)}$$

"Russell and Rao" atstumas:

$$d_8(x,y) = PM/N \quad \textbf{(1.17)}$$

"Bray and Curtis" atstumas:

$$d_9(x,y) = X/(PAX + 2PP) \quad \textbf{(1.18)}$$

"Kulcynski1" atstumas:

$$d_{10}(x,y) = PM/X \quad \textbf{(1.19)}$$

kur:

$X = \sum_{j=1}^v w_j \partial_{x,y}^j$, $\partial_{x,y}^j = 1$, jeigu $x_j \neq y_j$ ir ne abu x_j bei y_j yra 0, $\partial_{x,y}^j = 0$, kitais atvejais

$PM = \sum_{j=1}^v w_j \partial_{x,y}^j$, $\partial_{x,y}^j = 1$, jeigu $x_j = y_j$ ir abu x_j bei y_j yra 1, $\partial_{x,y}^j = 0$, kitais atvejais

$PX = \sum_{j=1}^v w_j \partial_{x,y}^j$, $\partial_{x,y}^j = 1$, jeigu $x_j \neq y_j$ ir abu x_j bei y_j yra 1, $\partial_{x,y}^j = 0$, kitais atvejais

$$PP = PM + PX$$

$$P = PM + X$$

$PAX = \sum_{j=1}^v w_j \partial_{x,y}^j$, $\partial_{x,y}^j = 1$, jeigu $x_j \neq y_j$ ir arba x_j yra 1 bei y_j yra 0, arba x_j yra 0 bei y_j yra 1,

$\partial_{x,y}^j = 0$, kitais atvejais

$N = \sum_{j=1}^v w_j$ (visų porų skaičius)

1.3. NEHIERARCHINIAI KLASTERIZAVIMO METODAI

1.3.1. NEHIERARCHINIŲ METODŲ APIBENDRINIMAS

Jeigu stebėjimų skaičius yra didelis, tai taikyti hierarchinius klasterizavimo metodus tampa nebepraktiška. Tuomet galima taikyti nehierarchinius metodus.

Nehierarchiniai metodai paprastai taikomi tada, kai iš anksto žinomas klasterių skaičius ir norima tiriamus objektus klasterizuoti [1].

Vienas iš šio metodo privalumų yra tas, kad mums nereikia skaičiuoti atstumų tarp visų subjektų porų. Taip pat šis metodas yra efektyvesnis ir praktiškesnis kai turime labai didelius duomenų kiekius.

Pats svarbiausias nehierarchinio klasterizavimo metodas yra *k*-vidurkių metodas. Jis veikia taip. Užduodami pradiniai taškai, kurie laikomi klasterių vidurkiais. Visi stebėjimai priskiriami laikiniems klasteriams pagal mažiausią atstumą iki užduotų klasterių vidurkių. Užduotų klasterių vidurkiai keičiami laikinų klasterių vidurkiais ir procesas kartojamas kol klasteriai stabilizuojasi. Klasterizavimas yra paremtas Euklidiniu atstumu, ir stebėjimai esantys arti vienas kito priskiriami tam pačiam klasteriui, o stebėjimai nutolę vienas nuo kito – skirtingiems klasteriams.

Paprasčiausias nehierarchinio klasterizavivo algoritmas:

- Pasirinkti tam tikrą skaičių *k* taškų kaip pradinius taškus.
- Suformuoti *k* klasterius priskiriant visus taškus artimiausiems vidurio taškams.
- Perskaičiuoti kiekvieno klasterio vidurio taškus kol vidurio taškai nebesikeičia.

Atstumas apskaičiuojamas taip:

$$d = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2(m_i, x) \qquad \textbf{(1.20)}$$

čia:

k – klasterių skaičius,

x – duomenų taškas klasteryje C_i,

m_i – atitinkamas klasterio C_i taškas.

Šis metodas susiduria su su problemomis:

- kai klasteriai yra skirtingų dydžių;
- kai klasteriai yra skirtingo tankio;
- kai klasteriai nėra rutulio formos.

1.3.2. KLASTERIŲ SKAIČIAUS NUSTATYMAS

Viena iš problemų, su kuria dažnai susiduriama klasterinėje analizėje – tai klasterių skaičiaus nustatymas. Jie gali būti nustatomi pagal tam tikrus kriterijus. Vieni iš populiariausių kriterijų yra šie: Šarlio kubinis klasterizavimo kriterijus (CCC), Pseudo-F Statistika (PSF) bei Pseudo-T^2 Statistika (PST-2) [10].

Šarlio kubinis klasterizavimo kriterijus (CCC).

Taikant šį kriterijų tikrinamos tokios hipotezės:

H_0 : turimi stebėjimiai yra pasiskirstę pagal daugiamatį tolygųjį skirstinį.

H_1 : turimi stebėjimai yra pasiskirstę pagal daugiamatį normalųjį skirstinį, kai visų projekcijų skirstiniai yra su vienodom dispersijom.

Teigiami CCC dydžiai reiškia, kad H_0 yra atmesta.

Norint apkaičiuoti CCC, pirmiausia turi būti apskaičiuota dispersija:

$$E(R^2) \cong 1 - \left[\frac{\sum_{j=1}^{d^*}\frac{1}{n+u_j}+\sum_{j=d^*+1}^{d}\frac{u_j^2}{n+u_j}}{\sum_{j=1}^{d}u_j^2}\right]\left[\frac{(n-q)^2}{n}\right]\left[1+\frac{4}{n}\right] \quad \textbf{(1.21)}$$

čia n - stebėjimių skaičius, q - klasterių skaičius, s_j - hiperkubo ilgis j -tos projekcijos kryptimi ir

$$u_j = \frac{s_j}{c} \quad \textbf{(1.22)}$$

kai

$$c = \left(\frac{v}{q}\right)^{\frac{1}{d}} \text{ ir } v = \prod_{i=1}^{d} s_i \quad \textbf{(1.23)}$$

d^* parenkamas didžiausias sveikasis skaičius mažesnis už q, bet toks, kad u_{d^*} būtų nemažesnis už vienetą.

CCC yra apskaičiuojamas:

$$CCC = ln\left[\frac{1-E(R^2)}{1-R^2}\right]\frac{\sqrt{\frac{nd^*}{2}}}{(0{,}001+E(R^2))^{1{,}2}} \quad \textbf{(1.24)}$$

čia

$$R^2 = 1 - \frac{d^*+\sum_{j=d^*+1}^{d}u_j^2}{\sum_{j=1}^{d}u_j^2} \quad \textbf{(1.25)}$$

Jei CCC reikšmės yra tarp 0 ir 2, tai galima klasterinė struktūra. Jei reikšmės yra didesnės už 2, tai rodo, kad parinktas tinkamas klasterių skaičius. Didelės neigiamos CCC reikšmės gali būti gautos dėl išskirčių.

Pseudo-F Statistika (PSF)

Pseudo-F Statistika matuoja išskaidymą tarp klasterių hierarchijos dabartiniame lygyje. Ji yra klasterių skaičiaus indikatorius. Statistika yra pasiskirsčiusi pagal Fišerio skirstinį su $v(g-1)$ ir $v(n-g)$ laisvės laipsniais, kur v yra kintamųjų skaičius su 2 prielaidomis:

1. klasterizavimo metodas skiria stebėjimus nuo klasterių atsitiktinai;
2. duomenys nepriklausomai paimami iš daugiamačio normaliojo skirstinio.

PSF yra apskaičiuojama taip:

$$PSF = \frac{(\sum_{i=1}^{n}\|x_i-\bar{x}\|^2)/(g-1)}{(\sum_{j=1}^{g}\sum_{i\in C_k}\|x_i-\overline{x_k}\|^2)/(n-g)} \quad \textbf{(1.26)}$$

kur $\bar{x}$ yra vektoriaus vidurkis, $\bar{x}_k$ yra klasterio k vektoriaus vidurkis, g yra klasterių skaičius bet kuriame duotame hierarchijos lygyje, n yra stebėjimų skaičius, x_i yra i-tasis stebėjimas.

PSF taip pat gali būti apibrėžiamas taip:

$$PSF = \frac{R^2/(g-1)}{(1-R^2)/(n-g)} \quad \textbf{(1.27)}$$

Pseudo-T^2 Statistika (PST-2)

Pseudo-T^2 Statistika lygina dviejų daugiamačių populiacijų vidurkius. PST-2 gali būti naudojama nuspręsti ar gali būti sujungiami 2 klasteriai ar ne. Jeigu PST-2 reikšmė yra didelė, klasterių vidurkiai yra pastebimai skirtingi, iš to seka, kad klasteriai neturėtų būti jungiami. Ir atvirkščiai, jeigu PST-2 reikšmė yra maža, tada klasteriai gali būti sujungiami.

Pseudo-T^2 Statistika yra pasiskirsčiusi pagal Fišerio skirstinį su v ir $v(n_K + n_L - 2)$ laisvės laipsniais.

PST2 jungiant klasterius C_k ir C_l į klasterį C_m yra apskaičiuojama taip:

$$PST2 = \frac{w_m - w_k - w_l}{[(w_k+w_l)/(n_k+n_l-2)]} \quad \textbf{(1.28)}$$

kur n_l yra stebėjimų skaičius klasteryje i ir

$$w_k = \sum_{i\in C_k}\|x_i - \bar{x}_k\|. \quad \textbf{(1.29)}$$

1.4. MONTE-KARLO MODELIAVIMAS

Realių binarinių duomenų praktikoje yra ganėtinai mažai. Tai galėtų būti įvairios lentelės su informacija apie žmogaus lytį, jo savybes, priklausymą tam tikroms grupėms ir panašiai, kuriose kintamasis įgyja reikšmę 1 arba 0.

Todėl darbe yra naudojamas Monte-Karlo modeliavimas, kurio pagalba modeliuojami binarinių duomenų mišiniai, panaudojant binominį skirstinį.

Monte-Karlo metodas – skaičiavimo algoritmas, pagrįstas statistiniu modeliavimu ir gautų rezultatų apdorojimu statistiniais metodais. Šis metodas leidžia brangiai kainuojančius bandymus pakeisti modeliavimu kompiuteriais ir labai sumažina tyrimų trukmę. Monte-Karlo metodai dažniausiai naudojami matematinių sistemų modeliavimui, kai neįmanoma gauti tikslių rezultatų naudojant deterministinį algoritmą [11].

Norint atlikti labai sudėtingą skaičiavimą, reikalaujantį ištyrinėti didelę duomenų erdvę, galima tą patį skaičiavimą atlikti tik su keletu atsitiktinai pasirinktų duomenų. Atsitiktinai parinkti duomenys dažniausiai būna „tipiški“, todėl natūralu tikėtis, kad ir atliktas skaičiavimas ne itin daug skirsis nuo tikslaus [12].

Binariniams duomenims modeliuoti naudojamas binominis skirstinys. Binominis skirstinys – tai dichotomine matavimų skale matuojamų požymių reikšmių skirstinys (pasiskirstymo dėsnis). Šis skirstinys yra diskretus, jis apibūdinamas parametrais n ir p. Parametras $n \geq 0$ reiškia bandymų kiekį, o p – požymio tikimybę įgyti vieną iš dviejų galimų reikšmių (požymio tikimybė įgauti antrąją reikšmę yra lygi $1 - p$). Binominio skirstinio pasiskirstymo tankio funkcija yra:

$$f(k,n,p) = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \quad \textbf{(1.30)}$$

čia $k = 0,1,2,\dots,n$.

Darbe naudojami įvairūs binarinių duomenų mišiniai iš 2, 3 ir 4 skirtingų tikimybių ir tankių:

$$f_X(x) = \sum_{i=1}^{q} p_i f_{X_i}(x),\ q = 2,3,4 \quad \textbf{(1.31)}$$

1.5. BINARINIŲ DUOMENŲ KLASTERIZAVIMAS KITŲ AUTORIŲ DARBUOSE

1.5.1. BENDRAS KLASTERIZAVIMO MODELIS BINARINIAMS DUOMENIMS

Modelis binarinių duomenų klasterizavimui [7] apibrėžiamas taip:

$$W = AXB^T + E \quad \textbf{(1.32)}$$

čia:

Matrica E reiškia paklaidos dedamąją,

AXB^T charakterizuoja matricos W informaciją, kuri gali būti apibrėžiama klasterinėmis struktūromis,

A ir B nurodo klasterių narių skaičių duomenų taškams ir savybėms,

X nurodo klasterių vaizdavimą.

1.4 lentelėje pateikti visi matavimai.

1.4 lentelė

Binarinio klasterizavimo modelio matavimai

$W=(w_{ij})_{nxm}$	Dvejetainių duomenų rinkinys
$D=(d_1,d_2,\ldots,d_n)$	Duomenų taškų rinkinys
$F=(f_1,f_2,\ldots,f_m)$	Savybių rinkinys
K	Klasterių skaičius duomenų taškams
C	Klasterių skaičius savybėms
$P=\{P_1,P_2,\ldots,P_K\}$	D padalijimas į K skaičių klasterių
$i \in P_k,\ 1\leq k\leq K$	i-tasis duomenų taškas klasteryje P_k
$p_1,p_2,\ldots,p_K$	K duomenų klasterių dydžiai
$Q=\{Q_1,Q_2,\ldots,Q_C\}$	F padalijimas į C skaičių klasterių
$j \in Q_c,\ 1\leq c\leq C$	j-toji savybė klasteryje Q_c
$q_1,q_2,\ldots,q_C$	C savybių klasterių dydžiai
$A=(a_{ik})_{nxK}$	Matrica, nurodanti duomenų narių skaičių
$B=(b_{jc})_{mxC}$	Matrica, nurodanti savybių narių skaičių
$X=(x_{kc})_{KxC}$	Matrica, nurodanti ryšį tarp duomenų ir savybių
$Trace(M)$	Matricos M pėdsakas

Pažymime $\widehat{W}$ kaip AXB^T aproksimaciją. Klasterizavimo tikslas bus sumažinti aproksimacijos paklaidą.

$$O(A,X,B) = \left\|W - \widehat{W}\right\|_F^2 = Trace\left[\left(W - \widehat{W}\right)\left(W - \widehat{W}\right)^T\right] = \sum_{i=1}^{n}\ \sum_{j=1}^{m}\left(w_{ij} - \widehat{w}_{ij}\right)^2 =$$

$$\sum_{i=1}^{n}\ \sum_{j=1}^{m}\left(w_{ij} - \sum_{k=1}^{K}\ \sum_{c=1}^{C} a_{ik}b_{jc}x_{kc}\right)^2 \quad \textbf{(1.33)}$$

$$\|M\|_F = \sqrt{\Sigma_{ij}M_{ij}^2}. \quad \textbf{(1.34)}$$

Bendra optimizavimo procedūra:

Tarkime

$$A = (a_{ik}),\ (a_{ik})\epsilon\{0,1\},\ \sum_{k=1}^{K} a_{ik} = 1 \quad \textbf{(1.35)}$$

$$B = (b_{jc}),\ (b_{jc})\epsilon\{0,1\},\ \sum_{c=1}^{C} b_{jc} = 1 \quad \textbf{(1.36)}$$

A ir B aprašo duomenų ir savybių narių skaičius. Iš ankstesnių lygčių gauname:

$$O(A,X,B) = \left\|W - \widehat{W}\right\|_F^2 =$$

$$= \sum_{i=1}^{n}\ \sum_{j=1}^{m}\left(w_{ij} - \sum_{k=1}^{K}\ \sum_{c=1}^{C} a_{ik}b_{jc}x_{kc}\right)^2 = \sum_{k=1}^{K}\ \sum_{c=1}^{C}\ \sum_{i\epsilon P_k}\ \sum_{j\epsilon Q_c}(w_{ij} - x_{kc})^2 \quad \textbf{(1.37)}$$

Fiksuotiems P_k ir Q_c lengva patikrinti, kad optimalus X yra gaunamas taip:

$$x_{kc} = \frac{1}{p_k q_c}\sum_{i\epsilon P_k}\ \sum_{j\epsilon Q_c} w_{ij} \quad \textbf{(1.38)}$$

Kitais žodžiais, X galima interpretuoti kaip centroidų matricą, ir ji nusako ryšį tarp duomenų klasterių ir savybių klasterių.

$O(A,X,B)$ gali būti minimizuojamas tokia procedūra [9]:

1. Duota *X* ir *B*, tada savybių padalijimas *Q* yra koreguojamas

$$\hat{a}_{ik} = \begin{cases} 1, jeigu \sum_{c=1}^{C}\ \sum_{j\epsilon Q_c}(w_{ij} - x_{kj})^2 < \sum_{c=1}^{C}\ \sum_{j\epsilon Q_c}(w_{ij} - x_{lj})^2, kai\ l = 1,\dots,K, l \neq k \\ 0, kitais\ atvejais \end{cases} \quad \textbf{(1.39)}$$

2. Duota *X* ir *A*, tada duomenų padalijimas *P* yra koreguojamas

$$\hat{b}_{jc} = \begin{cases} 1, jeigu \sum_{k=1}^{K}\ \sum_{i\epsilon P_k}(w_{ij} - x_{ic})^2 < \sum_{k=1}^{K}\ \sum_{i\epsilon P_k}(w_{ij} - x_{il})^2, kai\ l = 1,\dots,C, l \neq c \\ 0, kitais\ atvejais \end{cases} \quad \textbf{(1.40)}$$

3. Duota *A* ir *B*, tada *X* gali būti apskaičiuojamas pagal formulę

$$x_{kc} = \frac{1}{p_k q_c}\sum_{i\epsilon P_k}\ \sum_{j\epsilon Q_c} w_{ij} \quad \textbf{(1.41)}$$

Algoritmas procedūrai atlikti:

Įvedame (W_{nxm}, K ir C)
Išvedame: A ir B
begin
1 Aprašome A ir B
2 Suskaičiuojame X
3 Kartojame ciklą, kol nėra sutinkamas stop kriterijus
begin
1 Atnaujinti A
2 Atnaujinti B
3 Perskaičiuoti X

end
4 Išvedame A ir B

end

1.5.2. BLOK-DIAGONALINIS KLASTERIZAVIMAS

Matrica *X* aprašo ryšį tarp duomenų klasterių ir savybių klasterių. Turint binarinius duomenis dažniausiai ryšys tarp duomenų ir savybių yra simetrinis. Pasinaudodami tuo galime sukurti bendro modelio variaciją [7], kur *X* vienetinė matrica. Tada turime, kad *C=K*, duomenys ir savybės turi tą patį skaičių klasterių.

Tada AB^T gali būti interpretuojamas kaip *W* aproksimacija. Klasterizavimo tikslas tada bus surasti *(A,B)*, kuri minimizuoja kvadratinį skirtumą tarp *W* ir jos aproksimacijos AB^T.

$$O(A,B) = \|W - AB^T\|_F^2 \tag{1.42}$$

Optimizavimo procedūra:

$$O(A,B) = \|W - AB^T\|_F^2 = \sum_{i=1}^{n} \sum_{j=1}^{m} \left(w_{ij} - \sum_{k=1}^{K} a_{ik} b_{kj}\right)^2 = \sum_{i=1}^{n} \sum_{k=1}^{K} a_{ik} \sum_{j=1}^{m} \left(w_{ij} - b_{kj}\right)^2 =$$
$$\sum_{i=1}^{n} \sum_{k=1}^{K} a_{ik} \sum_{j=1}^{m} \left(w_{ij} - y_{kj}\right)^2 + \sum_{k=1}^{K} n_k \sum_{j=1}^{m} \left(y_{kj} - b_{kj}\right)^2 \tag{1.43}$$

kur $y_{kj} = \frac{1}{n_k} \sum_{i=1}^{n} a_{ik} w_{ij}$, o $n_k = \sum_{i=1}^{n} a_{ik}$

Kai duota *B*, naujos *A* reikšmės gali būti apskaičiuojamos priskiriant kiekvieną duomenų tašką artimiausiam klasteriui:

$$\hat{a}_{ik} = \begin{cases} 1, jeigu \sum_{j=1}^{m} (w_{ij} - b_{kj})^2 < \sum_{j=1}^{m} (w_{ij} - b_{ij})^2, l = 1, \dots, K, l \neq k \\ 0, kitais\ atvejais \end{cases} \tag{1.44}$$

Kai *A* yra fiksuotas, $O_{A,B}$ gali būti minimizuojamas *B* atžvilgiu minimizuojant šitokiu būdu:

$$O'(B) = \sum_{k=1}^{K} n_k \sum_{j=1}^{m} \left(y_{kj} - b_{kj}\right)^2 \quad \textbf{(1.45)}$$

y_{kj} gali būti suprantamas kaip tikimybė, kad *j*-toji savybė yra *k*-tajame klasteryje. Kadangi kiekvienas b_{kj} yra binarinis, t.y. arba 0 arba 1, tai $O'(B)$ minimizuojamas taip:

$$b_{kj} = \begin{cases} 1, jeigu\ y_{kj} > 1/2 \\ 0, kitais\ atvejais \end{cases} \quad \textbf{(1.46)}$$

Po kiekvienos iteracijos, mes paskaičiuojame $O(A,B)$ reikšmes. Jeigu reikšmė yra sumažėjusi, tada mes kartojame procesą, priešingu atveju – procesas pasiekė vietinį minimumą.

