Tartu Ülikool
Loodus- ja täppisteaduste valdkond
Matemaatika ja statistika instituut

Villem Lassmann

Kahe klasterdamismeetodi võrdlus TÜ Eesti Geenivaramu metaboloomika andmestiku näitel

Matemaatilise statistika eriala

Bakalaureusetöö (9 EAP)

Juhendaja: PhD Krista Fischer

Tartu 2019

Kahe klasterdamismeetodi võrdlus TÜ Eesti Geenivaramu metaboloomika andmestiku näitel

Lühikokkuvõte

Käesoleva bakalaureusetöö eesmärgiks on võrrelda kahte klasterdamismeetodit ning nende rakendamise tulemusel saadud klastrite informatsiooni suremusele, põhinedes metaboolika andmetele. Esmalt uuritakse meetodite klastrite erinevusi vanuse, soo ning kehamassiindeksi põhjal. Seejärel kasutatakse klastreid logistilise regressiooni mudelis, kus uuritakse suremust viie aasta jooksul. Viimasena kirjeldatakse mudelis statistiliselt oluliste tunnuste klastreite keskväärtusi ning keskväärtuste erinevust andmestiku keskmisest.

CERCS teaduseriala: P160 - Statistika, operatsioonanalüüs, programmeerimine, finants- ja kindlustusmatemaatika.

Märksõnad: Klasteranalüüs, kõrgemõõtmelised andmed, lineaarne regressioon, metaboloomika

Comparing two clustering methods on UT Estonian Gene Bank metabolomics dataset

Abstract

The aim of this thesis is to compare two clustering methods and the resulting clusters after applying the methods, including how much information the cluster give about dying, based on a metabolomics dataset. Firstly all of the clusters are compared based on age, sex and body mass index. After that the clusters are used in a logistic regression model, which predicts death in the next five years. Lastly the averages of variables in the clusters, which are statistically important in the model, are compared against the averages of the whole dataset.

CERCS research specialisation: P160 - Statistics, operations research, programming, actuarial mathematics

Keywords: Cluster analysis, high-dimensional data, linear regression, metabolomics

Sisukord

Sissejuhatus

Tänapäeval on saanud normiks, et kogutakse hulgaliselt andmeid. Sama tendents laieneb ka meditsiinile, kus on samuti tekkinud võimalus koguda suurandmeid, mille käsitsi uurimine on äärmiselt ressursirohke. Üheks selliseks andmete hulgaks on metaboloomika andmestik, mis on saadud inimeste vereproove tuumamagnetresonantsspektroskoopiaga uurides. Üheks võimalikuks meetodiks suuremahulisi andmeid uurida on kasutades klasterdamist, mis võimaldab andmetesse liialt süvenemata neid gruppeerida ning leida sarnasusi. Klasterdamismeetodeid on mitmeid ning nende valik sõltub peamiselt kasutusvaldkonnast. Kõige populaarsemad meetodid on k-keskmiste klasterdamine ning hierarhiline klasterdamine. Kuigi enamjaolt klasterdatakse vaatlusi, leidub vahel vajadus tunnuste jaotuse leidmiseks ning rakendatakse meetodeid tunnustele, leides nende klastrid. Üks vähem levinud klasterdamistüüpe on biklasterdamine, mis leiab vaatlusklastreid tunnuste alamhulkades ning tunnusklastreid vaatluste alamhulkades.

Käesoleva töö eesmärgiks on võrrelda kahte erinevat klasterdamismeetodit ning välja selgitada, kas tulemuseks saadud biklastrid sisaldavad informatsiooni inimese suremuse kohta. Esimene meetod on kombinatsioon k-keskmiste klasterdamisest ning hierarhilisest klasterdamisest, kus esmalt grupeeritakse andmestiku tunnused ning seejärel rakendatakse igas tunnuseklastris k-keskmiste klasterdamist. Teiseks meetodiks on spektraalne biklasterdamine, mis kasutab omaväärtusi ning -vektoreid optimaalsete klastrite leidmiseks.