Algoritmas procedūrai atlikti:

Įvedame: (W_{nxm} ir K)

Išvedame: A ir B
begin
1.1 Aprašome A
1.2 Paskaičiuojame B pagal auksčiau aprašytas lygtis
1.3 Paskaičiuojame O_0=O(A,B)
begin
2.1 Atnaujinam A kai duota B
2.2 Paskaičiuojam B kai duota A
2.3 Paskaičiuojam reikšmę O_1=O(A,B)
2.4 If O_1< O_0
2.4.1 O_0=O_1
2.4.2 Kartoti nuo 2.1
2.5 else
2.5.1 break
end
3 Grąžinti A ir B
end

1.6. PROGRAMINĖS ĮRANGOS APŽVALGA

Klasterizavimo metodams tirti darbe naudojama SAS programa. SAS (*Statistical Analysis System*) – tai integruota programinės įrangos produktų sistema, kuri leidžia vartotojams atlikti įvairius veiksmus: statistinę analizę, kokybės patobulinimus, duomenų apdorojimą, ataskaitų pateikimą, duomenų gavybą ir kita. Taip pat SAS turi daug verslo sprendimų, kurie įgalina didelės apimties programinės įrangos sprendimus tokiose srityse kaip IT valdymas, finansų valdymas, verslo analitikos ir kita.

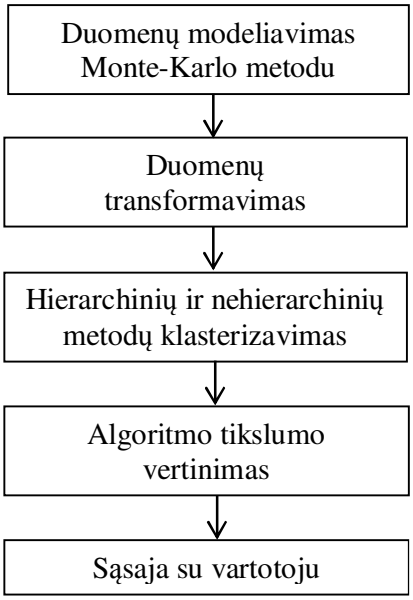
Šiuo metu SAS sistema yra viena iš galingiausių duomenų analizės sistemų, ji naudojama dideliam duomenų kiekiui apdoroti. SAS sistema turi privalumų prieš kitus paketus, tokius kaip SPSS, Statistica, Minitab ar Mathematica. SAS kalba turi gerai išvystytą makro kalbą, matricų kalbą.

SAS sistema gali būti valdoma komandų ir meniu pagalba. SAS programavimo kalba turi daug įvairių komandų, funkcijų, operatorių ir kitokių programavimo priemonių, kurios užtikrina norimo rezultato pasiekimą. Sistemos SAS programavimo kalba nėra sudėtinga, jos sintaksė yra labai panaši į kitų algoritminių kalbų sintaksę.

2. TIRIAMOJI DALIS IR REZULTATAI

2.1. TYRIMO SCHEMA

Klasterizavimo metodų efektyvumo tyrimas prasideda duomenų modeliavimu Monte-Karlo metodu. Tuomet modeliuoti duomenys transformuojami jiems sudarant atstumų matricas. Šie duomenys klasterizuojami bei tiriamas jų tikslumas skaičiuojant paklaidas. Galiausiai yra sukuriama sąsaja su vartotoju. Visą darbo eigos schemą matome 2.1 pav.



2.1 pav. Klasterizavimo metodų tyrimo schema

Hierarchinių metodų tyrimą sudaro 3 dalys:

- Metodų paklaidų palyginimas, kai tiriamas algoritmo tikslumas su skirtingais modeliuotų duomenų mišiniais.
- Metodų paklaidų palyginimas, kai tiriamas algoritmo tikslumas kai keičiame daugiamačių duomenų savybių skaičių.
- Metodų paklaidų palyginimas, kai tiriamas algoritmo tikslumas su skirtingais klasterių skaičiais.

Nehierarchinių metodų tyrimą sudaro 2 dalys:

- Paklaidų palyginimas, kai tiriamas skirtingų klasterių skaičius su skirtingai modeliuotais duomenų mišiniais.
- Paklaidų palyginimas, kai tiriamas tinkamas klasterių skaičiaus parinkimas.

2.2. PRADINIŲ REIKŠMIŲ PARINKIMAS

Pradiniai duomenys buvo modeliuoti Monte-Karlo modeliavimo metodu, panaudojus binominio skirstinio formulę:

$$f(k,n,p) = \frac{n!}{k!(n-k)!}p^k(1-p)^{n-k} \quad \textbf{(2.1)}$$

kur *n=1*, o tikimybė *p* buvo keičiama pagal norimus modeliuoti duomenis.

Kiekvienai hierarchinių metodų tyrimo daliai buvo tiriami 10 hierarchinių metodų su 10 skirtingų atstumų skaičiavimo būdų. Stebėjimų skaičius parinktas 120, o duomenų mišiniai padalinami į lygias dalis pagal skirtingas tikimybes.

Tiriant metodų tikslumą priklausomai nuo modeliuotų duomenų mišinių, savybių vektoriaus ilgis parinktas 10, o tikimybės *p* parinktos tokiu būdu:

- Kai tiriami 2 klasteriai, *p1=0,8*, *p2=0,2* (atsiskiriantys klasteriai), *p1=0,7*, *p2=0,3* (vidutiniškai persidengiantys klasteriai) ir *p1=0,6*, *p2=0,4* (smarkiai persidengiantys klasteriai).
- Kai tiriami 3 klasteriai, *p1=0,9*, *p2=0,5, p3=0,1* (atsiskiriantys klasteriai), *p1=0,75*, *p2=0,5, p3=0,25* (vidutiniškai persidengiantys klasteriai)) ir *p1=0,6*, *p2=0,5, p3=0,4* (smarkiai persidengiantys klasteriai).
- Kai tiriami 4 klasteriai, *p1=0,95*, *p2=0,65, p3=0,35, p4=0,05* (atsiskiriantys klasteriai), *p1=0,85*, *p2=0,6, p3=0,4, p4=0,15* (vidutiniškai persidengiantys klasteriai)) ir *p1=0,7*, *p2=0,55, p3=0,45, p4=0,3* (smarkiai persidengiantys klasteriai).

Tiriant metodų tikslumą priklausomai nuo duomenų savybių vektoriaus ilgio, duomenys buvo modeliuoti su vidutiniškai persidengiančiais klasteriais, o vektoriaus ilgis parinktas 5, 10 ir 20.

Tiriant metodų tikslumą nuo parenkamų klasterių skaičiaus, duomenų savybių vektoriaus ilgis buvo parinktas 15, duomenys buvo modeliuoti su atsiskiriančiais klasteriais, vidutiniškai persidengiančiais klasteriais ir smarkiai persidengiančiais klasteriais, kai klasterių skaičius lygus 2, 3 ir 4.

Nehierarchinių metodų tikslumo tyrime buvo modeliuota 1000 stebėjimų ir keičiami tokie parametrai:

- Klasterių skaičius (2, 3, 4).
- Duomenų savybių vektoriaus ilgis (50, 100, 200).
- Modeliuojamų duomenų mišiniai:

- Kai klasterių skaičius 2, *p1=0,9*, *p2=0,1* (labai gerai atsiskiriantys klasteriai), *p1=0,8*, *p2=0,2* (atsiskiriantys klasteriai), *p1=0,7*, *p2=0,3* (vidutiniškai persidengiantys klasteriai), *p1=0,6*, *p2=0,4* (smarkiai persidengiantys klasteriai) ir *p1=0,55*, *p2=0,45* (labai smarkiai persidengiantys klasteriai).
- Kai klasterių skaičius 3, *p1=0,9*, *p2=0,5, p3=0,1* (labai gerai atsiskiriantys klasteriai), *p1=0,8*, *p2=0,5, p3=0,2* (atsiskiriantys klasteriai), *p1=0,7*, *p2=0,5, p3=0,3* (vidutiniškai persidengiantys klasteriai), *p1=0,6*, *p2=0,5, p3=0,4* (smarkiai persidengiantys klasteriai) ir *p1=0,55*, *p2=0,5, p3=0,45* (labai smarkiai persidengiantys klasteriai).
- Kai klasterių skaičius 4, *p1=0,95*, *p2=0,65, p3=0,35, p4=0,05* (labai gerai atsiskiriantys klasteriai), *p1=0,85*, *p2=0,6, p3=0,4, p4=0,15* (atsiskiriantys klasteriai), *p1=0,75*, *p2=0,6, p3=0,4, p4=0,25* (vidutiniškai persidengiantys klasteriai), *p1=0,65*, *p2=0,55, p3=0,45, p4=0,35* (smarkiai persidengiantys klasteriai) ir *p1=0,58*, *p2=0,52, p3=0,48, p4=0,42* (labai smarkiai persidengiantys klasteriai).

Nehierarchinių metodų klasterių skaičiaus nustatymo tyrime buvo modeliuoti 2, 3 ir 4 klasteriai su blogai atskirtais klasteriais ir tiriama koks klasterių skaičius yra optimalus panaudojant Šarlio kubinį klasterizavimo kriterijų (CCC) bei Pseudo-F Statistiką (PSF).

2.3. DAUGIAMAČIŲ DUOMENŲ TRANSFORMAVIMAS Į ATSTUMŲ MATRICAS

Darbe Monte-Karlo metodu modeliuoti duomenys yra transformuojami į atstumų matricas taikant 10 skirtingų atstumų skaičiavimo būdų. Tam naudojama SAS procedūra *proc DISTANCE*. Vieni atstumų matai skaičiuoja panašumą tarp skirtingų porų, o kiti nepanašumą, todėl reikėjo visus atstumų matus suvienodinti, kad skaičiuotų nepanašumą (atstumą) tarp skirtingų stebėjimo porų.

„Hamming“, „Dmatch“, „Dsqmatch“, „Djaccard“, „Bray and Curtis“ bei „Kulcynski1” atstumų matai skaičiuoja nepanašumą, tuo tarpu „Roger and Tanimoto“, „Sokal and Sneath1”, Dice” bei “Russell and Rao” skaičiuoja panašumą tarp objektų, todėl šiuos 4 atstumų matus reikėjo pakeisti į nepanašumo matus, panaudojus formulę:

$$d = 1 - s \qquad \textbf{(2.2)}$$

čia s – atstumo mato panašumas, o d – atstumo mato nepanašumas.

2.4. METODŲ PAKLAIDŲ NUSTATYMAS

Sumodeliavus dichotominius duomenis ir transformavus juos į atstumų matricas jau galima atlikti klasterizavimą. Klasterizavimo metodų efektyvumas tiriamas pagal paklaidas, kurios gaunamos metodui neteisingai priskyrus esamą klasterį. Tiksliau sakant, kai buvo modeliuojami duomenys jiems pagal skirtingus duomenų mišinius buvo priskirti tikrieji klasteriai, o metodas pagal duomenų tarpusavio panašumą ir nepanašumą priskyrė tam klasterių įverčius. Tada buvo paskaičiuojamas stebėjimų skaičius su neteisingai priskirtais klasteriais ir apskaičiuota jų dalis nuo visų stebėjimų. Kuo paklaida mažesnė, tuo metodas tiksliau priskyrė teisingus klasterius.

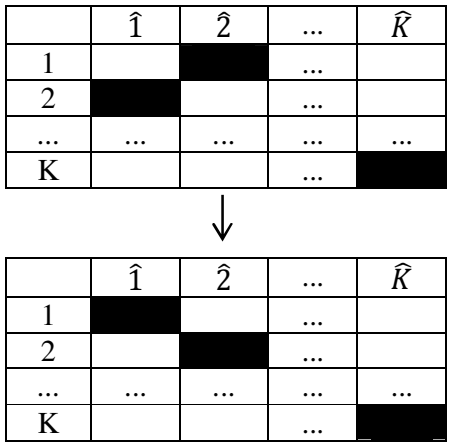
Metodų paklaidos buvo skaičiuojamos pagal šią formulę:

$$\partial = \frac{1}{n}\mathbb{1}\{\partial(x) \neq \hat{\partial}(x)\} \qquad \textbf{(2.3)}$$

kur $\partial(x)$ yra Monte-Karlo metodu sumodeliuotas klasteris pagal skirtingus duomenų mišinius, $\hat{\partial}(x)$ – klasterizavimo metodo priskirtas klasteris, n – stebėjimų skaičius.

Taip pat, metodui priskiriant klasterius buvo susidurta su klasterių tarpusavio sumaišymo problema, t.y. modeliuotų duomenų mišinių klasterių nesutapimas su metodo nustatytu klasteriu. Pvz. Jeigu buvo modeliuoti duomenys su 3 skirtingais duomenų mišiniais ir modeliuojant priskirti 3 klasteriai, pirmam duomenų mišiniui buvo priskirtas pirmas klasteris, antram – antras, o trečiam – trečias. Tuo tarpu klasterizavimo metodas priskirdamas klasterius pirmam duomenų mišiniui priskyrė antrą klasterį, antram duomenų mišiniui – pirmą klasterį, o trečiam – trečią. Matome, kad buvo sumaišyti pirmas su antru klasteriu ir, neatlikus klasterių transformacijų, skaičiuodami metodo paklaidą gausime neteisingą rezultatą.

Klasterių transformacijos buvo atliktos tokiu būdu:



2.2 pav. Klasterių transformacijų schema

2.5. KLASTERIZAVIMO METODŲ REZULTATAI

2.5.1. HIERARCHINIŲ KLASTERIZAVIMO METODŲ PAKLAIDOS

Hierarchinių metodų tyrimą sudaro 3 dalys. Buvo tirtos metodų paklaidos priklausomai nuo:

- modeliuotų duomenų mišinių;
- duomenų savynių vektoriaus ilgio;
- klasterių skaičiaus.

1 dalyje buvo modeliuoti duomenys su gerai atsiskiriančiais klasteriais, su vidutiniškai persidengiančiais klasteriais ir smarkiai persidengiančiais klasteriais (skirstinių mišiniai aprašyti 2.2. skyrelyje). Duomenų savybių vektoriaus ilgis parinktas 10, stebėjimų skaičius – 120, o klasterių skaičius 2, 3 ir 4.

Kai klasterių skaičius lygus 2 ir modeliuoti duomenys su atsiskiriančiais klasteriais, geriausius rezultatus parodė tolimiausio kaimyno („complete“) metodas (2.1. lentelė). Su daugeliu atstumų skaičiavimo būdu metodo paklaida 1,7%. Taip pat gerus rezultatus parodė beta-jautrus („flexible“) bei Ward‘o („ward“) metodai. Jų mažiausia paklaida taip pat 1,7% su tam tikrais atstumų matais. Taip pat geri rezultatai gauti ir su vidutinio atstumo metodu („average“), atstumo tarp centrų („centroid“) bei dviejų pakopų tankių („twostage“) metodais. Tuo tarpu tankių („density“), McQuitt‘y („mcquitty“), medianinis („median“) bei artimiausio kaimyno („single“) metodai parodė labai prastus rezultatus su visais atstumų skaičiavimo būdais.

2.1. lentelė

Metodų paklaidos su atsiskiriančiais klasteriais, kai klasterių skaičius = 2

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	braycurtis	k1
average	3,3%	3,3%	3,3%	3,3%	2,5%	2,5%	2,5%	49,2%	49,2%	49,2%
centroid	3,3%	3,3%	3,3%	3,3%	3,3%	3,3%	3,3%	49,2%	49,2%	49,2%
complete	**1,7%**	**1,7%**	**1,7%**	**1,7%**	**1,7%**	**1,7%**	**1,7%**	49,2%	49,2%	49,2%
density	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%
flexible	5,0%	**1,7%**	5,0%	2,5%	5,0%	5,0%	5,0%	**1,7%**	**1,7%**	**1,7%**
mcquitty	45,8%	45,8%	45,8%	45,8%	45,8%	45,8%	45,8%	49,2%	49,2%	49,2%
median	49,2%	49,2%	49,2%	49,2%	41,7%	41,7%	41,7%	49,2%	49,2%	49,2%
single	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%
twostage	2,5%	4,2%	2,5%	2,5%	2,5%	2,5%	2,5%	49,2%	49,2%	49,2%
ward	3,3%	**1,7%**	3,3%	2,5%	2,5%	2,5%	2,5%	**1,7%**	**1,7%**	**1,7%**

Kai klasterių skaičius lygus 2 ir modeliuoti duomenys su vidutiniškai persidengiančiais klasteriais, geriausius rezultatus parodė beta-jautrus („flexible“) metodas su „Russell and Rao“ atstumų matu. Paklaida lygi 10,0% (2.2. lentelė). Taip pat mažos šio metodo paklaidos ir su

simetriniais atstumų matais. Gerus rezultatus parodė ir vidutinio atstumo („average“) metodas su simetriniais atstumų matais.

2.2. lentelė

Metodų paklaidos su vidutiniškai persidengiančiais klasteriais, kai klasterių skaičius = 2

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	braycurtis	k1
average	**14,2%**	**14,2%**	**14,2%**	**14,2%**	**12,5%**	48,3%	48,3%	49,2%	48,3%	48,3%
centroid	**14,2%**	49,2%	15,0%	49,2%	**14,2%**	48,3%	48,3%	49,2%	48,3%	48,3%
complete	17,5%	17,5%	17,5%	17,5%	17,5%	46,7%	46,7%	48,3%	46,7%	46,7%
density	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%
flexible	**12,5%**	**14,2%**	**12,5%**	**12,5%**	**12,5%**	25,0%	33,3%	**10,0%**	33,3%	33,3%
mcquitty	27,5%	20,8%	30,8%	34,2%	18,3%	46,7%	46,7%	47,5%	46,7%	46,7%
median	48,3%	50,0%	50,0%	49,2%	48,3%	46,7%	48,3%	49,2%	48,3%	48,3%
single	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%
twostage	19,2%	41,7%	20,0%	21,7%	17,5%	40,8%	42,5%	49,2%	42,5%	42,5%
ward	15,8%	15,0%	15,8%	**11,7%**	**13,3%**	15,0%	31,7%	**13,3%**	31,7%	31,7%

Kai klasterių skaičius lygus 2 ir modeliuoti duomenys su smarkiai persidengiančiais klasteriais, geriausius rezultatus parodė tolimiausio kaimyno (“complete”) metodas ir Ward’o (“ward”) metodai su simetriniais atstumų matais. (2.2. lentelė). Tuo tarpu didžiausios paklaidos gautos su atstumo tarp centrų („centroid“), tankių („density“), medianiniu („median“), artimiausio kaimyno („single“) ir dviejų pakopų tankių („twostage“) metodais.

2.3. lentelė

Metodų paklaidos su smarkiai persidengiančiais klasteriais, kai klasterių skaičius = 2

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	braycurtis	k1
average	36,7%	38,3%	36,7%	36,7%	36,7%	48,3%	45,8%	49,2%	45,8%	45,8%
centroid	48,3%	49,2%	48,3%	49,2%	48,3%	49,2%	48,3%	49,2%	48,3%	48,3%
complete	**34,2%**	**34,2%**	**34,2%**	**34,2%**	**34,2%**	47,5%	47,5%	49,2%	47,5%	47,5%
density	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%
flexible	37,5%	37,5%	37,5%	37,5%	**33,3%**	38,3%	42,5%	**30,8%**	42,5%	42,5%
mcquitty	38,3%	38,3%	38,3%	38,3%	43,3%	42,5%	42,5%	48,3%	42,5%	42,5%
median	49,2%	49,2%	49,2%	49,2%	47,5%	49,2%	49,2%	49,2%	49,2%	49,2%
single	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%	49,2%
twostage	49,2%	48,3%	49,2%	49,2%	49,2%	48,3%	49,2%	50,0%	49,2%	49,2%
ward	**30,8%**	**31,7%**	**30,8%**	**33,3%**	37,5%	43,3%	41,7%	47,5%	41,7%	41,7%

Kai klasterių skaičius lygus 3 ir modeliuoti duomenys su atsiskiriančiais klasteriais, geriausius rezultatus parodė beta-jautrus („flexible“) metodas su visais atstumų matais. Mažiausia paklaida 17,5% (2.4. lentelė). Taip pat geri rezultati stebėti ir su Ward‘o („ward) metodu, mažiausia paklaida 20,8% su “Roger and Tanimoto” atstumu. Didžiausios paklaidos gautos su tankių (“density”) bei su artimiausio kaimyno (“single”) metodais, kurios siekia 65,8%.

2.4. lentelė

Metodų paklaidos su atsiskiriančiais klasteriais, kai klasterių skaičius = 3

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	braycurtis	k1
average	28,3%	28,3%	28,3%	28,3%	28,3%	28,3%	28,3%	65,8%	65,8%	65,8%
centroid	33,3%	32,5%	33,3%	31,7%	31,7%	31,7%	31,7%	65,8%	65,8%	65,8%
complete	31,7%	31,7%	31,7%	31,7%	31,7%	31,7%	31,7%	65,8%	65,8%	65,8%
density	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%
flexible	**17,5%**	**18,3%**	**17,5%**	**18,3%**	**17,5%**	**17,5%**	**17,5%**	**20,0%**	**20,0%**	**20,0%**
mcquitty	28,3%	30,0%	31,7%	30,0%	38,3%	38,3%	38,3%	65,8%	65,8%	65,8%
median	60,0%	65,8%	60,0%	63,3%	38,3%	38,3%	38,3%	65,8%	65,8%	65,8%
single	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%	65,8%
twostage	35,0%	35,0%	35,0%	35,0%	35,0%	35,0%	35,0%	65,8%	65,8%	65,8%
ward	24,2%	29,2%	24,2%	20,8%	24,2%	24,2%	24,2%	35,0%	35,0%	35,0%

Kai klasterių skaičius lygus 3 ir modeliuoti duomenys su vidutiniškai persidengiančiais klasteriais, geriausius rezultatus parodė McQuitty‘s („mcquitty“) metodas su „Dmatch“ ir „Roger and Tanimoto“ atstumais (34,2%), tačiau paklaidos su nesimetriniais matais buvo didelės (2.5. lentelė). Taip pat geri rezultatai pastebėti su beta-jautriu („flexible“) metodu ir nesimetriniais atstumų matais bei su Ward‘o („ward“) metodu ir simetriniais atstumų matais.

2.5. lentelė

Metodų paklaidos su vidutiniškai persidengiančiais klasteriais, kai klasterių skaičius = 3

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	braycurtis	k1
average	42,5%	41,7%	42,5%	42,5%	42,5%	63,3%	64,2%	65,0%	64,2%	64,2%
centroid	39,2%	65,8%	39,2%	65,0%	**38,3%**	64,2%	64,2%	65,8%	64,2%	64,2%
complete	40,0%	40,0%	40,0%	40,0%	40,0%	59,2%	59,2%	62,5%	59,2%	59,2%
density	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,8%	65,0%	65,0%
flexible	40,0%	40,8%	40,0%	**37,5%**	43,3%	**37,5%**	**38,3%**	**35,8%**	**38,3%**	**38,3%**
mcquitty	41,7%	**34,2%**	41,7%	**34,2%**	40,8%	63,3%	63,3%	64,2%	63,3%	63,3%
median	65,8%	65,8%	65,8%	65,0%	39,2%	63,3%	63,3%	65,8%	63,3%	63,3%
single	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,8%	65,0%	65,0%
twostage	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,8%	65,0%	65,0%
ward	39,2%	**38,3%**	39,2%	**35,0%**	**36,7%**	41,7%	50,8%	40,8%	50,8%	50,8%

Kai klasterių skaičius lygus 3 ir modeliuoti duomenys su smarkiai persidengiančiais klasteriais, mažiausios paklaidos buvo gautos su Ward’o (“ward”) metodu ir simetriniais atstumų matais (2.6. lentelė). Palyginus geri rezultatai gauti ir su tankių („density“) bei su beta-jautriu („flexible“) metodu.

2.6. lentelė

Metodų paklaidos su smarkiai persidengiančiais klasteriais, kai klasterių skaičius = 3

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	braycurtis	k1
average	63,3%	58,3%	63,3%	61,7%	**52,5%**	62,5%	61,7%	65,0%	61,7%	61,7%
centroid	62,5%	65,8%	62,5%	65,0%	58,3%	65,0%	63,3%	65,0%	63,3%	63,3%
complete	57,5%	57,5%	57,5%	57,5%	57,5%	61,7%	61,7%	64,2%	61,7%	61,7%
density	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%	65,0%
flexible	**56,7%**	**53,3%**	**56,7%**	59,2%	57,5%	58,3%	61,7%	59,2%	61,7%	61,7%
mcquitty	59,2%	65,0%	60,0%	65,0%	60,0%	60,8%	59,2%	60,8%	59,2%	59,2%
median	61,7%	65,8%	61,7%	65,8%	60,0%	65,0%	60,0%	65,8%	60,0%	60,0%
single	65,8%	65,8%	65,8%	65,8%	65,8%	65,0%	65,0%	65,8%	65,0%	65,0%
twostage	**56,7%**	65,0%	58,3%	**56,7%**	**55,8%**	60,8%	62,5%	65,0%	62,5%	62,5%
ward	**55,0%**	**55,8%**	**55,0%**	**53,3%**	**57,5%**	**60,0%**	**60,8%**	**56,7%**	**60,8%**	**60,8%**

Kai klasterių skaičius lygus 4 ir modeliuoti duomenys su atsiskiriančiais klasteriais, geriausius rezultatus parodė McQuitty‘s („mcquitty“) metodas su „Dsqmatch“ atstumų matu 30,0% (2.7. lentelė), tačiau su nesimetriniais atstumų matais šio metodo paklaidos buvo labai didelės. Su visais atstumų matais gerus rezultatus parodė beta-jautrus („flexible“) bei Ward‘o („ward“) metodai. Tuo tarpu blogiausi rezultatai gauti su tankių („density“) ir su artimiausio kaimyno („single“) metodu.

2.7. lentelė

Metodų paklaidos su atsiskiriančiais klasteriais, kai klasterių skaičius = 4

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	braycurtis	k1
average	42,5%	39,2%	42,5%	42,5%	43,3%	43,3%	43,3%	74,2%	74,2%	74,2%
centroid	47,5%	48,3%	47,5%	46,7%	41,7%	41,7%	41,7%	74,2%	74,2%	74,2%
complete	42,5%	42,5%	42,5%	42,5%	42,5%	42,5%	42,5%	74,2%	74,2%	74,2%
density	73,3%	74,2%	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%
flexible	36,7%	**34,2%**	36,7%	35,8%	**35,0%**	**35,0%**	**35,0%**	**32,5%**	**32,5%**	**32,5%**
mcquitty	**30,0%**	**33,3%**	**30,0%**	**33,3%**	45,0%	45,0%	45,0%	74,2%	74,2%	74,2%
median	45,8%	71,7%	45,8%	72,5%	45,0%	45,0%	45,0%	74,2%	74,2%	74,2%
single	74,2%	74,2%	74,2%	74,2%	74,2%	74,2%	74,2%	74,2%	74,2%	74,2%
twostage	48,3%	49,2%	48,3%	48,3%	48,3%	48,3%	48,3%	73,3%	73,3%	73,3%
ward	36,7%	35,8%	36,7%	**30,8%**	**35,0%**	**35,0%**	**35,0%**	52,5%	52,5%	52,5%

Kai klasterių skaičius lygus 4 ir modeliuoti duomenys su vidutiniškai persidengiančiais klasteriais, geriausius rezultatus parodė beta-jautrus („flexible“) metodas su tam tikrais nesimetriniais atstumų matais bei McQuitty‘s („mcquitty“) su tam tikrais simetriniais atstumų matais (2.8. lentelė).

2.8. lentelė

Metodų paklaidos su vidutiniškai persidengiančiais klasteriais, kai klasterių skaičius = 4

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	braycurtis	k1
average	45,0%	**41,7%**	45,0%	44,2%	49,2%	49,2%	49,2%	74,2%	74,2%	74,2%
centroid	47,5%	53,3%	50,0%	50,8%	47,5%	47,5%	47,5%	74,2%	74,2%	74,2%
complete	44,2%	44,2%	44,2%	44,2%	44,2%	44,2%	44,2%	74,2%	74,2%	74,2%
density	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%
flexible	45,0%	45,8%	45,0%	45,0%	45,0%	45,0%	45,0%	**39,2%**	**39,2%**	**39,2%**
mcquitty	**43,3%**	45,8%	**41,7%**	45,8%	45,8%	45,8%	45,8%	74,2%	74,2%	74,2%
median	47,5%	72,5%	47,5%	63,3%	49,2%	49,2%	49,2%	74,2%	74,2%	74,2%
single	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	74,2%	74,2%	74,2%
twostage	48,3%	49,2%	48,3%	48,3%	48,3%	48,3%	48,3%	73,3%	73,3%	73,3%
ward	45,0%	49,2%	45,0%	**43,3%**	**42,5%**	**42,5%**	**42,5%**	51,7%	51,7%	51,7%

Kai klasterių skaičius lygus 4 ir modeliuoti duomenys su smarkiai persidengiančiais klasteriais, mažiausios paklaidos gautos su beta-jautriu (“flexible”) metodu (52,5%) (2.9 lentelė). Taip pat gerus rezultatus parodė tas pats beta-jautrus („flexible“) metodas ir tolimiausio kaimyno („complete“) metodas su simetriniais atstumų matais.

2.9. lentelė

Metodų paklaidos su smarkiai persidengiančiais klasteriais, kai klasterių skaičius = 4

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	braycurtis	k1
average	61,7%	59,2%	61,7%	59,2%	61,7%	72,5%	72,5%	74,2%	72,5%	72,5%
centroid	59,2%	73,3%	59,2%	72,5%	**56,7%**	73,3%	71,7%	74,2%	71,7%	71,7%
complete	**57,5%**	**57,5%**	**57,5%**	**57,5%**	**57,5%**	69,2%	69,2%	71,7%	69,2%	69,2%
density	72,5%	72,5%	72,5%	72,5%	72,5%	74,2%	74,2%	72,5%	74,2%	74,2%
flexible	**52,5%**	61,7%	**52,5%**	**52,5%**	**57,5%**	60,8%	65,0%	**52,5%**	65,0%	65,0%
mcquitty	58,3%	59,2%	58,3%	59,2%	61,7%	68,3%	69,2%	69,2%	69,2%	69,2%
median	67,5%	71,7%	67,5%	70,8%	61,7%	67,5%	66,7%	74,2%	66,7%	66,7%
single	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	73,3%	74,2%	73,3%	73,3%
twostage	69,2%	72,5%	69,2%	71,7%	72,5%	74,2%	74,2%	72,5%	74,2%	74,2%
ward	58,3%	61,7%	58,3%	58,3%	**57,5%**	63,3%	63,3%	61,7%	63,3%	63,3%

Apibendrinus rezultatus galime pastebėti, kad duomenų mišinių parinkimas ir klasterių priskyrimas pagal duomenų mišinius turi labai didelę įtaką metodų paklaidoms. Kuo klasterių modeliuotų duomenų mišiniai labiau skiriasi pagal binominio skirstinio tikimybes, tuo metodams “lengviau sekasi” nustatyti tikruosius klasterius ir paklaidos su gerai atskirtais klasteriais gaunamos mažesnės nei su blogai atskirtais klasteriais. 2.10. lentelėje matome metodų paklaidų priklausomybę nuo modeliuotų duomenų mišinių parinkimo. Paklaidos buvo apskaičiuotos kaip vidurkis tarp visų atstumo matų.

2.10. lentelė

Metodų paklaidų priklausomybė nuo duomenų mišinių parinkimo

Paklaidos	Atsikiriantys klasteriai (2)	Vidutiniš persidengiantys klasteriai (2)	Smarkiai persidengiantys klasteriai (2)	Atsikiriantys klasteriai (3)	Vidutiniš persidengiantys klasteriai (3)	Smarkiai persidengiantys klasteriai (3)	Atsikiriantys klasteriai (4)	Vidutiniš persidengiantys klasteriai (4)	Smarkiai persidengiantys klasteriai (4)
average	16,8%	31,2%	42,0%	39,6%	53,3%	61,2%	51,9%	54,6%	66,8%
centroid	17,1%	38,4%	48,7%	42,3%	57,0%	63,4%	53,8%	56,7%	68,3%
complete	15,9%	32,3%	41,0%	41,9%	49,9%	59,8%	52,0%	53,2%	63,6%
density	49,2%	49,2%	49,2%	65,8%	65,1%	65,0%	73,4%	73,3%	73,2%
flexible	3,4%	19,9%	38,0%	18,4%	39,0%	58,6%	34,6%	43,3%	58,5%
mcquitty	46,8%	36,6%	41,5%	43,2%	51,0%	60,8%	48,4%	53,7%	64,2%
median	46,9%	48,7%	49,0%	56,2%	62,1%	62,6%	59,3%	60,1%	68,1%
single	49,2%	49,2%	49,2%	65,8%	65,1%	65,5%	74,2%	73,6%	73,4%
twostage	16,7%	33,8%	49,1%	44,2%	65,1%	60,6%	55,9%	55,9%	72,4%
ward	2,3%	19,5%	38,0%	27,6%	42,3%	57,6%	40,3%	46,5%	60,9%

2 dalyje buvo tiriama kaip keičiasi metodų paklaidos nuo duomenų savybių vektoriaus ilgio. Pavyzdžiui turime duomenų masyvą su 120 stebėjimų ir mus domina ar turint kiekvieno stebėjimo vis daugiau savybių metodų paklaidos didėja ar mažėja. Stebėjimų skaičius buvo parinktas 120 ir stebėjimai modeliuoti su vidutiniškai persidengiančiais klasteriais. Kaip matome iš 2.11. lentelės daugeliui metodų savybių skaičiaus padidėjimas lemia mažesnes paklaidas.

2.11. lentelė

Metodų paklaidų priklausomybė nuo duomenų savybių vektoriaus ilgio

Paklaidos	Savybių skaičius =5 klasterių skaičius =2	Savybių skaičius =10 klasterių skaičius =2	Savybių skaičius =15 klasterių skaičius =2	Savybių skaičius =5 klasterių skaičius =3	Savybių skaičius =10 klasterių skaičius =3	Savybių skaičius =15 klasterių skaičius =3	Savybių skaičius =5 klasterių skaičius =4	Savybių skaičius =10 klasterių skaičius =4	Savybių skaičius =15 klasterių skaičius =4
average	29,3%	31,2%	29,4%	48,3%	53,3%	46,9%	53,2%	54,6%	50,8%
centroid	29,0%	38,4%	36,7%	48,5%	57,0%	56,4%	54,2%	56,7%	61,6%
complete	33,4%	32,3%	30,4%	53,0%	49,9%	54,7%	54,9%	53,2%	51,4%
density	35,8%	49,2%	49,2%	65,2%	65,1%	65,7%	68,1%	73,3%	73,3%
flexible	25,7%	19,9%	15,3%	42,4%	39,0%	34,9%	42,5%	43,3%	32,6%
mcquitty	41,6%	36,6%	31,0%	47,7%	51,0%	45,9%	54,9%	53,7%	52,6%
median	32,2%	48,7%	41,5%	49,4%	62,1%	64,4%	57,4%	60,1%	63,0%
single	49,2%	49,2%	49,2%	65,8%	65,1%	65,8%	64,3%	73,6%	74,2%
twostage	35,8%	33,8%	27,8%	55,9%	65,1%	62,5%	61,7%	55,9%	56,8%
ward	24,3%	19,5%	19,6%	40,6%	42,3%	38,9%	44,6%	46,5%	38,9%

3 dalyje buvo tiriama kaip keičiasi metodų paklaidos nuo parenkamų klasterių skačiaus. Stebėjimų skaičius buvo parinktas 120, duomenų savybių vektoriaus ilgis 15, ir tirta metodų paklaidų priklausomybė nuo klasterių skaičiaus kai modeliuoti duomenys su atsiskiriančiais klasteriais,

vidutiniškai persidengiančiais klasteriais ir smarkiai persidengiančiais klasteriais. Kaip matome iš 2.12. lentelės klasterių skaičiaus parinkimas turi labai didelę įtaką metodų paklaidoms. Mažiausios paklaidos gaunamos kai tiriami 2 klasteriai, o blogiausios kai tiriami 4 klasteriai.

2.12. lentelė

Metodų paklaidų priklausomybė nuo klasterių skaičiaus parinkimo

Paklaidos	Klasterių skaičius =2, atskiriantys klasteriai	Klasterių skaičius =3, atskiriantys klasteriai	Klasterių skaičius =4, atskiriantys klasteriai	Klasterių skaičius =2, vidutin. persidengiantys klasteriai	Klasterių skaičius =3, vidutin. persidengiantys klasteriai	Klasterių skaičius =4, vidutin. persidengiantys klasteriai	Klasterių skaičius =2, smarkiai persidengiantys klasteriai	Klasterių skaičius =3, smarkiai persidengiantys klasteriai	Klasterių skaičius =4, smarkiai persidengiantys klasteriai
average	15,3%	38,4%	49,6%	29,4%	46,9%	50,8%	39,5%	60,8%	60,3%
centroid	15,2%	41,6%	54,3%	36,7%	56,4%	61,6%	48,8%	65,1%	70,7%
complete	15,3%	36,1%	48,5%	30,4%	54,7%	51,4%	43,4%	60,5%	61,1%
density	49,2%	65,8%	73,6%	49,2%	65,7%	73,3%	49,2%	65,7%	73,9%
flexible	0,7%	10,3%	31,3%	15,3%	34,9%	32,6%	39,1%	57,9%	57,4%
mcquitty	15,6%	40,2%	50,6%	31,0%	45,9%	52,6%	40,6%	60,3%	59,6%
median	48,8%	54,5%	64,8%	41,5%	64,4%	63,0%	49,1%	64,3%	72,0%
single	49,2%	65,8%	74,2%	49,2%	65,8%	74,2%	49,2%	65,8%	73,5%
twostage	15,6%	42,5%	56,7%	27,8%	62,5%	56,8%	44,1%	60,5%	69,8%
ward	0,9%	20,2%	33,8%	19,6%	38,9%	38,9%	34,8%	58,0%	55,7%

2.5.2. NEHIERARCHINIŲ KLASTERIZAVIMO METODŲ PAKLAIDOS

Nehierarchinių metodų tikslumo tyrimas susidėjo iš 2 dalių. Pirmoje dalyje buvo tiriama k-vidurkių metodo paklaidų priklausomybė nuo klasterių skaičiaus, duomenų savybių vektoriaus ilgio bei modeliuojamų duomenų mišinių. Tiriant metodo paklaidas buvo modeliuota 1000 stebėjimų ir keičiami tokie parametrai:

- klasterių skaičius (2, 3, 4);
- duomenų savybių vektoriaus ilgis (50, 100, 200);
- modeliuojamų duomenų mišiniai, kurių tikimybės aprašytos 2.2. skyriuje.

Kaip matome iš 2.13. lentelės didėjant klasterių skaičiui metodo paklaidos didėja. Didėjant duomenų savybių vektoriaus ilgiui metodų paklaidos mažėja. O priklausomybė nuo modeliuotų duomenų mišinių ir jiems priskirtų klasterių labai pastebima. Kuo klasteriai labiau atskirti vienas nuo kito, tuo metodo paklaidos mažesnės.

2.13. lentelė

k-vidurkių metodo paklaidų tyrimas

	Klasterių skaičius =2, savybių skaičius =50	Klasterių skaičius =3, savybių skaičius =50	Klasterių skaičius =4, savybių skaičius =50	Klasterių skaičius =2, savybių skaičius =100	Klasterių skaičius =3, savybių skaičius =100	Klasterių skaičius =4, savybių skaičius =100	Klasterių skaičius =2, savybių skaičius =200	Klasterių skaičius =3, savybių skaičius =200	Klasterių skaičius =4, savybių skaičius =200
Labai gerai atsiskiriantys klasteriai	0,0%	0,0%	0,4%	0,0%	0,0%	0,0%	0,0%	0,0%	0,0%
Atsiskiriantys klasteriai	0,0%	0,4%	2,7%	0,0%	0,1%	0,2%	0,0%	0,0%	0,1%
Vidut. Persidengiantys klasteriai	0,0%	2,6%	12,8%	0,0%	0,5%	18,7%	0,0%	0,1%	1,5%
Smarkiai persidengiantys klasteriai	2,5%	19,0%	17,8%	0,5%	7,8%	37,1%	0,2%	3,2%	12,4%
Labai smarkiai persidengiantys klasteriai	7,7%	32,3%	38,2%	3,8%	31,4%	32,1%	2,6%	21,7%	38,9%

Antroje dalyje buvo tiriamas tinkamas klasterių skaičiaus parinkimas ir analizuojami Šarlio kubinis klasterizavimo kriterijus (CCC) bei Pseudo-F Statistika (PSF). Iš 2.14 lentelės matome, kad CCC parametras yra tikslesnis.

2.14. lentelė

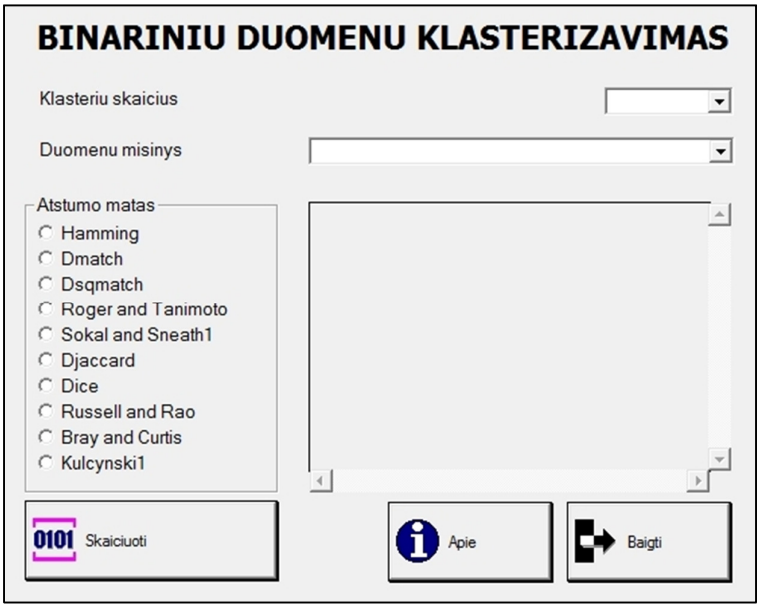
Klasterių skaičiaus nustatymo tyrimas

	CCC, Klasterių skaičius =2	PSF, Klasterių skaičius =2	CCC, Klasterių skaičius =3	PSF, Klasterių skaičius =3	CCC, Klasterių skaičius =4	PSF, Klasterių skaičius =4
2 klasteriai	**223,58**	**52,54**	51,161	**44,88**	49,067	**83,29**
3 klasteriai	160,142	29,55	**57,648**	28,33	57,601	48,9
4 klasteriai	134,885	22,01	46,979	19,59	59,421	34,63
5 klasteriai	117,26	17,79	49,276	16,63	**64,09**	28,09
6 klasteriai	105,821	15,21	49,82	14,58	62,214	23,09

3. PROGRAMINĖ REALIZACIJA IR INSTRUKCIJA VARTOTOJUI

SAS programos pagalba sukurti algoritmai, skirti analizuoti hierarchinius ir nehierarchinius klasterizavimo algoritmus. Kadangi darbo tikslas buvo nustatyti klasterizavimo metodų efektyvumą esant įvairioms duomenų aibėms, todėl modeliuota daug skirtingų duomenų mišinių. Taip pat tiriama metodų efektyvumas nuo pasirinktų klasterių skaičiaus, nuo duomenų savybių vektoriaus ilgio, nuo taikomo atstumų mato.