Töö esimeses peatükis on antud ülevaade klasterdamisest, klasterdamismeetoditest, biklasterdamisest ning spektraalsest biklasterdamisest. Samuti on välja toodud teised statistilised meetodid, mida on töös kasutatud. Teises peatükis rakendatakse mõlemat meetodit Eesti Geenivaramu andmestikule ning võrreldakse meetodite tulemusi. Seejärel koostatakse logistilise regressiooni mudel ning uuritakse lähemalt mudelis olulisi tunnuseid. Kõik töös tehtud arvutused ning joonised on koostatud statistikapaketiga R.

1. Meetodite ülevaade

1.1 Klasterdamine

Järgnevad kolm peatükki põhinevad raamatul (James et al. 2013).
Klasterdamine viitab väga suurele meetodite hulgale, mille eesmärgiks on leida andmestikust alamhulki või klastreid. Klasterdada soovitakse nii, et andmestikust leitud klastrite sees olevad vaatlused oleksid üksteisele võimalikult sarnased ning klastrid oleksid üksteisest võimalikult erinevad. Selleks defineeritakse ka, mis määrab klastrisisese sarnasuse ning klastrite vahelise erinevuse, mis sõltub enamasti kasutusvaldkonnast.
Klasterdamist kasutatakse eelkõige suuremahuliste andmestike puhul (nii tunnuste kui vaatluste arvu mõttes) .Näiteks uuritakse erinevate patsientide geenide ekspressioonitaset ning soovitakse teada, kas leidub mingi peidetud struktuur. Klasterdamise abil leitakse erinevad grupid ning selle põhjal saab analüüsida, kas kindlal grupil on kõrgem või väiksem soodumus mõneks haiguseks. Teiseks kasutusalaks võib tuua turunduse, kus soovitakse näiteks teada millistele sihtgruppidele millist reklaami teha, grupeerides nad erinevate majanduslike näitajate (aastane sissetulek, peresuurus, jne) kaupa.
Kuna klasterdamist kasutatakse mitmetes valdkondades, siis leidub ka väga palju klasterdamismeetodeid. Kaks kõige populaarsemat meetodit on k-keskmiste klasterdamine ning hierarhiline klasterdamine. K-keskmiste puhul leitakse täpselt k klastrit, kuid hierarhilise klasterdamise puhul on väljundiks enamasti puu-kujuline graafik ehk dendrogramm, mis jagab kõik vaatlused või tunnused eri gruppidesse ning siis sarnasuse näitajate puhul hakkab neid ühendama.

1.1.1 K-keskmiste klasterdamine

K-keskmiste klasterdamine jagab kõik vaatlused või tunnused k-ks ühisosata klastriks. Defineeritakse, et $C_1, \ldots, C_K$ on hulgad, mis sisaldavad algseid vaatlusi. Hulkadel on kaks omadust

1. $C_1 \cup \cdots \cup C_K = \{1, \ldots, n\}$. Iga vaatlus kuulub vähemalt ühte hulka.

2. $C_K \cap C_{K'} = \emptyset \quad \forall K \neq K'$. Ühelgil hulkade paaril ei leidu ühisosa, ehk iga vaatlus kuulub täpselt ühte gruppi.

K-keskmiste meetodi puhul on klastri kvaliteet defineeritud läbi klastrisisese varieeruvuse. Klastrisisest varieeruvust klastris C_k näitab suurus $W(C_k)$, mis mõõdab seda, kui lähedal on klastris olevad vaatlused üksteisele. Sellest tulenevalt tahetakse minimiseerida

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\} \tag{1.1}$$

Ehk tahetakse jagada kõik vaatlused klastritesse nii, et kõikide klastrisiseste varieeruvuste summa oleks võimalikult väike.