Vartotojui šiuos algoritmus nėra patogu naudoti, kadangi reikia keisti tam tikrus parametrus. Todėl buvo sukurta sąsaja su vartotoju, kuri apjungia ir valdo tyrimui naudojamas SAS makro komandas. Vartotojas gali pasirinkti kiek klasterių norima analizuoti, kokius duomenų mišinius naudoti bei kokius atstumų matus taikyti nustatant klasterizavimo metodų efektyvumą.



3.1 pav. Vartotojo sąsaja

Vartotojas norėdamas atlikti tyrimą turi pasirinkti keletą tyrimo parametrų:

1. klasterių skaičių;
2. modeliuotų duomenų mišinius;
3. taikomą atstumo matą.

Klasterių skaičiaus tyrimas apima 2 klasterių tyrimą, 3 klasterių tyrimą bei 4 klasterių tyrimą. Modeliuojamus duomenis galima pasirinkti iš 3 mišinių tipų: su atsiskiriančiais klasteriais, su vidutiniškai persidengiančiais klasteriais bei su smarkiai persidengiančiais klasteriais. Taikomą atstumo matą vartotojas gali pasirinkti iš 10 siūlomų simetrinių ir nesimetrinių atstumo matų:

„Hamming“, „Dmatch“, „Dsqmatch“, Roger and Tanimoto“, „Sokal and Sneath1”, “Djaccard”, “Dice”, Russell and Rao”, Bray and Curtis” bei “Kulcynski1”.

Pasirinkus parametrus, vartotojui parodoma, kuris metodas yra tinkamiausias su pasirinktu atstumų matu bei esant tam tikriems duomenims ir klasterių skaičiui.

Jei vartotojas nori atlikti skaičiavimus nepasirinkęs parametrų, tuomet į ekraną išvedami pranešimai informuojantys, kad vartotojas privalo pasirinkti parametrus.

Taip pat sukurta informacinė programos dalis, kurioje pateikiama informacija apie sukurtą programą.

Norint paleisti programą, reikia programos „BinaryClustering“ katalogą nukopijuoti į C diską. Programa paleidžiama iš *binaryclustering.lnk* failo.

3.1 lentelė

Sukurtos programos ir jų paskirtis

Programos vardas	Paskirtis
BinaryClustering	Pagrindinė programos dalis, kuri valdo visas posistemes.
Apie	Informuoja vartotoją apie programą.
**.sas*	Atlieka algoritmus pagal pasirinktus parametrus.

IŠVADOS

1. Atlikus hierarchinių klasterizavimo metodų tikslumo tyrimą binariniams duomenimis buvo parodyta, kad efektyviausi metodai yra beta-jautrus („flexible“) bei Ward‘o („ward“). Jų gautos paklaidos yra mažiausios palyginus su kitais hierarchiniais metodais. Metodai buvo efektyvūs ir su simetriniais atstumų matais, ir su nesimetriniais atstumų matais.
2. Didžiausios paklaidos gautos atliekant klasterizavimą su tankių („density“) ir artimiausio kaimyno („single“) metodu, kurių efektyvumas aptinkant teisingus klasterius binariniams duomenims yra labai mažas.
3. Atliktas tyrimas nustatyti kaip skiriasi hierarchinių metodų paklaidos, kai modeliuojami duomenų mišiniai su atsiskiriančiais klasteriais, vidutiniškai persidengiančiais klasteriais ir smarkiai persidengiančiais klasteriais. Metodų paklaidų skirtumas tarp šių mišinių yra labai ženklus ir geriausi rezultatai gaunami su modeliuotais duomenų mišiniais, kuriuose klasteriai yra gerai atskirti.
4. Atliktas tyrimas nustatyti kaip skiriasi hierarchinių metodų paklaidos, kai modeliuojant duomenis parenkamas tam tikras klasterių skaičius, kuris priskiriamas pagal duomenų mišinių pasiskirstymo tankius. Esant 2 klasteriams paklaidos yra mažesnės nei esant 3 klasteriams, analogiškai esant 3 klasteriams paklaidos mažesnės nei esant 4 klasteriams.
5. Nustatyta, kad didinant duomenų savybių vektoriaus ilgį modeliuotiems duomenims, klasterizavimo metodų tikslumas gerėja ir paklaidos mažėja.
6. Atlikus nehierarchinių klasterizavimo metodų analizę nustatyta, kad metodų paklaidos mažesnės tada, kai tiriamas mažesnis klasterių skaičius, kai modeliuotų duomenų mišiniai yra su atsiskiriančiais klasteriais bei kai parenkamas didesnis duomenų savybių vektoriaus ilgis.

REKOMENDACIJOS

Atliekant magistrinį darbą šia tematika buvo įvykdytos visos apsibrėžtos užduotys, tačiau norint labiau įsigilinti į klasterizavimo metodų binariniams duomenims efektyvumą yra galimi tolesni tyrimai. Galima būtų pabandyti sukurti tam tikrą klasių skaičiaus nustatymo algoritmą ar pasiūlyti naują atstumo matą su kuriuo geriau veiktų tam tikri klasterizavimo metodai. Taip pat galima atlikti platesnius tyrimus kaip metodų paklaidos reaguoja smarkiai didinant klasterių skaičių.

Kita galimybė tolimesniems tyrimams būtų ieškoti realių binarinių duomenų, jiems pritaikyti klasterizavimo algoritmus ir žiūrėti ar tinkamai buvo nustatyti klasteriai. Taip pat galima būtų transformuoti nebinarinius duomenis į binarinius pagal tam tikrus algoritmus. Pvz. nustatyti tam tikras sąlygas pagal kurias duomenis vienu atveju būtų paverčiami į 1, o kitu atveju į 0.

Šiame darbe didesnis dėmesys skirtas hierarchiniams klasterizavimo algoritmams, tačiau taip pat įdomu būtų giliau paanalizuoti kitus nehierarchinius metodus, kurie nebuvo nagrinėti šiame darbe, tokius kaip neparametrinis klasterizavimas ar kitus.

ŠALTINIAI IR LITERATŪRA

1. Applied clustering Techniques Course Notes by David Yeo, 2003.
2. SAS® Programming I: Essentials Course Notes by Michelle Buchecker, Sarah Calhoun and Larry Stewart, 2004.
3. SAS® Programming II: Manipulating Data with the DATA Step Course Notes by Jemshaid Cheema and Melinda Thielbar, 2004.
4. SAS® Macro Language Course Notes by Jim Simon, 2004.
5. SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc. 209-246
6. SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc. 1483-1531
7. Tao Li, A General Model for Clustering Binary Data, *KDD '05 Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining.*
8. Tao Li ,Shenghuo Zhu, On Clustering Binary Data, *International Conference On Data Mining (SDM'05)*, 2005
9. Tao Li, A Unified View On Clustering Binary Data, *Volume 62, Number 3 (2006), 199-215, DOI: 10.1007/s10994-005-5316-9*
10. Šmidtaitė, R. NETIESINIŲ STATISTIKŲ TAIKYMAS ATSITIKTINIŲ VEKTORIŲ PASISIKIRSTYMO TANKIŲ VERTINIME, Magistro darbas, vadovas dr. T. Ruzgas, Kauno technologijos universitetas, taikomosios matematikos katedra, 2008, p. 136.
11. http://lt.wikipedia.org/wiki/Monte_Karlo_metodas
12. http://www.math24.lt/teorija,monte-karlo-metodas.html
13. http://lt.wikipedia.org/wiki/Matavimimų_skalė

1 PRIEDAS. REZULTATŲ LENTELĖS

Rezultatai, kai buvo tiriamas hierarchinių metodų tikslumas, priklausomai nuo modeliuotų duomenų mišinių parinkimo, kai duomenų savybių vektoriaus ilgis lygus 10.

2 klasteriai:

1 lentelė

„Hamming“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,03333	0,14167	0,36667
centroid	0,03333	0,14167	0,48333
complete	0,01667	0,175	0,34167
density	0,49167	0,49167	0,49167
flexible	0,05	0,125	0,375
mcquitty	0,45833	0,275	0,38333
median	0,49167	0,48333	0,49167
single	0,49167	0,49167	0,49167
twostage	0,025	0,19167	0,49167
ward	0,03333	0,15833	0,30833

2 lentelė

„Dmatch“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,03333	0,14167	0,38333
centroid	0,03333	0,49167	0,49167
complete	0,01667	0,175	0,34167
density	0,49167	0,49167	0,49167
flexible	0,01667	0,14167	0,375
mcquitty	0,45833	0,20833	0,38333
median	0,49167	0,5	0,49167
single	0,49167	0,49167	0,49167
twostage	0,04167	0,41667	0,48333
ward	0,01667	0,15	0,31667

3 lentelė

„Dsqmatch“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,03333	0,14167	0,36667
centroid	0,03333	0,15	0,48333
complete	0,01667	0,175	0,34167
density	0,49167	0,49167	0,49167
flexible	0,05	0,125	0,375
mcquitty	0,45833	0,30833	0,38333
median	0,49167	0,5	0,49167
single	0,49167	0,49167	0,49167
twostage	0,025	0,2	0,49167
ward	0,03333	0,15833	0,30833

4 lentelė

„Roger and Tanimoto“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,03333	0,14167	0,36667
centroid	0,03333	0,49167	0,49167
complete	0,01667	0,175	0,34167
density	0,49167	0,49167	0,49167
flexible	0,025	0,125	0,375
mcquitty	0,45833	0,34167	0,38333
median	0,49167	0,49167	0,49167
single	0,49167	0,49167	0,49167
twostage	0,025	0,21667	0,49167
ward	0,025	0,11667	0,33333

5 lentelė

„Sokal and Sneath1“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,025	0,125	0,36667
centroid	0,03333	0,14167	0,48333
complete	0,01667	0,175	0,34167
density	0,49167	0,49167	0,49167
flexible	0,05	0,125	0,33333
mcquitty	0,45833	0,18333	0,43333
median	0,41667	0,48333	0,475
single	0,49167	0,49167	0,49167
twostage	0,025	0,175	0,49167
ward	0,025	0,13333	0,375

6 lentelė

„Djaccard“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,025	0,48333	0,48333
centroid	0,03333	0,48333	0,49167
complete	0,01667	0,46667	0,475
density	0,49167	0,49167	0,49167
flexible	0,05	0,25	0,38333
mcquitty	0,45833	0,46667	0,425
median	0,41667	0,46667	0,49167
single	0,49167	0,49167	0,49167
twostage	0,025	0,40833	0,48333
ward	0,025	0,15	0,43333

7 lentelė

„Dice“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,025	0,48333	0,45833
centroid	0,03333	0,48333	0,48333
complete	0,01667	0,46667	0,475
density	0,49167	0,49167	0,49167
flexible	0,05	0,33333	0,425
mcquitty	0,45833	0,46667	0,425
median	0,41667	0,48333	0,49167
single	0,49167	0,49167	0,49167
twostage	0,025	0,425	0,49167
ward	0,025	0,31667	0,41667

8 lentelė

„Russell and Rao“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,49167	0,49167	0,49167
centroid	0,49167	0,49167	0,49167
complete	0,49167	0,48333	0,49167
density	0,49167	0,49167	0,49167
flexible	0,01667	0,1	0,30833
mcquitty	0,49167	0,475	0,48333
median	0,49167	0,49167	0,49167
single	0,49167	0,49167	0,49167
twostage	0,49167	0,49167	0,5
ward	0,01667	0,13333	0,475

9 lentelė

„Bray and Curtis“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,49167	0,48333	0,45833
centroid	0,49167	0,48333	0,48333
complete	0,49167	0,46667	0,475
density	0,49167	0,49167	0,49167
flexible	0,01667	0,33333	0,425
mcquitty	0,49167	0,46667	0,425
median	0,49167	0,48333	0,49167
single	0,49167	0,49167	0,49167
twostage	0,49167	0,425	0,49167
ward	0,01667	0,31667	0,41667

10 lentelė

„Kulcynski1“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,49167	0,48333	0,45833
centroid	0,49167	0,48333	0,48333
complete	0,49167	0,46667	0,475
density	0,49167	0,49167	0,49167
flexible	0,01667	0,33333	0,425
mcquitty	0,49167	0,46667	0,425
median	0,49167	0,48333	0,49167
single	0,49167	0,49167	0,49167
twostage	0,49167	0,425	0,49167
ward	0,01667	0,31667	0,41667

3 klasteriai:

11 lentelė

„Hamming“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,28333	0,425	0,63333
centroid	0,33333	0,39167	0,625
complete	0,31667	0,4	0,575
density	0,65833	0,65	0,65
flexible	0,175	0,4	0,56667
mcquitty	0,28333	0,41667	0,59167
median	0,6	0,65833	0,61667
single	0,65833	0,65	0,65833
twostage	0,35	0,65	0,56667
ward	0,24167	0,39167	0,55

12 lentelė

„Dmatch“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,28333	0,41667	0,58333
centroid	0,325	0,65833	0,65833
complete	0,31667	0,4	0,575
density	0,65833	0,65	0,65
flexible	0,18333	0,40833	0,53333
mcquitty	0,3	0,34167	0,65
median	0,65833	0,65833	0,65833
single	0,65833	0,65	0,65833
twostage	0,35	0,65	0,65
ward	0,29167	0,38333	0,55833

13 lentelė

„Dsqmatch“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,28333	0,425	0,63333
centroid	0,33333	0,39167	0,625
complete	0,31667	0,4	0,575
density	0,65833	0,65	0,65
flexible	0,175	0,4	0,56667
mcquitty	0,31667	0,41667	0,6
median	0,6	0,65833	0,61667
single	0,65833	0,65	0,65833
twostage	0,35	0,65	0,58333
ward	0,24167	0,39167	0,55

14 lentelė

„Roger and Tanimoto“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,28333	0,425	0,61667
centroid	0,31667	0,65	0,65
complete	0,31667	0,4	0,575
density	0,65833	0,65	0,65
flexible	0,18333	0,375	0,59167
mcquitty	0,3	0,34167	0,65
median	0,63333	0,65	0,65833
single	0,65833	0,65	0,65833
twostage	0,35	0,65	0,56667
ward	0,20833	0,35	0,53333

15 lentelė

„Sokal and Sneath1“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,28333	0,425	0,525
centroid	0,31667	0,38333	0,58333
complete	0,31667	0,4	0,575
density	0,65833	0,65	0,65
flexible	0,175	0,43333	0,575
mcquitty	0,38333	0,40833	0,6
median	0,38333	0,39167	0,6
single	0,65833	0,65	0,65833
twostage	0,35	0,65	0,55833
ward	0,24167	0,36667	0,575

16 lentelė

„Djaccard“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,28333	0,63333	0,625
centroid	0,31667	0,64167	0,65
complete	0,31667	0,59167	0,61667
density	0,65833	0,65	0,65
flexible	0,175	0,375	0,58333
mcquitty	0,38333	0,63333	0,60833
median	0,38333	0,63333	0,65
single	0,65833	0,65	0,65
twostage	0,35	0,65	0,60833
ward	0,24167	0,41667	0,6

17 lentelė

„Dice“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,28333	0,64167	0,61667
centroid	0,31667	0,64167	0,63333
complete	0,31667	0,59167	0,61667
density	0,65833	0,65	0,65
flexible	0,175	0,38333	0,61667
mcquitty	0,38333	0,63333	0,59167
median	0,38333	0,63333	0,6
single	0,65833	0,65	0,65
twostage	0,35	0,65	0,625
ward	0,24167	0,50833	0,60833

18 lentelė

„Russell and Rao“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,65833	0,65	0,65
centroid	0,65833	0,65833	0,65
complete	0,65833	0,625	0,64167
density	0,65833	0,65833	0,65
flexible	0,2	0,35833	0,59167
mcquitty	0,65833	0,64167	0,60833
median	0,65833	0,65833	0,65833
single	0,65833	0,65833	0,65833
twostage	0,65833	0,65833	0,65
ward	0,35	0,40833	0,56667

19 lentelė

„Bray and Curtis“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,65833	0,64167	0,61667
centroid	0,65833	0,64167	0,63333
complete	0,65833	0,59167	0,61667
density	0,65833	0,65	0,65
flexible	0,2	0,38333	0,61667
mcquitty	0,65833	0,63333	0,59167
median	0,65833	0,63333	0,6
single	0,65833	0,65	0,65
twostage	0,65833	0,65	0,625
ward	0,35	0,50833	0,60833

20 lentelė

„Kulcynski1“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,65833	0,64167	0,61667
centroid	0,65833	0,64167	0,63333
complete	0,65833	0,59167	0,61667
density	0,65833	0,65	0,65
flexible	0,2	0,38333	0,61667
mcquitty	0,65833	0,63333	0,59167
median	0,65833	0,63333	0,6
single	0,65833	0,65	0,65
twostage	0,65833	0,65	0,625
ward	0,35	0,50833	0,60833

4 klasteriai:

21 lentelė

„Hamming“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,425	0,45	0,61667
centroid	0,475	0,475	0,59167
complete	0,425	0,44167	0,575
density	0,73333	0,73333	0,725
flexible	0,36667	0,45	0,525
mcquitty	0,3	0,43333	0,58333
median	0,45833	0,475	0,675
single	0,74167	0,73333	0,73333
twostage	0,48333	0,48333	0,69167
ward	0,36667	0,45	0,58333

22 lentelė

„Dmatch“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,39167	0,41667	0,59167
centroid	0,48333	0,53333	0,73333
complete	0,425	0,44167	0,575
density	0,74167	0,73333	0,725
flexible	0,34167	0,45833	0,61667
mcquitty	0,33333	0,45833	0,59167
median	0,71667	0,725	0,71667
single	0,74167	0,73333	0,73333
twostage	0,49167	0,49167	0,725
ward	0,35833	0,49167	0,61667

23 lentelė

„Dsqmatch“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,425	0,45	0,61667
centroid	0,475	0,5	0,59167
complete	0,425	0,44167	0,575
density	0,73333	0,73333	0,725
flexible	0,36667	0,45	0,525
mcquitty	0,3	0,41667	0,58333
median	0,45833	0,475	0,675
single	0,74167	0,73333	0,73333
twostage	0,48333	0,48333	0,69167
ward	0,36667	0,45	0,58333

24 lentelė

„Roger and Tanimoto“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,425	0,44167	0,59167
centroid	0,46667	0,50833	0,725
complete	0,425	0,44167	0,575
density	0,73333	0,73333	0,725
flexible	0,35833	0,45	0,525
mcquitty	0,33333	0,45833	0,59167
median	0,725	0,63333	0,70833
single	0,74167	0,73333	0,73333
twostage	0,48333	0,48333	0,71667
ward	0,30833	0,43333	0,58333

25 lentelė

„Sokal and Sneath1“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,43333	0,49167	0,61667
centroid	0,41667	0,475	0,56667
complete	0,425	0,44167	0,575
density	0,73333	0,73333	0,725
flexible	0,35	0,45	0,575
mcquitty	0,45	0,45833	0,61667
median	0,45	0,49167	0,61667
single	0,74167	0,73333	0,73333
twostage	0,48333	0,48333	0,725
ward	0,35	0,425	0,575

26 lentelė

„Djaccard“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,43333	0,49167	0,725
centroid	0,41667	0,475	0,73333
complete	0,425	0,44167	0,69167
density	0,73333	0,73333	0,74167
flexible	0,35	0,45	0,60833
mcquitty	0,45	0,45833	0,68333
median	0,45	0,49167	0,675
single	0,74167	0,73333	0,73333
twostage	0,48333	0,48333	0,74167
ward	0,35	0,425	0,63333

27 lentelė

„Dice“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,43333	0,49167	0,725
centroid	0,41667	0,475	0,71667
complete	0,425	0,44167	0,69167
density	0,73333	0,73333	0,74167
flexible	0,35	0,45	0,65
mcquitty	0,45	0,45833	0,69167
median	0,45	0,49167	0,66667
single	0,74167	0,73333	0,73333
twostage	0,48333	0,48333	0,74167
ward	0,35	0,425	0,63333

28 lentelė

„Russell and Rao“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,74167	0,74167	0,74167
centroid	0,74167	0,74167	0,74167
complete	0,74167	0,74167	0,71667
density	0,73333	0,73333	0,725
flexible	0,325	0,39167	0,525
mcquitty	0,74167	0,74167	0,69167
median	0,74167	0,74167	0,74167
single	0,74167	0,74167	0,74167
twostage	0,73333	0,73333	0,725
ward	0,525	0,51667	0,61667

29 lentelė

„Bray and Curtis“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,74167	0,74167	0,725
centroid	0,74167	0,74167	0,71667
complete	0,74167	0,74167	0,69167
density	0,73333	0,73333	0,74167
flexible	0,325	0,39167	0,65
mcquitty	0,74167	0,74167	0,69167
median	0,74167	0,74167	0,66667
single	0,74167	0,74167	0,73333
twostage	0,73333	0,73333	0,74167
ward	0,525	0,51667	0,63333

30 lentelė

„Kulcynski1“ atstumas

Metodas	Atsiskiriantys klasteriai	Vidutiniškai persidengiantys klasteriai	Smarkiai persidengiantys klasteriai
average	0,74167	0,74167	0,725
centroid	0,74167	0,74167	0,71667
complete	0,74167	0,74167	0,69167
density	0,73333	0,73333	0,74167
flexible	0,325	0,39167	0,65
mcquitty	0,74167	0,74167	0,69167
median	0,74167	0,74167	0,66667
single	0,74167	0,74167	0,73333
twostage	0,73333	0,73333	0,74167
ward	0,525	0,51667	0,63333

Rezultatai, kai buvo tiriamas hierarchinių metodų tikslumas priklausomai nuo duomenų savybių vektoriaus ilgio su vidutiniškai persidengiančiais klasteriais.

2 klasteriai:

31 lentelė

„Hamming“ atstumas

Metodas	5	10	15
average	0,20833	0,14167	0,11667
centroid	0,20833	0,14167	0,075
complete	0,26667	0,175	0,14167
density	0,3	0,49167	0,49167
flexible	0,20833	0,125	0,11667
mcquitty	0,38333	0,275	0,125
median	0,20833	0,48333	0,10833
single	0,49167	0,49167	0,49167
twostage	0,3	0,19167	0,05833
ward	0,21667	0,15833	0,10833

32 lentelė

„Dmatch“ atstumas

Metodas	5	10	15
average	0,20833	0,14167	0,1
centroid	0,175	0,49167	0,49167
complete	0,26667	0,175	0,14167
density	0,3	0,49167	0,49167
flexible	0,29167	0,14167	0,11667
mcquitty	0,38333	0,20833	0,18333
median	0,325	0,5	0,49167
single	0,49167	0,49167	0,49167
twostage	0,3	0,41667	0,08333
ward	0,325	0,15	0,19167

33 lentelė

„Dsqmatch“ atstumas

Metodas	5	10	15
average	0,20833	0,14167	0,11667
centroid	0,20833	0,15	0,075
complete	0,26667	0,175	0,14167
density	0,3	0,49167	0,49167
flexible	0,20833	0,125	0,11667
mcquitty	0,38333	0,30833	0,11667
median	0,20833	0,5	0,10833
single	0,49167	0,49167	0,49167
twostage	0,3	0,2	0,05833
ward	0,21667	0,15833	0,10833

34 lentelė

„Roger and Tanimoto“ atstumas

Metodas	5	10	15
average	0,20833	0,14167	0,125
centroid	0,20833	0,49167	0,49167
complete	0,26667	0,175	0,14167
density	0,3	0,49167	0,49167
flexible	0,25833	0,125	0,10833
mcquitty	0,38333	0,34167	0,18333
median	0,38333	0,49167	0,49167
single	0,49167	0,49167	0,49167
twostage	0,3	0,21667	0,06667
ward	0,26667	0,11667	0,13333

35 lentelė

„Sokal and Sneath1“ atstumas

Metodas	5	10	15
average	0,20833	0,125	0,11667
centroid	0,20833	0,14167	0,075
complete	0,26667	0,175	0,14167
density	0,3	0,49167	0,49167
flexible	0,29167	0,125	0,11667
mcquitty	0,38333	0,18333	0,10833
median	0,20833	0,48333	0,49167
single	0,49167	0,49167	0,49167
twostage	0,3	0,175	0,05833
ward	0,25	0,13333	0,10833

36 lentelė

„Djaccard“ atstumas

Metodas	5	10	15
average	0,20833	0,48333	0,46667
centroid	0,20833	0,48333	0,49167
complete	0,26667	0,46667	0,46667
density	0,3	0,49167	0,49167
flexible	0,29167	0,25	0,125
mcquitty	0,38333	0,46667	0,46667
median	0,20833	0,46667	0,49167
single	0,49167	0,49167	0,49167
twostage	0,3	0,40833	0,49167
ward	0,25	0,15	0,13333

37 lentelė

„Dice“ atstumas

Metodas	5	10	15
average	0,20833	0,48333	0,46667
centroid	0,20833	0,48333	0,49167
complete	0,26667	0,46667	0,46667
density	0,3	0,49167	0,49167
flexible	0,29167	0,33333	0,125
mcquitty	0,38333	0,46667	0,46667
median	0,20833	0,48333	0,49167
single	0,49167	0,49167	0,49167
twostage	0,3	0,425	0,49167
ward	0,25	0,31667	0,275

38 lentelė

„Russell and Rao“ atstumas

Metodas	5	10	15
average	0,49167	0,49167	0,475
centroid	0,49167	0,49167	0,49167
complete	0,49167	0,48333	0,475
density	0,49167	0,49167	0,49167
flexible	0,24167	0,1	0,08333
mcquitty	0,49167	0,475	0,49167
median	0,49167	0,49167	0,49167
single	0,49167	0,49167	0,49167
twostage	0,49167	0,49167	0,49167
ward	0,21667	0,13333	0,15833

39 lentelė

„Bray and Curtis“ atstumas

Metodas	5	10	15
average	0,49167	0,48333	0,46667
centroid	0,49167	0,48333	0,49167
complete	0,49167	0,46667	0,46667
density	0,49167	0,49167	0,49167
flexible	0,24167	0,33333	0,125
mcquitty	0,49167	0,46667	0,46667
median	0,49167	0,48333	0,49167
single	0,49167	0,49167	0,49167
twostage	0,49167	0,425	0,49167
ward	0,21667	0,31667	0,275

40 lentelė

„Kulcynski1“ atstumas

Metodas	5	10	15
average	0,49167	0,48333	0,49167
centroid	0,49167	0,48333	0,49167
complete	0,49167	0,46667	0,45833
density	0,49167	0,49167	0,49167
flexible	0,24167	0,33333	0,49167
mcquitty	0,49167	0,46667	0,49167
median	0,49167	0,48333	0,49167
single	0,49167	0,49167	0,49167
twostage	0,49167	0,425	0,49167
ward	0,21667	0,31667	0,46667

3 klasteriai:

41 lentelė

„Hamming“ atstumas

Metodas	5	10	15
average	0,40833	0,425	0,31667
centroid	0,44167	0,39167	0,375
complete	0,475	0,4	0,45833
density	0,65	0,65	0,65833
flexible	0,43333	0,4	0,30833
mcquitty	0,4	0,41667	0,3
median	0,41667	0,65833	0,63333
single	0,65833	0,65	0,65833
twostage	0,51667	0,65	0,65833
ward	0,475	0,39167	0,3

42 lentelė

„Dmatch“ atstumas

Metodas	5	10	15
average	0,40833	0,41667	0,33333
centroid	0,4	0,65833	0,65833
complete	0,475	0,4	0,45833
density	0,65	0,65	0,65
flexible	0,4	0,40833	0,35833
mcquitty	0,4	0,34167	0,33333
median	0,46667	0,65833	0,65
single	0,65833	0,65	0,65833
twostage	0,51667	0,65	0,65
ward	0,375	0,38333	0,275

43 lentelė

„Dsqmatch“ atstumas

Metodas	5	10	15
average	0,40833	0,425	0,31667
centroid	0,44167	0,39167	0,375
complete	0,475	0,4	0,45833
density	0,65	0,65	0,65833
flexible	0,43333	0,4	0,30833
mcquitty	0,4	0,41667	0,30833
median	0,41667	0,65833	0,63333
single	0,65833	0,65	0,65833
twostage	0,51667	0,65	0,65833
ward	0,475	0,39167	0,3

44 lentelė

„Roger and Tanimoto“ atstumas

Metodas	5	10	15
average	0,40833	0,425	0,31667
centroid	0,44167	0,65	0,65
complete	0,475	0,4	0,45833
density	0,65	0,65	0,65833
flexible	0,4	0,375	0,29167
mcquitty	0,4	0,34167	0,33333
median	0,46667	0,65	0,64167
single	0,65833	0,65	0,65833
twostage	0,51667	0,65	0,65833
ward	0,43333	0,35	0,275

45 lentelė

„Sokal and Sneath1“ atstumas

Metodas	5	10	15
average	0,40833	0,425	0,31667
centroid	0,38333	0,38333	0,325
complete	0,475	0,4	0,45833
density	0,65	0,65	0,65833
flexible	0,43333	0,43333	0,30833
mcquitty	0,4	0,40833	0,34167
median	0,4	0,39167	0,61667
single	0,65833	0,65	0,65833
twostage	0,51667	0,65	0,33333
ward	0,29167	0,36667	0,3

46 lentelė

„Djaccard“ atstumas

Metodas	5	10	15
average	0,40833	0,63333	0,60833
centroid	0,38333	0,64167	0,65
complete	0,475	0,59167	0,63333
density	0,65	0,65	0,65833
flexible	0,43333	0,375	0,40833
mcquitty	0,4	0,63333	0,58333
median	0,4	0,63333	0,65833
single	0,65833	0,65	0,65833
twostage	0,51667	0,65	0,65833
ward	0,29167	0,41667	0,39167

47 lentelė

„Dice“ atstumas

Metodas	5	10	15
average	0,40833	0,64167	0,60833
centroid	0,38333	0,64167	0,65
complete	0,475	0,59167	0,63333
density	0,65	0,65	0,65833
flexible	0,43333	0,38333	0,41667
mcquitty	0,4	0,63333	0,59167
median	0,4	0,63333	0,65
single	0,65833	0,65	0,65833
twostage	0,51667	0,65	0,65833
ward	0,29167	0,50833	0,525

48 lentelė

„Russell and Rao“ atstumas

Metodas	5	10	15
average	0,65833	0,65	0,65833
centroid	0,65833	0,65833	0,65833
complete	0,65833	0,625	0,64167
density	0,65833	0,65833	0,65833
flexible	0,425	0,35833	0,25833
mcquitty	0,65833	0,64167	0,61667
median	0,65833	0,65833	0,65833
single	0,65833	0,65833	0,65833
twostage	0,65833	0,65833	0,65833
ward	0,475	0,40833	0,475

49 lentelė

„Bray and Curtis“ atstumas

Metodas	5	10	15
average	0,65833	0,64167	0,60833
centroid	0,65833	0,64167	0,65
complete	0,65833	0,59167	0,63333
density	0,65833	0,65	0,65833
flexible	0,425	0,38333	0,41667
mcquitty	0,65833	0,63333	0,59167
median	0,65833	0,63333	0,65
single	0,65833	0,65	0,65833
twostage	0,65833	0,65	0,65833
ward	0,475	0,50833	0,525

50 lentelė

„Kulcynski1“ atstumas

Metodas	5	10	15
average	0,65833	0,64167	0,60833
centroid	0,65833	0,64167	0,65
complete	0,65833	0,59167	0,63333
density	0,65833	0,65	0,65833
flexible	0,425	0,38333	0,41667
mcquitty	0,65833	0,63333	0,59167
median	0,65833	0,63333	0,65
single	0,65833	0,65	0,65833
twostage	0,65833	0,65	0,65833
ward	0,475	0,50833	0,525

4 klasteriai:

51 lentelė

„Hamming“ atstumas

Metodas	5	10	15
average	0,44167	0,45	0,40833
centroid	0,44167	0,475	0,50833
complete	0,46667	0,44167	0,41667
density	0,65833	0,73333	0,73333
flexible	0,38333	0,45	0,31667
mcquitty	0,46667	0,43333	0,45
median	0,46667	0,475	0,70833
single	0,6	0,73333	0,74167
twostage	0,56667	0,48333	0,49167
ward	0,425	0,45	0,35

52 lentelė

„Dmatch“ atstumas

Metodas	5	10	15
average	0,44167	0,41667	0,425
centroid	0,54167	0,53333	0,73333
complete	0,46667	0,44167	0,41667
density	0,65833	0,73333	0,73333
flexible	0,45833	0,45833	0,325
mcquitty	0,46667	0,45833	0,43333
median	0,71667	0,725	0,73333
single	0,6	0,73333	0,74167
twostage	0,56667	0,49167	0,51667
ward	0,5	0,49167	0,325

53 lentelė

„Dsqmatch“ atstumas

Metodas	5	10	15
average	0,44167	0,45	0,41667
centroid	0,44167	0,5	0,50833
complete	0,46667	0,44167	0,41667
density	0,65833	0,73333	0,73333
flexible	0,38333	0,45	0,31667
mcquitty	0,46667	0,41667	0,39167
median	0,46667	0,475	0,71667
single	0,6	0,73333	0,74167
twostage	0,56667	0,48333	0,49167
ward	0,425	0,45	0,35

54 lentelė

„Roger and Tanimoto“ atstumas

Metodas	5	10	15
average	0,44167	0,44167	0,38333
centroid	0,44167	0,50833	0,73333
complete	0,46667	0,44167	0,41667
density	0,65833	0,73333	0,73333
flexible	0,5	0,45	0,325
mcquitty	0,46667	0,45833	0,43333
median	0,46667	0,63333	0,71667
single	0,6	0,73333	0,74167
twostage	0,56667	0,48333	0,5
ward	0,40833	0,43333	0,36667

55 lentelė

„Sokal and Sneath1“ atstumas

Metodas	5	10	15
average	0,44167	0,49167	0,40833
centroid	0,44167	0,475	0,48333
complete	0,46667	0,44167	0,41667
density	0,65833	0,73333	0,73333
flexible	0,38333	0,45	0,35833
mcquitty	0,46667	0,45833	0,44167
median	0,46667	0,49167	0,4
single	0,6	0,73333	0,74167
twostage	0,56667	0,48333	0,49167
ward	0,40833	0,425	0,375

56 lentelė

„Djaccard“ atstumas

Metodas	5	10	15
average	0,44167	0,49167	0,40833
centroid	0,44167	0,475	0,48333
complete	0,46667	0,44167	0,41667
density	0,65833	0,73333	0,73333
flexible	0,38333	0,45	0,35833
mcquitty	0,46667	0,45833	0,44167
median	0,46667	0,49167	0,4
single	0,6	0,73333	0,74167
twostage	0,56667	0,48333	0,49167
ward	0,40833	0,425	0,375

57 lentelė

„Dice“ atstumas

Metodas	5	10	15
average	0,44167	0,49167	0,40833
centroid	0,44167	0,475	0,48333
complete	0,46667	0,44167	0,41667
density	0,65833	0,73333	0,73333
flexible	0,38333	0,45	0,35833
mcquitty	0,46667	0,45833	0,44167
median	0,46667	0,49167	0,4
single	0,6	0,73333	0,74167
twostage	0,56667	0,48333	0,49167
ward	0,40833	0,425	0,375

58 lentelė

„Russell and Rao“ atstumas

Metodas	5	10	15
average	0,74167	0,74167	0,74167
centroid	0,74167	0,74167	0,74167
complete	0,74167	0,74167	0,74167
density	0,73333	0,73333	0,73333
flexible	0,45833	0,39167	0,3
mcquitty	0,74167	0,74167	0,74167
median	0,74167	0,74167	0,74167
single	0,74167	0,74167	0,74167
twostage	0,73333	0,73333	0,73333
ward	0,49167	0,51667	0,45833

59 lentelė

„Bray and Curtis“ atstumas

Metodas	5	10	15
average	0,74167	0,74167	0,74167
centroid	0,74167	0,74167	0,74167
complete	0,74167	0,74167	0,74167
density	0,73333	0,73333	0,73333
flexible	0,45833	0,39167	0,3
mcquitty	0,74167	0,74167	0,74167
median	0,74167	0,74167	0,74167
single	0,74167	0,74167	0,74167
twostage	0,73333	0,73333	0,73333
ward	0,49167	0,51667	0,45833

60 lentelė

„Kulcynski1“ atstumas

Metodas	5	10	15
average	0,74167	0,74167	0,74167
centroid	0,74167	0,74167	0,74167
complete	0,74167	0,74167	0,74167
density	0,73333	0,73333	0,73333
flexible	0,45833	0,39167	0,3
mcquitty	0,74167	0,74167	0,74167
median	0,74167	0,74167	0,74167
single	0,74167	0,74167	0,74167
twostage	0,73333	0,73333	0,73333
ward	0,49167	0,51667	0,45833

Rezultatai, kai buvo tiriamas hierarchinių metodų tikslumas nuo parenkamų klasterių skaičiaus, kai duomenų savybių vektoriaus ilgis lygus 15.

Atsiskiriantys klasteriai:

61 lentelė

„Hamming“ atstumas

Metodas	2	3	4
average	0,00833	0,225	0,38333
centroid	0,00833	0,31667	0,45
complete	0,00833	0,23333	0,375
density	0,49167	0,65833	0,73333
flexible	0,00833	0,11667	0,35
mcquitty	0,00833	0,26667	0,33333
median	0,48333	0,60833	0,725
single	0,49167	0,65833	0,74167
twostage	0,00833	0,325	0,49167
ward	0,00833	0,075	0,31667

62 lentelė

„Dmatch“ atstumas

Metodas	2	3	4
average	0,00833	0,25833	0,40833
centroid	0	0,325	0,48333
complete	0,00833	0,23333	0,375
density	0,49167	0,65833	0,73333
flexible	0,00833	0,11667	0,21667
mcquitty	0,025	0,31667	0,43333
median	0,49167	0,65	0,725
single	0,49167	0,65833	0,74167
twostage	0,03333	0,325	0,49167
ward	0,01667	0,15	0,21667

63 lentelė

„Dsqmatch“ atstumas

Metodas	2	3	4
average	0,00833	0,225	0,38333
centroid	0,00833	0,31667	0,45
complete	0,00833	0,23333	0,375
density	0,49167	0,65833	0,73333
flexible	0,00833	0,11667	0,35
mcquitty	0,025	0,275	0,33333
median	0,48333	0,60833	0,725
single	0,49167	0,65833	0,74167
twostage	0,00833	0,325	0,49167
ward	0,00833	0,075	0,31667

64 lentelė

„Roger and Tanimoto“ atstumas

Metodas	2	3	4
average	0,00833	0,23333	0,40833
centroid	0	0,3	0,475
complete	0,00833	0,23333	0,375
density	0,49167	0,65833	0,73333
flexible	0,01667	0,10833	0,19167
mcquitty	0,025	0,38333	0,43333
median	0,49167	0,65833	0,73333
single	0,49167	0,65833	0,74167
twostage	0,00833	0,325	0,49167
ward	0,00833	0,09167	0,275

65 lentelė

„Sokal and Sneath1“ atstumas

Metodas	2	3	4
average	0,00833	0,30833	0,38333
centroid	0,00833	0,30833	0,45
complete	0,00833	0,23333	0,375
density	0,49167	0,65833	0,73333
flexible	0,00833	0,1	0,275
mcquitty	0	0,26667	0,43333
median	0,48333	0,31667	0,45
single	0,49167	0,65833	0,74167
twostage	0,00833	0,325	0,49167
ward	0,00833	0,1	0,35

66 lentelė

„Djaccard“ atstumas

Metodas	2	3	4
average	0,00833	0,30833	0,38333
centroid	0,00833	0,30833	0,45
complete	0,00833	0,23333	0,375
density	0,49167	0,65833	0,73333
flexible	0,00833	0,1	0,275
mcquitty	0	0,26667	0,43333
median	0,48333	0,31667	0,45
single	0,49167	0,65833	0,74167
twostage	0,00833	0,325	0,49167
ward	0,00833	0,1	0,35

67 lentelė

„Dice“ atstumas

Metodas	2	3	4
average	0,00833	0,30833	0,38333
centroid	0,00833	0,30833	0,45
complete	0,00833	0,23333	0,375
density	0,49167	0,65833	0,73333
flexible	0,00833	0,1	0,275
mcquitty	0	0,26667	0,43333
median	0,48333	0,31667	0,45
single	0,49167	0,65833	0,74167
twostage	0,00833	0,325	0,49167
ward	0,00833	0,1	0,35

68 lentelė

„Russell and Rao“ atstumas

Metodas	2	3	4
average	0,49167	0,65833	0,74167
centroid	0,49167	0,65833	0,74167
complete	0,49167	0,65833	0,74167
density	0,49167	0,65833	0,74167
flexible	0	0,09167	0,4
mcquitty	0,49167	0,65833	0,74167
median	0,49167	0,65833	0,74167
single	0,49167	0,65833	0,74167
twostage	0,49167	0,65833	0,74167
ward	0,00833	0,44167	0,4

69 lentelė

„Bray and Curtis“ atstumas

Metodas	2	3	4
average	0,49167	0,65833	0,74167
centroid	0,49167	0,65833	0,74167
complete	0,49167	0,65833	0,74167
density	0,49167	0,65833	0,74167
flexible	0	0,09167	0,4
mcquitty	0,49167	0,65833	0,74167
median	0,49167	0,65833	0,74167
single	0,49167	0,65833	0,74167
twostage	0,49167	0,65833	0,74167
ward	0,00833	0,44167	0,4