Selleks, et klastreid saaks määrata, peab defineerima kuidas arvutatakse klastrisiseseid varieeruvusi. Selleks on samuti palju erinevaid võimalusi, kuid kõige populaarsem on Eukleidiline kaugus. Defineeritakse, et

$$W(C_K) = \frac{1}{|C_k|} \sum_{i,i' \in C_K} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2, \tag{1.2}$$

kus $|C_k|$ tähistab vaatluste arvu k-ndas klastris. Teiste sõnadega, klastrisisene varieeruvus k-nda klastri jaoks on summa üle kõikide vaatluspaaride eukleidilise kauguse ruudu, jagatud vaatluste arvuga klastris.

Kombineerides valemid (1.1) ja (1.2) saadakse järgnev avaldis

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_K} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\} \tag{1.3}$$

Järgmisena peab leidma algoritmi, millega minimeerida (1.3), ehk jaotada kõik vaatlused klastritesse nii, et (1.3) oleks minimiseeritud. Seda on raske teha väga täpselt, kuna klastritesse jagamiseks on pea K^n erinevat võimalust. Ehk mida suurem on K ja n, seda aega nõudvamaks ja keerulisemaks jaotamine läheb. Üks võimalus selleks on kasutada algoritmi, millega leitakse lokaalne miinimum. Algoritm on järgnev

1. Suvaliselt määrata igale vaatlusele suvaline täisarv vahemikus $[1, \ldots, K]$. Need on iga vaatluse algsed klastrid.

2. Korratakse, kuni vaatluste klastrid enam ei muutu.

   (a) Iga klastri $j = 1, \ldots, K$ jaoks arvutada klastri keskpunkt. Klastri keskpunktiks loetakse vektorit, mis koosneb iga p tunnuse keskväärtusest: $\bar{x}_j = (\bar{x}_{j1}, \ldots, \bar{x}_{kp})$

(b) Määratakse igale vaatlusele eukleidilise kauguse mõttes kõige lähemal olev klaster uueks klastriks.

Antud algoritm garanteerib, et iga sammuga väheneb valemis (1.3) otsitav väärtus. Seda saab näidata valemiga

$$\frac{1}{|C_k|}\sum_{i,i'\in C_K}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2 = 2\sum_{i\in C_k}\sum_{j=1}^{p}(x_{ij}-\bar{x}_{kj})^2, \tag{1.4}$$

kus $\bar{x}_{kj} = \frac{1}{|C_k|}\sum_{i\in C_k} x_{ij}$ on tunnuse j keskväärtus klastris C_k. Algoritmi sammul 2(a) on klastri keskmised iga tunnuse jaoks konstandid, mis minimiseerivad summa ruutude varieeruvust. Sammul 2(b) klastrite ümber määramisel saab (1.4) ainult väheneda. Kuna algoritm töötab nii kaua, kuni ümber määramist enam ei toimu, siis (1.4) väheneb igal sammul ning me saavutame lokaalse miinimumi (1.3).
Kuna k-klasterdamine leiab lokaalse miinimumi, siis tulemused sõltuvad mingil määral algselt valitud klastritest. Seetõttu on vajalik algoritmi mitmekordne jooksutamine, mille tulemusel saab valida kõike väiksema klastrite sisese variatsiooniga tulemuse. Teine probleem k-klasterdamisega on, et selle kasutamisel tuleb ära määrata K, ehk mitu klastrit väljastatakse. Kõige lihtsam lahendus probleemile on kasutada fikseeritud klastrite arvu. Teiseks võimaluseks on rakendada meetodit andmestikule mitu korda järjest, suurendades iga kord klastrite arvu, kuni klastrisisene varieeruvus väga ei muutu. See arv valitaks optimaalseks klastrite arvuks. Viimast lähenemist nimetatakse küünarnuki meetodiks.

1.1.2 Hierarhiline klasterdamine

K-keskmiste kõige suurem probleem on, et tuleb valida klastrite arv. Hierarhiline klasterdamine on alternatiivne meetod, mis ei nõua klastrite arvu valikut. Teine põhjus hierarhilise klasterdamise valikuks on selle väljundiks olev lihtsasti loetav dendrogramm.