70 lentelė

„Kulcynski1“ atstumas

Metodas	2	3	4
average	0,49167	0,65833	0,74167
centroid	0,49167	0,65833	0,74167
complete	0,49167	0,65833	0,74167
density	0,49167	0,65833	0,74167
flexible	0	0,09167	0,4
mcquitty	0,49167	0,65833	0,74167
median	0,49167	0,65833	0,74167
single	0,49167	0,65833	0,74167
twostage	0,49167	0,65833	0,74167
ward	0,00833	0,44167	0,4

Vidutiniškai persidengiantys klasteriai:

71 lentelė

„Hamming“ atstumas

Metodas	2	3	4
average	0,11667	0,31667	0,40833
centroid	0,075	0,375	0,50833
complete	0,14167	0,45833	0,41667
density	0,49167	0,65833	0,73333
flexible	0,11667	0,30833	0,31667
mcquitty	0,125	0,3	0,45
median	0,10833	0,63333	0,70833
single	0,49167	0,65833	0,74167
twostage	0,05833	0,65833	0,49167
ward	0,10833	0,3	0,35

72 lentelė

„Dmatch“ atstumas

Metodas	2	3	4
average	0,1	0,33333	0,425
centroid	0,49167	0,65833	0,73333
complete	0,14167	0,45833	0,41667
density	0,49167	0,65	0,73333
flexible	0,11667	0,35833	0,325
mcquitty	0,18333	0,33333	0,43333
median	0,49167	0,65	0,73333
single	0,49167	0,65833	0,74167
twostage	0,08333	0,65	0,51667
ward	0,19167	0,275	0,325

73 lentelė

„Dsqmatch“ atstumas

Metodas	2	3	4
average	0,11667	0,31667	0,41667
centroid	0,075	0,375	0,50833
complete	0,14167	0,45833	0,41667
density	0,49167	0,65833	0,73333
flexible	0,11667	0,30833	0,31667
mcquitty	0,11667	0,30833	0,39167
median	0,10833	0,63333	0,71667
single	0,49167	0,65833	0,74167
twostage	0,05833	0,65833	0,49167
ward	0,10833	0,3	0,35

74 lentelė

„Roger and Tanimoto“ atstumas

Metodas	2	3	4
average	0,125	0,31667	0,38333
centroid	0,49167	0,65	0,73333
complete	0,14167	0,45833	0,41667
density	0,49167	0,65833	0,73333
flexible	0,10833	0,29167	0,325
mcquitty	0,18333	0,33333	0,43333
median	0,49167	0,64167	0,71667
single	0,49167	0,65833	0,74167
twostage	0,06667	0,65833	0,5
ward	0,13333	0,275	0,36667

75 lentelė

„Sokal and Sneath1“ atstumas

Metodas	2	3	4
average	0,11667	0,31667	0,40833
centroid	0,075	0,325	0,48333
complete	0,14167	0,45833	0,41667
density	0,49167	0,65833	0,73333
flexible	0,11667	0,30833	0,35833
mcquitty	0,10833	0,34167	0,44167
median	0,49167	0,61667	0,4
single	0,49167	0,65833	0,74167
twostage	0,05833	0,33333	0,49167
ward	0,10833	0,3	0,375

76 lentelė

„Djaccard“ atstumas

Metodas	2	3	4
average	0,46667	0,60833	0,40833
centroid	0,49167	0,65	0,48333
complete	0,46667	0,63333	0,41667
density	0,49167	0,65833	0,73333
flexible	0,125	0,40833	0,35833
mcquitty	0,46667	0,58333	0,44167
median	0,49167	0,65833	0,4
single	0,49167	0,65833	0,74167
twostage	0,49167	0,65833	0,49167
ward	0,13333	0,39167	0,375

77 lentelė

„Dice“ atstumas

Metodas	2	3	4
average	0,46667	0,60833	0,40833
centroid	0,49167	0,65	0,48333
complete	0,46667	0,63333	0,41667
density	0,49167	0,65833	0,73333
flexible	0,125	0,41667	0,35833
mcquitty	0,46667	0,59167	0,44167
median	0,49167	0,65	0,4
single	0,49167	0,65833	0,74167
twostage	0,49167	0,65833	0,49167
ward	0,275	0,525	0,375

78 lentelė

„Russell and Rao“ atstumas

Metodas	2	3	4
average	0,475	0,65833	0,74167
centroid	0,49167	0,65833	0,74167
complete	0,475	0,64167	0,74167
density	0,49167	0,65833	0,73333
flexible	0,08333	0,25833	0,3
mcquitty	0,49167	0,61667	0,74167
median	0,49167	0,65833	0,74167
single	0,49167	0,65833	0,74167
twostage	0,49167	0,65833	0,73333
ward	0,15833	0,475	0,45833

79 lentelė

„Bray and Curtis“ atstumas

Metodas	2	3	4
average	0,46667	0,60833	0,74167
centroid	0,49167	0,65	0,74167
complete	0,46667	0,63333	0,74167
density	0,49167	0,65833	0,73333
flexible	0,125	0,41667	0,3
mcquitty	0,46667	0,59167	0,74167
median	0,49167	0,65	0,74167
single	0,49167	0,65833	0,74167
twostage	0,49167	0,65833	0,73333
ward	0,275	0,525	0,45833

80 lentelė

„Kulcynski1“ atstumas

Metodas	2	3	4
average	0,49167	0,60833	0,74167
centroid	0,49167	0,65	0,74167
complete	0,45833	0,63333	0,74167
density	0,49167	0,65833	0,73333
flexible	0,49167	0,41667	0,3
mcquitty	0,49167	0,59167	0,74167
median	0,49167	0,65	0,74167
single	0,49167	0,65833	0,74167
twostage	0,49167	0,65833	0,73333
ward	0,46667	0,525	0,45833

Smarkiai persidengiantys klasteriai:

81 lentelė

„Hamming“ atstumas

Metodas	2	3	4
average	0,30833	0,575	0,49167
centroid	0,48333	0,65833	0,725
complete	0,4	0,59167	0,525
density	0,49167	0,65833	0,74167
flexible	0,35833	0,58333	0,55833
mcquitty	0,35833	0,59167	0,55
median	0,49167	0,65	0,69167
single	0,49167	0,65833	0,73333
twostage	0,38333	0,59167	0,675
ward	0,325	0,55	0,49167

82 lentelė

„Dmatch“ atstumas

Metodas	2	3	4
average	0,30833	0,51667	0,48333
centroid	0,49167	0,65	0,73333
complete	0,4	0,59167	0,525
density	0,49167	0,65	0,74167
flexible	0,46667	0,55	0,53333
mcquitty	0,36667	0,575	0,525
median	0,49167	0,65	0,725
single	0,49167	0,65833	0,73333
twostage	0,4	0,6	0,58333
ward	0,33333	0,58333	0,525

83 lentelė

„Dsqmatch“ atstumas

Metodas	2	3	4
average	0,30833	0,575	0,49167
centroid	0,48333	0,65833	0,725
complete	0,4	0,59167	0,525
density	0,49167	0,65833	0,74167
flexible	0,325	0,58333	0,55833
mcquitty	0,375	0,625	0,53333
median	0,49167	0,65	0,7
single	0,49167	0,65833	0,73333
twostage	0,38333	0,59167	0,675
ward	0,325	0,55	0,49167

84 lentelė

„Roger and Tanimoto“ atstumas

Metodas	2	3	4
average	0,30833	0,575	0,46667
centroid	0,49167	0,65	0,725
complete	0,4	0,59167	0,525
density	0,49167	0,65833	0,74167
flexible	0,46667	0,56667	0,53333
mcquitty	0,36667	0,575	0,525
median	0,49167	0,65	0,725
single	0,49167	0,65833	0,73333
twostage	0,39167	0,59167	0,675
ward	0,28333	0,53333	0,50833

85 lentelė

„Sokal and Sneath1“ atstumas

Metodas	2	3	4
average	0,26667	0,58333	0,5
centroid	0,48333	0,625	0,54167
complete	0,4	0,59167	0,525
density	0,49167	0,65833	0,73333
flexible	0,29167	0,53333	0,53333
mcquitty	0,36667	0,56667	0,55833
median	0,48333	0,64167	0,725
single	0,49167	0,65833	0,73333
twostage	0,39167	0,58333	0,675
ward	0,43333	0,525	0,49167

86 lentelė

„Djaccard“ atstumas

Metodas	2	3	4
average	0,49167	0,65	0,70833
centroid	0,49167	0,65	0,725
complete	0,46667	0,63333	0,69167
density	0,49167	0,65833	0,74167
flexible	0,35	0,6	0,59167
mcquitty	0,375	0,58333	0,625
median	0,49167	0,64167	0,725
single	0,49167	0,65833	0,73333
twostage	0,49167	0,6	0,74167
ward	0,3	0,6	0,59167

87 lentelė

„Dice“ atstumas

Metodas	2	3	4
average	0,49167	0,65	0,70833
centroid	0,49167	0,65833	0,70833
complete	0,46667	0,63333	0,69167
density	0,49167	0,65833	0,74167
flexible	0,44167	0,59167	0,59167
mcquitty	0,43333	0,60833	0,6
median	0,49167	0,625	0,71667
single	0,49167	0,65833	0,73333
twostage	0,49167	0,60833	0,74167
ward	0,36667	0,60833	0,6

88 lentelė

„Russell and Rao“ atstumas

Metodas	2	3	4
average	0,49167	0,65	0,725
centroid	0,49167	0,65	0,73333
complete	0,46667	0,56667	0,71667
density	0,49167	0,65	0,725
flexible	0,275	0,53333	0,50833
mcquitty	0,49167	0,65	0,7
median	0,49167	0,65	0,73333
single	0,49167	0,65833	0,74167
twostage	0,49167	0,65	0,725
ward	0,29167	0,59167	0,55833

89 lentelė

„Bray and Curtis“ atstumas

Metodas	2	3	4
average	0,49167	0,65	0,70833
centroid	0,49167	0,65833	0,70833
complete	0,46667	0,63333	0,69167
density	0,49167	0,65833	0,74167
flexible	0,44167	0,59167	0,59167
mcquitty	0,43333	0,60833	0,6
median	0,49167	0,625	0,71667
single	0,49167	0,65833	0,73333
twostage	0,49167	0,61667	0,74167
ward	0,36667	0,60833	0,6

90 lentelė

„Kulcynski1“ atstumas

Metodas	2	3	4
average	0,48333	0,65	0,74167
centroid	0,48333	0,65	0,74167
complete	0,475	0,625	0,69167
density	0,49167	0,65833	0,74167
flexible	0,49167	0,65833	0,74167
mcquitty	0,49167	0,65	0,74167
median	0,49167	0,65	0,74167
single	0,49167	0,65833	0,74167
twostage	0,49167	0,61667	0,74167
ward	0,45	0,65	0,70833

Rezultatai, kai buvo tiriamas nehierarchinių metodų tikslumas, priklausomai nuo modeliuotų duomenų mišinių parinkimo ir parenkamų klasterių skaičiaus.

Duomenų savybių vektoriaus ilgis lygus 50:

91 lentelė

	2	3	4
Labai gerai atsiskiriantys klasteriai	0	0	0,004
Atsiskiriantys klasteriai	0	0,004	0,027
Vidutiniškai persidengiantys klasteriai	0	0,026	0,128
Smarkiai persidengiantys klasteriai	0,025	0,19	0,178
Labai smarkiai persidengiantys klasteriai	0,077	0,323	0,382

Duomenų savybių vektoriaus ilgis lygus 100:

92 lentelė

	2	3	4
Labai gerai atsiskiriantys klasteriai	0	0	0
Atsiskiriantys klasteriai	0	0,001	0,002
Vidutiniškai persidengiantys klasteriai	0	0,005	0,187
Smarkiai persidengiantys klasteriai	0,005	0,078	0,371
Labai smarkiai persidengiantys klasteriai	0,038	0,314	0,321

Duomenų savybių vektoriaus ilgis lygus 200:

93 lentelė

	2	3	4
Labai gerai atsiskiriantys klasteriai	0	0	0
Atsiskiriantys klasteriai	0	0	0,001
Vidutiniškai persidengiantys klasteriai	0	0,001	0,015
Smarkiai persidengiantys klasteriai	0,002	0,032	0,124
Labai smarkiai persidengiantys klasteriai	0,026	0,217	0,389

Rezultatai, kai buvo tiriamas metodų tikslumas, priklausomai nuo modeliuotų duomenų mišinių parinkimo ir duomenų savybių vektoriaus ilgio.

2 klasteriai:

94 lentelė

	50	100	200
Labai gerai atsiskiriantys klasteriai	0	0	0
Atsiskiriantys klasteriai	0	0	0
Vidutiniškai persidengiantys klasteriai	0	0	0
Smarkiai persidengiantys klasteriai	0,025	0,005	0,002
Labai smarkiai persidengiantys klasteriai	0,077	0,038	0,026

3 klasteriai:

95 lentelė

	50	100	200
Labai gerai atsiskiriantys klasteriai	0	0	0
Atsiskiriantys klasteriai	0,004	0,001	0
Vidutiniškai persidengiantys klasteriai	0,026	0,005	0,001
Smarkiai persidengiantys klasteriai	0,19	0,078	0,032
Labai smarkiai persidengiantys klasteriai	0,323	0,314	0,217

4 klasteriai:

96 lentelė

	50	100	200
Labai gerai atsiskiriantys klasteriai	0,004	0	0
Atsiskiriantys klasteriai	0,027	0,002	0,001
Vidutiniškai persidengiantys klasteriai	0,128	0,187	0,015
Smarkiai persidengiantys klasteriai	0,178	0,371	0,124
Labai smarkiai persidengiantys klasteriai	0,382	0,321	0,389

Rezultatai, kai buvo tiriama nehierarchinių metodų klasterių skaičiaus nustatymo problema.

Modeliuoti 2 klasteriai:

97 lentelė

	CCC	PSF
1	0	-
2	223,58	52,54
3	160,142	29,55
4	134,885	22,01
5	117,26	17,79
6	105,821	15,21

Modeliuoti 3 klasteriai:

98 lentelė

	CCC	PSF
1	0	-
2	51,161	44,88
3	57,648	28,33
4	46,979	19,59
5	49,276	16,63
6	49,82	14,58

Modeliuoti 4 klasteriai:

99 lentelė

	CCC	PSF
1	0	-
2	49,067	83,29
3	57,601	48,9
4	59,421	34,63
5	64,09	28,09
6	62,214	23,09

2 PRIEDAS. KONFERENCIJOS PUBLIKACIJOS MEDŽIAGA

KLASTERIZAVIMO METODŲ TAIKOMŲ BINARINIAMS DUOMENIMS TYRIMAS

Darius Tamašauskas, dr. Tomas Ruzgas
Kauno technologijos universitetas

Darbe tiriami hierarchiniai klasterizavimo metodai taikant įvairius binarinių daugiamačių duomenų mišinius, gautus Monte Carlo modeliavimo pagalba.

SAS paketo pagalba buvo sukurti algoritmai, kurie įvertintų ar duomenys buvo tinkamai suklasterizuoti pagal iš anksto nustatytus klasterius ir kaip rezultatas grąžintos paklaidos, kiek procentų kintamųjų buvo netinkamai suklasterizuoti.

Panaudosime pavyzdį su 3 daugiamačiais duomenų masyvais (150 stebėjimų su 8 skirtingais požymiais) su gerai atskirtais klasteriais, vidutiniškai atskirtais klasteriais bei blogai atskirtais klasteriais. Duomenys buvo sukurti Monte Carlo modeliavimo pagalba su įvairiais binarinių duomenų mišiniais, pagal formulę:

$$f(k,n,p) = \frac{n!}{k!\,(n-k)!}p^k(1-p)^{n-k}$$

čia $k = 0,1,2,\dots,n$.

Skirtingų mišinių duomenų blokams priskirti atskiri mums žinomi klasteriai.

Prieš duomenis klasterizuojant jiems buvo sukurtos atstumų matricos, atstumus apskaičiuojant skirtingais atstumų matais. Kadangi buvo tiriami binariniai duomenys, atstumai buvo skaičiuojami tik nominalioje matavimų skalėje, kurioje binarinis kintamasis gali įgyti 2 reikšmes: 1 arba 0.

Nominalioje matavimų skalėje buvo panaudoti 5 simetriniai atstumų matai: Hamming, Dmatch, Dsqmatch, Roger and Tanimoto bei Sokal and Sneath 1. Taip pat 5 nesimetriniai atstumų matai: Djaccard, Dice, Russell and Rao, Bray and Curtis bei Kulcynski 1.

Su kiekvienu iš šių atstumų matų paskaičiuotos atstumų matricos, kurių duomenys buvo klasterizuojami su hierarchiniais metodais: Average, Centroid, Complete, Density, Fexible, McQuitty, Median, Single, Twostage,Ward.

Atstumas tarp klasterių *K* ir *L* Ward metodu yra apskaičiuojamas taip:

$$D_{KL} = \frac{\|\bar{x}_K - \bar{x}_L\|^2}{(\frac{1}{n_K} + \frac{1}{n_L})}$$

kur $\bar{x}$ yra klasterio C vektoriaus vidurkis, o n yra stebėjimų skaičius.

Gauti rezultatai pateikiami lentelėse.

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	brayc urtis	k1
average	30%	31%	30%	31%	30%	30%	30%	66%	66%	66%
centroid	30%	29%	30%	30%	30%	30%	30%	66%	66%	66%
complete	30%	30%	30%	30%	30%	30%	30%	65%	65%	65%
density	66%	66%	66%	66%	66%	66%	66%	66%	66%	66%
flexible	26%	15%	26%	24%	26%	26%	26%	**26%**	**26%**	**26%**
mcquitty	22%	22%	22%	22%	22%	22%	22%	66%	66%	66%
median	23%	34%	23%	63%	**20%**	**20%**	**20%**	66%	66%	66%
single	66%	66%	66%	66%	66%	66%	66%	66%	66%	66%
twostage	32%	32%	32%	32%	32%	32%	32%	66%	66%	66%
ward	**14%**	**5%**	**14%**	**6%**	25%	25%	25%	29%	29%	29%

1 lentelė. Paklaidos, esant duomenų mišiniui su 3 gerai atskirtais klasteriais

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	brayc urtis	k1
average	34%	34%	34%	34%	34%	34%	34%	66%	66%	66%
centroid	34%	34%	34%	33%	34%	34%	34%	66%	66%	66%
complete	36%	36%	36%	36%	36%	36%	36%	66%	66%	66%
density	66%	66%	66%	66%	66%	66%	66%	66%	66%	66%
flexible	**30%**	24%	**30%**	**22%**	**32%**	**32%**	**32%**	**42%**	**42%**	**42%**
mcquitty	33%	33%	33%	33%	35%	35%	35%	66%	66%	66%
median	38%	34%	38%	64%	43%	43%	43%	66%	66%	66%
single	65%	65%	65%	65%	65%	65%	65%	66%	66%	66%
twostage	48%	48%	47%	48%	48%	48%	48%	66%	66%	66%
ward	33%	**22%**	33%	33%	**32%**	**32%**	**32%**	**42%**	**42%**	**42%**

2 lentelė. Paklaidos, esant duomenų mišiniui su 3 vidutiniškai atskirtais klasteriais

Paklaidos	hamming	dmatch	dsqmatch	rt	ss1	djaccard	dice	rr	brayc urtis	k1
average	50%	50%	50%	50%	53%	53%	53%	66%	66%	66%
centroid	60%	65%	60%	64%	46%	46%	46%	66%	66%	66%
complete	44%	**44%**	44%	44%	**44%**	**44%**	**44%**	66%	66%	66%
density	66%	66%	66%	66%	66%	66%	66%	65%	65%	65%
flexible	42%	48%	42%	48%	49%	49%	49%	51%	51%	51%
mcquitty	48%	**44%**	48%	44%	45%	45%	45%	66%	66%	66%
median	46%	64%	46%	44%	58%	58%	58%	66%	66%	66%
single	66%	66%	66%	66%	66%	66%	66%	66%	66%	66%
twostage	66%	61%	60%	60%	60%	60%	60%	65%	65%	65%
ward	**40%**	51%	**40%**	**40%**	48%	48%	48%	**45%**	**45%**	**45%**

3 lentelė. Paklaidos, esant duomenų mišiniui su 3 blogai atskirtais klasteriais

Matome, kad su gerai atskirtais klasteriais geriausi rezultatai gaunami naudojant Ward metodą, o su vidutiniškai ir blogai atskirtais klasteriais gerus rezultatus gauname ir su Flexible bei Complete metodais.

Literatūra

1. Applied clustering Techniques Course Notes by David Yeo, 2003.
2. SAS Institute Inc. 2008. SAS/STAT® 9.2 User's Guide. Cary, NC: SAS Institute Inc.

3 PRIEDAS. PROGRAMŲ TEKSTAI

Hierarchinio klasterizavimo algoritmas, kai tiriami 2 klasteriai.

```
%macro montecarlo();

       data duomenys(drop=i j);
       array masyvas{&arrayx} j1-j&arrayx;

       do i=1 to round(&count/2);
            Number='No'||left(i);
               c=1;
            do j=1 to &arrayx;
               masyvas{j}=ranbin(i,&nn,&p1);
               end;
          output;
          end;

          do i=round(&count/2)+1 to &count;
             Number='No'||left(i);
               c=2;
             do j=1 to &arrayx;
               masyvas{j}=ranbin(i,&nn,&p2);
               end;
          output;
          end;

       run;

%mend;

%macro dist();

       %if &distmethod=hamming             /* OK (d) Hamming distance */
        or &distmethod=dmatch              /* OK (d) Simple matching coeficient transformed
to Euclidean distance */
        or &distmethod=dsqmatch            /* OK (d) Simple matching coeficient transformed
to squared Euclidean distance */
        or &distmethod=rt                  /* NE (s) Roger and Tanimoto */
        or &distmethod=ss1                 /* NE (s) Sokal and Sneath */
       %then %do;
           proc distance data=duomenys method=&distmethod out=atstumas;
                var nominal(j1--j&arrayx);
                   id Number;
            run;
          %end;
          %else %if &distmethod=djaccard      /* OK (d) Jaccard dissimilarity coeficient */
                 or &distmethod=dice          /* NE (s) Dice coefficient */
                      or &distmethod=rr            /* NE (s) Russel and Rao */
                      or &distmethod=braycurtis    /* OK (d) Bray-Curtis coefficient */
                      or &distmethod=k1            /* OK (d) Kulcynski 1 */
       %then %do;
              proc distance data=duomenys method=&distmethod out=atstumas;
                var anominal(j1--j&arrayx);
                   id Number;
            run;
          %end;

       %if &distmethod=rt                  /* s -> d */
        or &distmethod=ss1                 /* s -> d */
           or &distmethod=dice                /* s -> d */
           or &distmethod=rr                  /* s -> d */
       %then %do;
           data atstumasnew (type=distance);
                    set atstumas;
                    drop z;
                    array matrica{&count} No1-No&count;
```

```
                do z=1 to &count;
                   matrica{z}=1-matrica{z};
                end;
             run;
          %end;
          %else %if &distmethod=hamming       /* d -> d */
         or &distmethod=dmatch              /* d -> d */
         or &distmethod=dsqmatch            /* d -> d */
         or &distmethod=djaccard            /* d -> d */
              or &distmethod=braycurtis          /* d -> d */
              or &distmethod=k1                  /* d -> d */
        %then %do;
              data atstumasnew (type=distance);
                   set atstumas;
                   drop z;
                   array matrica{&count} No1-No&count;
                   do z=1 to &count;
                      matrica{z}=matrica{z};
                   end;
             run;
        %end;

%mend;

%macro hclustering(clmethod=, k=);

          %if &k NE %then %do;
           proc cluster data=atstumasnew method=&clmethod k=&k pseudo
outtree=dendrograma&clmethod noprint;
                id Number;
           run;
          %end;
          %else %do;
           proc cluster data=atstumasnew method=&clmethod pseudo
outtree=dendrograma&clmethod noprint;
                id Number;
           run;
          %end;

        proc tree data=dendrograma&clmethod horizontal out=treeout&clmethod n=2 noprint;
             id Number;
        run;

          proc sort data=treeout&clmethod out=treeout2&clmethod;
               by Number;
          run;

          proc sort data=duomenys out=duomenys2;
               by Number;
          run;

          data klasteriai&clmethod;
               merge duomenys2 treeout2&clmethod;
                    by Number;
                    keep Number c cluster;
          run;

          data transformacijos&clmethod;
               set klasteriai&clmethod;
            transformacija1=cluster;
                       if transformacija1='1' then transformacija1='1';
                       else if transformacija1='2' then transformacija1='2';
                    transformacija2=cluster;
                       if transformacija2='1' then transformacija2='2';
                       else if transformacija2='2' then transformacija2='1';

                  tr1=transformacija1;
                         if transformacija1-c='0' then tr1=0;
                         else tr1=1;
                    tr2=transformacija2;
```

```
                    if transformacija2-c='0' then tr2=0;
                    else tr2=1;
          run;

          proc means data=transformacijos&clmethod mean noprint;
               var tr1-tr2;
                    output out=stats&clmethod;
          run;

          data vidurkiai&clmethod(keep=error _STAT_);
               set stats&clmethod;
                    where _STAT_='MEAN';
                    Error=min(of tr1-tr2);
          run;

          data paklaida&clmethod;
               set vidurkiai&clmethod(rename=(_STAT_=Method));
                    if Method="MEAN" then Method="&clmethod";
       run;

%mend;

options nonumber nodate;

%let p1=0.8;            /* Tikimybe */
%let p2=0.2;            /* Tikimybe */
%let nn=1;               /* n */
%let count=120;          /* Generuoja stebejimu skaiciu */
%let arrayx=10;           /* Generuoja savybiu skaiciu */

%montecarlo;

%let distmethod=hamming;     /* Parenka atitinkama atstumo skaiciavimo metoda */

%dist;

%hclustering(clmethod=average);
%hclustering(clmethod=centroid);
%hclustering(clmethod=complete);
%hclustering(clmethod=density, k=4);
%hclustering(clmethod=flexible);
%hclustering(clmethod=mcquitty);
%hclustering(clmethod=median);
%hclustering(clmethod=single);
%hclustering(clmethod=twostage, k=4);
%hclustering(clmethod=ward);

data paklaidos;
     set paklaidaaverage
           paklaidacentroid
              paklaidacomplete
              paklaidadensity
              paklaidaflexible
              paklaidamcquitty
              paklaidamedian
              paklaidasingle
              paklaidatwostage
              paklaidaward;
run;

proc print data=paklaidos noobs;
     title "Paklaidos, kai naudojamas atstumas: &distmethod";
run;
```

Hierarchinio klasterizavimo algoritmas, kai tiriami 3 klasteriai.

```
%macro montecarlo();
```

```
       data duomenys(drop=i j);
       array masyvas{&arrayx} j1-j&arrayx;

       do i=1 to round(&count/3);
            Number='No'||left(i);
               c=1;
            do j=1 to &arrayx;
               masyvas{j}=ranbin(i,&nn,&p1);
               end;
          output;
          end;

       do i=round(&count/3)+1 to round((&count/3)*2);
            Number='No'||left(i);
               c=2;
            do j=1 to &arrayx;
               masyvas{j}=ranbin(i,&nn,&p2);
               end;
          output;
          end;

          do i=round((&count/3)*2)+1 to &count;
             Number='No'||left(i);
               c=3;
             do j=1 to &arrayx;
               masyvas{j}=ranbin(i,&nn,&p3);
               end;
          output;
          end;

       run;

%mend;

%macro dist();

       %if &distmethod=hamming             /* OK (d) Hamming distance */
        or &distmethod=dmatch              /* OK (d) Simple matching coeficient transformed
to Euclidean distance */
        or &distmethod=dsqmatch            /* OK (d) Simple matching coeficient transformed
to squared Euclidean distance */
        or &distmethod=rt                  /* NE (s) Roger and Tanimoto */
        or &distmethod=ss1                 /* NE (s) Sokal and Sneath */
       %then %do;
           proc distance data=duomenys method=&distmethod out=atstumas;
                var nominal(j1--j&arrayx);
                   id Number;
            run;
          %end;
          %else %if &distmethod=djaccard      /* OK (d) Jaccard dissimilarity coeficient */
                 or &distmethod=dice          /* NE (s) Dice coefficient */
                      or &distmethod=rr            /* NE (s) Russel and Rao */
                      or &distmethod=braycurtis    /* OK (d) Bray-Curtis coefficient */
                      or &distmethod=k1            /* OK (d) Kulcynski 1 */
       %then %do;
              proc distance data=duomenys method=&distmethod out=atstumas;
                var anominal(j1--j&arrayx);
                   id Number;
            run;
          %end;

       %if &distmethod=rt                  /* s -> d */
        or &distmethod=ss1                 /* s -> d */
           or &distmethod=dice                /* s -> d */
           or &distmethod=rr                  /* s -> d */
       %then %do;
           data atstumasnew (type=distance);
                    set atstumas;
                    drop z;
                    array matrica{&count} No1-No&count;
```

```
                do z=1 to &count;
                   matrica{z}=1-matrica{z};
                end;
             run;
          %end;
          %else %if &distmethod=hamming       /* d -> d */
         or &distmethod=dmatch              /* d -> d */
         or &distmethod=dsqmatch            /* d -> d */
         or &distmethod=djaccard            /* d -> d */
              or &distmethod=braycurtis          /* d -> d */
              or &distmethod=k1                  /* d -> d */
        %then %do;
              data atstumasnew (type=distance);
                   set atstumas;
                   drop z;
                   array matrica{&count} No1-No&count;
                   do z=1 to &count;
                      matrica{z}=matrica{z};
                   end;
                run;
        %end;

%mend;

%macro hclustering(clmethod=, k=);

          %if &k NE %then %do;
           proc cluster data=atstumasnew method=&clmethod k=&k pseudo
outtree=dendrograma&clmethod noprint;
                id Number;
           run;
          %end;
          %else %do;
           proc cluster data=atstumasnew method=&clmethod pseudo
outtree=dendrograma&clmethod noprint;
                id Number;
           run;
          %end;

        proc tree data=dendrograma&clmethod horizontal out=treeout&clmethod n=3 noprint;
             id Number;
        run;

          proc sort data=treeout&clmethod out=treeout2&clmethod;
               by Number;
          run;

          proc sort data=duomenys out=duomenys2;
               by Number;
          run;

          data klasteriai&clmethod;
               merge duomenys2 treeout2&clmethod;
                    by Number;
                    keep Number c cluster;
          run;

          data transformacijos&clmethod;
               set klasteriai&clmethod;
            transformacija1=cluster;
                      if transformacija1='1' then transformacija1='1';
                      else if transformacija1='2' then transformacija1='2';
                      else if transformacija1='3' then transformacija1='3';
                    transformacija2=cluster;
                      if transformacija2='1' then transformacija2='1';
                      else if transformacija2='2' then transformacija2='3';
                      else if transformacija2='3' then transformacija2='2';
            transformacija3=cluster;
                      if transformacija3='1' then transformacija3='2';
                      else if transformacija3='2' then transformacija3='1';
```

```
                    else if transformacija3='3' then transformacija3='3';
                 transformacija4=cluster;
                    if transformacija4='1' then transformacija4='2';
                    else if transformacija4='2' then transformacija4='3';
                    else if transformacija4='3' then transformacija4='1';
                 transformacija5=cluster;
                    if transformacija5='1' then transformacija5='3';
                    else if transformacija5='2' then transformacija5='1';
                    else if transformacija5='3' then transformacija5='2';
                 transformacija6=cluster;
                    if transformacija6='1' then transformacija6='3';
                    else if transformacija6='2' then transformacija6='2';
                    else if transformacija6='3' then transformacija6='1';

               tr1=transformacija1;
                     if transformacija1-c='0' then tr1=0;
                     else tr1=1;
                 tr2=transformacija2;
                     if transformacija2-c='0' then tr2=0;
                     else tr2=1;
                 tr3=transformacija3;
                     if transformacija3-c='0' then tr3=0;
                     else tr3=1;
                 tr4=transformacija4;
                     if transformacija4-c='0' then tr4=0;
                     else tr4=1;
                 tr5=transformacija5;
                     if transformacija5-c='0' then tr5=0;
                     else tr5=1;
                 tr6=transformacija6;
                     if transformacija6-c='0' then tr6=0;
                     else tr6=1;
         run;

         proc means data=transformacijos&clmethod mean noprint;
              var tr1-tr6;
                   output out=stats&clmethod;
         run;

         data vidurkiai&clmethod(keep=error _STAT_);
              set stats&clmethod;
                   where _STAT_='MEAN';
                   Error=min(of tr1-tr6);
         run;

         data paklaida&clmethod;
              set vidurkiai&clmethod(rename=(_STAT_=Method));
                   if Method="MEAN" then Method="&clmethod";
       run;

%mend;

options nonumber nodate;

%let p1=0.9;            /* Tikimybe */
%let p2=0.5;            /* Tikimybe */
%let p3=0.1;            /* Tikimybe */
%let nn=1;               /* n */
%let count=120;          /* Generuoja stebejimu skaiciu */
%let arrayx=10;           /* Generuoja savybiu skaiciu */

%montecarlo;

%let distmethod=hamming;     /* Parenka atitinkama atstumo skaiciavimo metoda */

%dist;

%hclustering(clmethod=average);
%hclustering(clmethod=centroid);
%hclustering(clmethod=complete);
```

```
%hclustering(clmethod=density, k=4);
%hclustering(clmethod=flexible);
%hclustering(clmethod=mcquitty);
%hclustering(clmethod=median);
%hclustering(clmethod=single);
%hclustering(clmethod=twostage, k=4);
%hclustering(clmethod=ward);

data paklaidos;
     set paklaidaaverage
           paklaidacentroid
              paklaidacomplete
              paklaidadensity
              paklaidaflexible
              paklaidamcquitty
              paklaidamedian
              paklaidasingle
              paklaidatwostage
              paklaidaward;
run;

proc print data=paklaidos noobs;
     title "Paklaidos, kai naudojamas atstumas: &distmethod";
run;
```

Hierarchinio klasterizavimo algoritmas, kai tiriami 4 klasteriai.

```
%macro montecarlo();

       data duomenys(drop=i j);
       array masyvas{&arrayx} j1-j&arrayx;

       do i=1 to round(&count/4);
             Number='No'||left(i);
                c=1;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p1);
                end;
          output;
          end;

       do i=round(&count/4)+1 to round((&count/4)*2);
             Number='No'||left(i);
                c=2;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p2);
                end;
          output;
          end;

       do i=round((&count/4)*2)+1 to round((&count/4)*3);
             Number='No'||left(i);
                c=3;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p3);
                end;
          output;
          end;

          do i=round((&count/4)*3)+1 to &count;
             Number='No'||left(i);
                c=4;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p4);
                end;
          output;
          end;
```

```
       run;

%mend;

%macro dist();

       %if &distmethod=hamming             /* OK (d) Hamming distance */
        or &distmethod=dmatch              /* OK (d) Simple matching coeficient transformed
to Euclidean distance */
        or &distmethod=dsqmatch            /* OK (d) Simple matching coeficient transformed
to squared Euclidean distance */
        or &distmethod=rt                  /* NE (s) Roger and Tanimoto */
        or &distmethod=ss1                 /* NE (s) Sokal and Sneath */
       %then %do;
           proc distance data=duomenys method=&distmethod out=atstumas;
                var nominal(j1--j&arrayx);
                   id Number;
           run;
          %end;
          %else %if &distmethod=djaccard      /* OK (d) Jaccard dissimilarity coeficient */
                 or &distmethod=dice          /* NE (s) Dice coefficient */
                      or &distmethod=rr            /* NE (s) Russel and Rao */
                      or &distmethod=braycurtis    /* OK (d) Bray-Curtis coefficient */
                      or &distmethod=k1            /* OK (d) Kulcynski 1 */
       %then %do;
              proc distance data=duomenys method=&distmethod out=atstumas;
                var anominal(j1--j&arrayx);
                   id Number;
           run;
          %end;

       %if &distmethod=rt                  /* s -> d */
        or &distmethod=ss1                 /* s -> d */
           or &distmethod=dice                /* s -> d */
           or &distmethod=rr                  /* s -> d */
       %then %do;
           data atstumasnew (type=distance);
                   set atstumas;
                   drop z;
                   array matrica{&count} No1-No&count;
                   do z=1 to &count;
                      matrica{z}=1-matrica{z};
                   end;
                run;
          %end;
          %else %if &distmethod=hamming       /* d -> d */
         or &distmethod=dmatch               /* d -> d */
         or &distmethod=dsqmatch             /* d -> d */
         or &distmethod=djaccard             /* d -> d */
               or &distmethod=braycurtis          /* d -> d */
               or &distmethod=k1                  /* d -> d */
       %then %do;
              data atstumasnew (type=distance);
                    set atstumas;
                    drop z;
                    array matrica{&count} No1-No&count;
                    do z=1 to &count;
                       matrica{z}=matrica{z};
                    end;
                 run;
       %end;

%mend;

%macro hclustering(clmethod=, k=);

          %if &k NE %then %do;
           proc cluster data=atstumasnew method=&clmethod k=&k pseudo
outtree=dendrograma&clmethod noprint;
                id Number;
```

```
          run;
         %end;
         %else %do;
          proc cluster data=atstumasnew method=&clmethod pseudo
outtree=dendrograma&clmethod noprint;
               id Number;
          run;
         %end;

      proc tree data=dendrograma&clmethod horizontal out=treeout&clmethod n=4 noprint;
           id Number;
      run;

         proc sort data=treeout&clmethod out=treeout2&clmethod;
              by Number;
         run;

         proc sort data=duomenys out=duomenys2;
              by Number;
         run;

         data klasteriai&clmethod;
              merge duomenys2 treeout2&clmethod;
                   by Number;
                   keep Number c cluster;
         run;

         data transformacijos&clmethod;
              set klasteriai&clmethod;
           transformacija1=cluster;
                      if transformacija1='1' then transformacija1='1';
                      else if transformacija1='2' then transformacija1='2';
                      else if transformacija1='3' then transformacija1='3';
                      else if transformacija1='4' then transformacija1='4';
                transformacija2=cluster;
                      if transformacija2='1' then transformacija2='1';
                      else if transformacija2='2' then transformacija2='2';
                      else if transformacija2='3' then transformacija2='4';
                      else if transformacija2='4' then transformacija2='3';
                transformacija3=cluster;
                      if transformacija3='1' then transformacija3='1';
                      else if transformacija3='2' then transformacija3='3';
                      else if transformacija3='3' then transformacija3='2';
                      else if transformacija3='4' then transformacija3='4';
                transformacija4=cluster;
                      if transformacija4='1' then transformacija4='1';
                      else if transformacija4='2' then transformacija4='3';
                      else if transformacija4='3' then transformacija4='4';
                      else if transformacija4='4' then transformacija4='2';
                   transformacija5=cluster;
                      if transformacija5='1' then transformacija5='1';
                      else if transformacija5='2' then transformacija5='4';
                      else if transformacija5='3' then transformacija5='2';
                      else if transformacija5='4' then transformacija5='3';
                   transformacija6=cluster;
                      if transformacija6='1' then transformacija6='1';
                      else if transformacija6='2' then transformacija6='4';
                      else if transformacija6='3' then transformacija6='3';
                      else if transformacija6='4' then transformacija6='2';
                   transformacija7=cluster;
                      if transformacija7='1' then transformacija7='2';
                      else if transformacija7='2' then transformacija7='1';
                      else if transformacija7='3' then transformacija7='3';
                      else if transformacija7='4' then transformacija7='4';
                transformacija8=cluster;
                      if transformacija8='1' then transformacija8='2';
                      else if transformacija8='2' then transformacija8='1';
                      else if transformacija8='3' then transformacija8='4';
                      else if transformacija8='4' then transformacija8='3';
                transformacija9=cluster;
```

```
     if transformacija9='1' then transformacija9='2';
     else if transformacija9='2' then transformacija9='3';
     else if transformacija9='3' then transformacija9='1';
     else if transformacija9='4' then transformacija9='4';
transformacija10=cluster;
     if transformacija10='1' then transformacija10='2';
     else if transformacija10='2' then transformacija10='3';
     else if transformacija10='3' then transformacija10='4';
     else if transformacija10='4' then transformacija10='1';
  transformacija11=cluster;
     if transformacija11='1' then transformacija11='2';
     else if transformacija11='2' then transformacija11='4';
     else if transformacija11='3' then transformacija11='1';
     else if transformacija11='4' then transformacija11='3';
  transformacija12=cluster;
     if transformacija12='1' then transformacija12='2';
     else if transformacija12='2' then transformacija12='4';
     else if transformacija12='3' then transformacija12='3';
     else if transformacija12='4' then transformacija12='1';
  transformacija13=cluster;
     if transformacija13='1' then transformacija13='3';
     else if transformacija13='2' then transformacija13='1';
     else if transformacija13='3' then transformacija13='2';
     else if transformacija13='4' then transformacija13='4';
transformacija14=cluster;
     if transformacija14='1' then transformacija14='3';
     else if transformacija14='2' then transformacija14='1';
     else if transformacija14='3' then transformacija14='4';
     else if transformacija14='4' then transformacija14='2';
transformacija15=cluster;
     if transformacija15='1' then transformacija15='3';
     else if transformacija15='2' then transformacija15='2';
     else if transformacija15='3' then transformacija15='1';
     else if transformacija15='4' then transformacija15='4';
transformacija16=cluster;
     if transformacija16='1' then transformacija16='3';
     else if transformacija16='2' then transformacija16='2';
     else if transformacija16='3' then transformacija16='4';
     else if transformacija16='4' then transformacija16='1';
  transformacija17=cluster;
     if transformacija17='1' then transformacija17='3';
     else if transformacija17='2' then transformacija17='4';
     else if transformacija17='3' then transformacija17='1';
     else if transformacija17='4' then transformacija17='2';
  transformacija18=cluster;
     if transformacija18='1' then transformacija18='3';
     else if transformacija18='2' then transformacija18='4';
     else if transformacija18='3' then transformacija18='2';
     else if transformacija18='4' then transformacija18='1';
  transformacija19=cluster;
     if transformacija19='1' then transformacija19='4';
     else if transformacija19='2' then transformacija19='1';
     else if transformacija19='3' then transformacija19='2';
     else if transformacija19='4' then transformacija19='3';
transformacija20=cluster;
     if transformacija20='1' then transformacija20='4';
     else if transformacija20='2' then transformacija20='1';
     else if transformacija20='3' then transformacija20='3';
     else if transformacija20='4' then transformacija20='2';
transformacija21=cluster;
     if transformacija21='1' then transformacija21='4';
     else if transformacija21='2' then transformacija21='2';
     else if transformacija21='3' then transformacija21='1';
     else if transformacija21='4' then transformacija21='3';
transformacija22=cluster;
     if transformacija22='1' then transformacija22='4';
     else if transformacija22='2' then transformacija22='2';
     else if transformacija22='3' then transformacija22='3';
     else if transformacija22='4' then transformacija22='1';
  transformacija23=cluster;
```

```
      if transformacija23='1' then transformacija23='4';
      else if transformacija23='2' then transformacija23='3';
      else if transformacija23='3' then transformacija23='1';
      else if transformacija23='4' then transformacija23='2';
   transformacija24=cluster;
      if transformacija24='1' then transformacija24='4';
      else if transformacija24='2' then transformacija24='3';
      else if transformacija24='3' then transformacija24='2';
      else if transformacija24='4' then transformacija24='1';

tr1=transformacija1;
       if transformacija1-c='0' then tr1=0;
       else tr1=1;
   tr2=transformacija2;
       if transformacija2-c='0' then tr2=0;
       else tr2=1;
   tr3=transformacija3;
       if transformacija3-c='0' then tr3=0;
       else tr3=1;
   tr4=transformacija4;
       if transformacija4-c='0' then tr4=0;
       else tr4=1;
   tr5=transformacija5;
       if transformacija5-c='0' then tr5=0;
       else tr5=1;
   tr6=transformacija6;
       if transformacija6-c='0' then tr6=0;
       else tr6=1;
   tr7=transformacija7;
       if transformacija7-c='0' then tr7=0;
       else tr7=1;
   tr8=transformacija8;
       if transformacija8-c='0' then tr8=0;
       else tr8=1;
   tr9=transformacija9;
       if transformacija9-c='0' then tr9=0;
       else tr9=1;
   tr10=transformacija10;
       if transformacija10-c='0' then tr10=0;
       else tr10=1;
   tr11=transformacija11;
       if transformacija11-c='0' then tr11=0;
       else tr11=1;
   tr12=transformacija12;
       if transformacija12-c='0' then tr12=0;
       else tr12=1;
tr13=transformacija13;
       if transformacija13-c='0' then tr13=0;
       else tr13=1;
   tr14=transformacija14;
       if transformacija14-c='0' then tr14=0;
       else tr14=1;
   tr15=transformacija15;
       if transformacija15-c='0' then tr15=0;
       else tr15=1;
   tr16=transformacija16;
       if transformacija16-c='0' then tr16=0;
       else tr16=1;
   tr17=transformacija17;
       if transformacija17-c='0' then tr17=0;
       else tr17=1;
   tr18=transformacija18;
       if transformacija18-c='0' then tr18=0;
       else tr18=1;
   tr19=transformacija19;
       if transformacija19-c='0' then tr19=0;
       else tr19=1;
   tr20=transformacija20;
       if transformacija20-c='0' then tr20=0;
       else tr20=1;
```

```
                    tr21=transformacija21;
                        if transformacija21-c='0' then tr21=0;
                        else tr21=1;
                    tr22=transformacija22;
                        if transformacija22-c='0' then tr22=0;
                        else tr22=1;
                    tr23=transformacija23;
                        if transformacija23-c='0' then tr23=0;
                        else tr23=1;
                    tr24=transformacija24;
                        if transformacija24-c='0' then tr24=0;
                        else tr24=1;
          run;

          proc means data=transformacijos&clmethod mean noprint;
               var tr1-tr24;
                    output out=stats&clmethod;
          run;

          data vidurkiai&clmethod(keep=error _STAT_);
               set stats&clmethod;
                    where _STAT_='MEAN';
                    Error=min(of tr1-tr24);
          run;

          data paklaida&clmethod;
               set vidurkiai&clmethod(rename=(_STAT_=Method));
                    if Method="MEAN" then Method="&clmethod";
       run;

%mend;

options nonumber nodate;

%let p1=0.95;            /* Tikimybe */
%let p2=0.65;            /* Tikimybe */
%let p3=0.35;            /* Tikimybe */
%let p4=0.05;            /* Tikimybe */
%let nn=1;               /* n */
%let count=120;          /* Generuoja stebejimu skaiciu */
%let arrayx=10;           /* Generuoja savybiu skaiciu */

%montecarlo;

%let distmethod=hamming;     /* Parenka atitinkama atstumo skaiciavimo metoda */

%dist;

%hclustering(clmethod=average);
%hclustering(clmethod=centroid);
%hclustering(clmethod=complete);
%hclustering(clmethod=density, k=4);
%hclustering(clmethod=flexible);
%hclustering(clmethod=mcquitty);
%hclustering(clmethod=median);
%hclustering(clmethod=single);
%hclustering(clmethod=twostage, k=4);
%hclustering(clmethod=ward);

data paklaidos;
     set paklaidaaverage
           paklaidacentroid
              paklaidacomplete
              paklaidadensity
              paklaidaflexible
              paklaidamcquitty
              paklaidamedian
              paklaidasingle
              paklaidatwostage
              paklaidaward;
```

```
run;

proc print data=paklaidos noobs;
     title "Paklaidos, kai naudojamas atstumas: &distmethod";
run;
```