Dendrogrammi interpreteerimine

Dendrogrammi kujutatakse tagurpidi puuna, kus puu *lehed* asuvad graafiku kõige alumises osas. Liikudes mööda puud üles, ehk y-telge mööda, hakkavad lehed liituma, moodustades *oksad*. Lehtede liitumine oksaks näitab, et antud vaatlused on

üksteisele sarnased. Mida madalamal liitumine toimub, seda sarnasemad vaatlused on. Vastupidiselt, mida kõrgemal liitumine toimub seda erinevamad vaatlused on. Teiste sõnadega, suvalise kahe vaatluse sarnasust saab mõõta selle põhjal kui kõrgel nad puus esimest korda liituvad. See tähendab, et kaks punkti, mis võivad tunduda algse andmestiku põhjal väärtustelt sarnased, võivad olla hierarhilise klasterdamise tulemusel kahes väga erinevas klastris.

Hierarhilise klasterdamise tulemusena saab valida dendrogrammilt $\{1, \ldots, n\}$ erinevat klastrit, olenevalt määratud sarnasuse kordajast, ehk kõrgusest y-teljel. See tähendab, et kõrguse määramine on tehniliselt sama rolliga, mis K määramine k-klasterdamises. Hierarhiline tähendab seda, et iga madalalt määratud klaster sisaldub mõnes kõrgemalt lõigatud klastris. Kõige kõrgemat klastrit, mis sisaldab kõiki vaatlusi, nimetatakse puu *juureks*.

Hierarhilise klasterdamise algoritm

Hierarhilise klasterdamise dendrogramm leitakse järgneva algoritmiga. Esmalt defineeritakse ära erinevuse näitaja kahe vaatluse vahel. Nagu ka k-klasterdamises, on hierarhilises klasterdamises kõige kasutatum mõõdik eukleidiline kaugus. Algselt on iga vaatlus eraldi klaster, ehk kokku n klastrit. Seejärel leitakse kaks klastrit, mis on kõige sarnasemad, ehk kõige väiksema erinevusega. Need klastrid liidetakse omavahel, mille tulemusel jääb alles $n - 1$ klastrit. Algoritmi jätkatakse kuni on alles ainult üks klaster. Hierarhiline klasterdamine koosneb seega järgmistest sammudest

1. Alustatkse n klastriga, erinevuse mõõdikuga (näiteks eukleidiline kaugus) ning kõikide $\binom{n}{2} = \frac{n(n-1)}{2}$ paaride erinevustega. Iga vaatlus on eraldi klastris.
2. Iga $i = n, n-1, \ldots, 2$ puhul
   (a) Üle kõikide paaride, leitakse kõige väiksema erinevusega paar, ehk kõige sarnasemad, ning liidetakse need kokku. Dendrogrammil on liitumise kõrguseks erinevuse suurus.
   (b) Lietakse uued klastrite vahelised erinevused üle kõikide $i-1$ alles jäänud klastri.

Järgmiseks jääb lahendada probleem, kuidas arvuatatakse kahe klastri erinevust, kui vähemalt ühes on rohkem kui üks element. Selleks lisatakse *linkage* ehk sidusus, mille kaudu defineeritakse kahe klastri vahelist erinevust. Neli enimkasutatavat tüüpi sidususi on:

1. *Complete linkage* (maksimaalne sidusus) ehk maksimaalne klastrite vaheline erinevus. Arvutatakse kõikide klastris A ja klastris B olevate elementide paaride vahelised kaugused. Seejärel loetakse klastrite vaheliseks erinevuseks maksimaalne erinevus.

2. *Single linkage* (minimaalne sidusus) ehk minimaalne klastrite vaheline erinevus. Sarnaselt eelnevale, leitakse kõikide klastris A olevate ja klastris B olevate elementide paaride vahelised kaugused. Seejärel loetakse erinevuseks minimaalne erinevus.