Nehierarchinio klasterizavimo algoritmas, kai tiriami 2 klasteriai.

```
%macro montecarlo();

       data duomenys(drop=i j);
       array masyvas{&arrayx} j1-j&arrayx;

       do i=1 to round(&count/2);
            Number='No'||left(i);
               c=1;
            do j=1 to &arrayx;
               masyvas{j}=ranbin(i,&nn,&p1);
               end;
          output;
          end;

          do i=round(&count/2)+1 to &count;
             Number='No'||left(i);
               c=2;
             do j=1 to &arrayx;
               masyvas{j}=ranbin(i,&nn,&p2);
               end;
          output;
          end;

       run;

%mend;

%macro fclustering(clnr=, itnr=);

       title "Klasteriu skaicius = &clnr";
       proc fastclus data=duomenys maxclus=&clnr maxiter=&itnr out=fast noprint;
              id Number;
          run;

          data fast1;
               set fast;
                    keep Number c cluster;
          run;

          data transformacijos;
               set fast1;
                    transformacija1=cluster;
                       if transformacija1='1' then transformacija1='1';
                       else if transformacija1='2' then transformacija1='2';
                    transformacija2=cluster;
                       if transformacija2='1' then transformacija2='2';
                       else if transformacija2='2' then transformacija2='1';

                  tr1=transformacija1;
                        if transformacija1-c='0' then tr1=0;
                        else tr1=1;
                    tr2=transformacija2;
                        if transformacija2-c='0' then tr2=0;
                        else tr2=1;
          run;

          proc means data=transformacijos mean noprint;
               var tr1-tr2;
                    output out=stats;
          run;
```

```
          data vidurkiai(keep=error _STAT_);
               set stats;
                    where _STAT_='MEAN';
                    Error=min(of tr1-tr2);
          run;

          data paklaida;
               set vidurkiai(rename=(_STAT_=Method));
                    if Method="MEAN" then Method="k-means";
       run;

%mend;

options nonumber nodate;

%let p1=0.9;            /* Tikimybe */
%let p2=0.1;            /* Tikimybe */
%let nn=1;               /* n */
%let count=1000;          /* Generuoja stebejimu skaiciu */
%let arrayx=50;           /* Generuoja savybiu skaiciu */

%montecarlo;

%fclustering(clnr=2,itnr=10);

proc print data=paklaida noobs;
run;
```

Nehierarchinio klasterizavimo algoritmas, kai tiriami 3 klasteriai.

```
%macro montecarlo();

       data duomenys(drop=i j);
       array masyvas{&arrayx} j1-j&arrayx;

       do i=1 to round(&count/3);
             Number='No'||left(i);
                c=1;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p1);
                end;
          output;
          end;

       do i=round(&count/3)+1 to round((&count/3)*2);
             Number='No'||left(i);
                c=2;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p2);
                end;
          output;
          end;

          do i=round((&count/3)*2)+1 to &count;
             Number='No'||left(i);
                c=3;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p3);
                end;
          output;
          end;

       run;

%mend;

%macro fclustering(clnr=, itnr=);
```

```
title "Klasteriu skaicius = &clnr";
proc fastclus data=duomenys maxclus=&clnr maxiter=&itnr out=fast noprint;
       id Number;
   run;

   data fast1;
        set fast;
             keep Number c cluster;
   run;

   data transformacijos;
        set fast1;
             transformacija1=cluster;
                if transformacija1='1' then transformacija1='1';
                else if transformacija1='2' then transformacija1='2';
                else if transformacija1='3' then transformacija1='3';
             transformacija2=cluster;
                if transformacija2='1' then transformacija2='1';
                else if transformacija2='2' then transformacija2='3';
                else if transformacija2='3' then transformacija2='2';
     transformacija3=cluster;
                if transformacija3='1' then transformacija3='2';
                else if transformacija3='2' then transformacija3='1';
                else if transformacija3='3' then transformacija3='3';
             transformacija4=cluster;
                if transformacija4='1' then transformacija4='2';
                else if transformacija4='2' then transformacija4='3';
                else if transformacija4='3' then transformacija4='1';
             transformacija5=cluster;
                if transformacija5='1' then transformacija5='3';
                else if transformacija5='2' then transformacija5='1';
                else if transformacija5='3' then transformacija5='2';
             transformacija6=cluster;
                if transformacija6='1' then transformacija6='3';
                else if transformacija6='2' then transformacija6='2';
                else if transformacija6='3' then transformacija6='1';

           tr1=transformacija1;
                  if transformacija1-c='0' then tr1=0;
                  else tr1=1;
             tr2=transformacija2;
                  if transformacija2-c='0' then tr2=0;
                  else tr2=1;
             tr3=transformacija3;
                  if transformacija3-c='0' then tr3=0;
                  else tr3=1;
             tr4=transformacija4;
                  if transformacija4-c='0' then tr4=0;
                  else tr4=1;
             tr5=transformacija5;
                  if transformacija5-c='0' then tr5=0;
                  else tr5=1;
             tr6=transformacija6;
                  if transformacija6-c='0' then tr6=0;
                  else tr6=1;
   run;

   proc means data=transformacijos mean noprint;
        var tr1-tr6;
             output out=stats;
   run;

   data vidurkiai(keep=error _STAT_);
        set stats;
             where _STAT_='MEAN';
             Error=min(of tr1-tr6);
   run;

   data paklaida;
```

```
              set vidurkiai(rename=(_STAT_=Method));
                   if Method="MEAN" then Method="k-means";
       run;

%mend;

options nonumber nodate;

%let p1=0.9;            /* Tikimybe */
%let p2=0.5;            /* Tikimybe */
%let p3=0.1;            /* Tikimybe */
%let nn=1;               /* n */
%let count=1000;          /* Generuoja stebejimu skaiciu */
%let arrayx=50;           /* Generuoja savybiu skaiciu */

%montecarlo;

%fclustering(clnr=3,itnr=10);

proc print data=paklaida noobs;
run;
```

Nehierarchinio klasterizavimo algoritmas, kai tiriami 4 klasteriai.

```
%macro montecarlo();

       data duomenys(drop=i j);
       array masyvas{&arrayx} j1-j&arrayx;

       do i=1 to round(&count/4);
             Number='No'||left(i);
                c=1;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p1);
                end;
          output;
          end;

       do i=round(&count/4)+1 to round((&count/4)*2);
             Number='No'||left(i);
                c=2;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p2);
                end;
          output;
          end;

       do i=round((&count/4)*2)+1 to round((&count/4)*3);
             Number='No'||left(i);
                c=3;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p3);
                end;
          output;
          end;

          do i=round((&count/4)*3)+1 to &count;
             Number='No'||left(i);
                c=4;
             do j=1 to &arrayx;
                masyvas{j}=ranbin(i,&nn,&p4);
                end;
          output;
          end;

       run;

%mend;
```

```
%macro fclustering(clnr=, itnr=);

       title "Klasteriu skaicius = &clnr";
       proc fastclus data=duomenys maxclus=&clnr maxiter=&itnr out=fast noprint;
              id Number;
         run;

         data fast1;
              set fast;
                   keep Number c cluster;
         run;

         data transformacijos;
              set fast1;
           transformacija1=cluster;
                      if transformacija1='1' then transformacija1='1';
                      else if transformacija1='2' then transformacija1='2';
                      else if transformacija1='3' then transformacija1='3';
                      else if transformacija1='4' then transformacija1='4';
                 transformacija2=cluster;
                      if transformacija2='1' then transformacija2='1';
                      else if transformacija2='2' then transformacija2='2';
                      else if transformacija2='3' then transformacija2='4';
                      else if transformacija2='4' then transformacija2='3';
                 transformacija3=cluster;
                      if transformacija3='1' then transformacija3='1';
                      else if transformacija3='2' then transformacija3='3';
                      else if transformacija3='3' then transformacija3='2';
                      else if transformacija3='4' then transformacija3='4';
                 transformacija4=cluster;
                      if transformacija4='1' then transformacija4='1';
                      else if transformacija4='2' then transformacija4='3';
                      else if transformacija4='3' then transformacija4='4';
                      else if transformacija4='4' then transformacija4='2';
                   transformacija5=cluster;
                      if transformacija5='1' then transformacija5='1';
                      else if transformacija5='2' then transformacija5='4';
                      else if transformacija5='3' then transformacija5='2';
                      else if transformacija5='4' then transformacija5='3';
                   transformacija6=cluster;
                      if transformacija6='1' then transformacija6='1';
                      else if transformacija6='2' then transformacija6='4';
                      else if transformacija6='3' then transformacija6='3';
                      else if transformacija6='4' then transformacija6='2';
                   transformacija7=cluster;
                      if transformacija7='1' then transformacija7='2';
                      else if transformacija7='2' then transformacija7='1';
                      else if transformacija7='3' then transformacija7='3';
                      else if transformacija7='4' then transformacija7='4';
                 transformacija8=cluster;
                      if transformacija8='1' then transformacija8='2';
                      else if transformacija8='2' then transformacija8='1';
                      else if transformacija8='3' then transformacija8='4';
                      else if transformacija8='4' then transformacija8='3';
                 transformacija9=cluster;
                      if transformacija9='1' then transformacija9='2';
                      else if transformacija9='2' then transformacija9='3';
                      else if transformacija9='3' then transformacija9='1';
                      else if transformacija9='4' then transformacija9='4';
                 transformacija10=cluster;
                      if transformacija10='1' then transformacija10='2';
                      else if transformacija10='2' then transformacija10='3';
                      else if transformacija10='3' then transformacija10='4';
                      else if transformacija10='4' then transformacija10='1';
                   transformacija11=cluster;
                      if transformacija11='1' then transformacija11='2';
                      else if transformacija11='2' then transformacija11='4';
                      else if transformacija11='3' then transformacija11='1';
                      else if transformacija11='4' then transformacija11='3';
```

```
   transformacija12=cluster;
      if transformacija12='1' then transformacija12='2';
      else if transformacija12='2' then transformacija12='4';
      else if transformacija12='3' then transformacija12='3';
      else if transformacija12='4' then transformacija12='1';
   transformacija13=cluster;
      if transformacija13='1' then transformacija13='3';
      else if transformacija13='2' then transformacija13='1';
      else if transformacija13='3' then transformacija13='2';
      else if transformacija13='4' then transformacija13='4';
transformacija14=cluster;
      if transformacija14='1' then transformacija14='3';
      else if transformacija14='2' then transformacija14='1';
      else if transformacija14='3' then transformacija14='4';
      else if transformacija14='4' then transformacija14='2';
transformacija15=cluster;
      if transformacija15='1' then transformacija15='3';
      else if transformacija15='2' then transformacija15='2';
      else if transformacija15='3' then transformacija15='1';
      else if transformacija15='4' then transformacija15='4';
transformacija16=cluster;
      if transformacija16='1' then transformacija16='3';
      else if transformacija16='2' then transformacija16='2';
      else if transformacija16='3' then transformacija16='4';
      else if transformacija16='4' then transformacija16='1';
   transformacija17=cluster;
      if transformacija17='1' then transformacija17='3';
      else if transformacija17='2' then transformacija17='4';
      else if transformacija17='3' then transformacija17='1';
      else if transformacija17='4' then transformacija17='2';
   transformacija18=cluster;
      if transformacija18='1' then transformacija18='3';
      else if transformacija18='2' then transformacija18='4';
      else if transformacija18='3' then transformacija18='2';
      else if transformacija18='4' then transformacija18='1';
   transformacija19=cluster;
      if transformacija19='1' then transformacija19='4';
      else if transformacija19='2' then transformacija19='1';
      else if transformacija19='3' then transformacija19='2';
      else if transformacija19='4' then transformacija19='3';
transformacija20=cluster;
      if transformacija20='1' then transformacija20='4';
      else if transformacija20='2' then transformacija20='1';
      else if transformacija20='3' then transformacija20='3';
      else if transformacija20='4' then transformacija20='2';
transformacija21=cluster;
      if transformacija21='1' then transformacija21='4';
      else if transformacija21='2' then transformacija21='2';
      else if transformacija21='3' then transformacija21='1';
      else if transformacija21='4' then transformacija21='3';
transformacija22=cluster;
      if transformacija22='1' then transformacija22='4';
      else if transformacija22='2' then transformacija22='2';
      else if transformacija22='3' then transformacija22='3';
      else if transformacija22='4' then transformacija22='1';
   transformacija23=cluster;
      if transformacija23='1' then transformacija23='4';
      else if transformacija23='2' then transformacija23='3';
      else if transformacija23='3' then transformacija23='1';
      else if transformacija23='4' then transformacija23='2';
   transformacija24=cluster;
      if transformacija24='1' then transformacija24='4';
      else if transformacija24='2' then transformacija24='3';
      else if transformacija24='3' then transformacija24='2';
      else if transformacija24='4' then transformacija24='1';

tr1=transformacija1;
       if transformacija1-c='0' then tr1=0;
       else tr1=1;
   tr2=transformacija2;
```

```
            if transformacija2-c='0' then tr2=0;
            else tr2=1;
        tr3=transformacija3;
            if transformacija3-c='0' then tr3=0;
            else tr3=1;
        tr4=transformacija4;
            if transformacija4-c='0' then tr4=0;
            else tr4=1;
        tr5=transformacija5;
            if transformacija5-c='0' then tr5=0;
            else tr5=1;
        tr6=transformacija6;
            if transformacija6-c='0' then tr6=0;
            else tr6=1;
        tr7=transformacija7;
            if transformacija7-c='0' then tr7=0;
            else tr7=1;
        tr8=transformacija8;
            if transformacija8-c='0' then tr8=0;
            else tr8=1;
        tr9=transformacija9;
            if transformacija9-c='0' then tr9=0;
            else tr9=1;
        tr10=transformacija10;
            if transformacija10-c='0' then tr10=0;
            else tr10=1;
        tr11=transformacija11;
            if transformacija11-c='0' then tr11=0;
            else tr11=1;
        tr12=transformacija12;
            if transformacija12-c='0' then tr12=0;
            else tr12=1;
    tr13=transformacija13;
            if transformacija13-c='0' then tr13=0;
            else tr13=1;
        tr14=transformacija14;
            if transformacija14-c='0' then tr14=0;
            else tr14=1;
        tr15=transformacija15;
            if transformacija15-c='0' then tr15=0;
            else tr15=1;
        tr16=transformacija16;
            if transformacija16-c='0' then tr16=0;
            else tr16=1;
        tr17=transformacija17;
            if transformacija17-c='0' then tr17=0;
            else tr17=1;
        tr18=transformacija18;
            if transformacija18-c='0' then tr18=0;
            else tr18=1;
        tr19=transformacija19;
            if transformacija19-c='0' then tr19=0;
            else tr19=1;
        tr20=transformacija20;
            if transformacija20-c='0' then tr20=0;
            else tr20=1;
        tr21=transformacija21;
            if transformacija21-c='0' then tr21=0;
            else tr21=1;
        tr22=transformacija22;
            if transformacija22-c='0' then tr22=0;
            else tr22=1;
        tr23=transformacija23;
            if transformacija23-c='0' then tr23=0;
            else tr23=1;
        tr24=transformacija24;
            if transformacija24-c='0' then tr24=0;
            else tr24=1;
run;
```

```
          proc means data=transformacijos mean noprint;
               var tr1-tr24;
                    output out=stats;
          run;

          data vidurkiai(keep=error _STAT_);
               set stats;
                    where _STAT_='MEAN';
                    Error=min(of tr1-tr24);
          run;

          data paklaida;
               set vidurkiai(rename=(_STAT_=Method));
                    if Method="MEAN" then Method="k-means";
       run;

%mend;

options nonumber nodate;

%let p1=0.95;            /* Tikimybe */
%let p2=0.65;            /* Tikimybe */
%let p3=0.35;            /* Tikimybe */
%let p4=0.05;            /* Tikimybe */
%let nn=1;               /* n */
%let count=1000;          /* Generuoja stebejimu skaiciu */
%let arrayx=50;           /* Generuoja savybiu skaiciu */

%montecarlo;

%fclustering(clnr=4,itnr=10);

proc print data=paklaida noobs;
run;
```