3. *Average linkage* (keskmine sidusus) ehk keskmine klastrite vaheline erinevus. Leitakse kõikide klastris A olevate ja klastris B olevate elementide paaride vahelised kaugused. Seejärel loetakse erinevuseks kõikide erinevuste keskmine.

4. *Centroid linkage* (keskväärtuste sidusus) ehk keskväärtuste erinevus. Leitakse klastrite A ja B keskpunktid ning seejärel arvutatakse kahe keskväärtuse vaheline erinevus.

Statistikas kasutatakse enamasti maksimaalset, minimaalset või keskmist sidusust. Tihti kasutatakse hierarhilises klasterdamises ka teisi erinevuse mõõdikuid, kui eukleidiline distants. Statistikas kasutatakse kauguse mõõdikutena tihtipeale Spearmani ja Pearsoni korrelatsioonikordajaid tunnuste vahel.
Erinevuse mõõdiku valik on eriti oluline, kuna erinevad mõõdikud võivad anda väga erinevaid dendrogramme. Erilist tähelepanu tuleb pöörata sellele, mis tüüpi andmetega on tegu ning mida uuritakse ja selle põhjal leida sobiv mõõdik.
Hierarhilise klasterdamise puhul on teatud olukordades kasulik andmeid eelnevalt normaliseerida, kuna tihti tuleb ette, et suure dispersiooniga tunnused mõjutavad rohkem klastrite vahelisi erinevusi, mõjutades sellega ka lõpliku dendrogrammi - on olukordi, kui see siiski pole soovitav. Normaliseerides andmed, antakse igale tunnusele võrdne kaal. Nagu ka erinevuse mõõdiku valikuga, oleneb normaliseerimine

sellest, mis on uurimuse eesmärk.

Statistikas tuleb ette ka väga palju olukordi, kus uuritakse vaatluste asemel tunnuseid. Tunnuste hierarhiliseks klasterdamiseks kasutatakse eelkõige kauguse mõõdikuna korrelatsiooni. Eelnevalt mainitud Spearmani ja Pearsoni korrelatsioonikordajad on kõige enam kasutusel olevad. Tunnuste klastrid võivad anda andmestiku kohta väga palju informatsiooni, kuid vahel kasutatakse ka tunnuste klasterdamist selleks, et koostada alamandmestik, kuhu võetakse väga tugevalt või nõrgalt korreleeritud andmed, mida edasi uurida.

1.2 Biklasterdamine

Biklasterdamise alla klassifitseeritakse klasterdamise algoritme, mis klasterdavad korraga nii vaatluseid kui ka tunnuseid. Võrreldes tavalise klasterdamisega, leiavad biklasterdamise algoritmid globaalse mudeli asemel lokaalse mudeli. Näiteks tavalise klasterdamise algoritmiga leitakse kõik tunnuste klastrid üle kõikide vaatluste. Samuti leitakse kõik vaatluste klastrid üle kõikide tunnuste. Biklasterdamises leitakse vaatluste klastrid üle tunnuste alamhulga ning tunnuste klastrid üle vaatluste alamhulga. Teiste sõnadega, biklasterdamise algoritmid leiavad vaatluste klastrid, mis on sarnased ainult mingis kindlas tunnuste alamhulgas. (Madeira ja Oliveira 2004)

1.2.1 Spektraalne biklasterdamine

Järgnev alapeatükk põhineb artiklil (Kluger et al. 2003).

Spektraalne biklasterdamine eeldab, et saame eristada vaatluste klastreid, mis on sarnased vaid teatud tunnuste rühma lõikes ning tunnuste klastreid, mis on korreleeritud teatud vaatluste klastri sees. Samuti eeldatakse, et andmestikul leidub peidetud klassifitseerimismaatriks. Andmestik viiakse lõpuks plokk-maatriks kujule, kus iga plokk iseloomustab ühte vaatlus-tunnus klassi. Eeldame, et andmestiku A vaatluse i ja tunnuse j väärtust saab ligikaudselt esitada kolme sõltumatu faktori põhjal: $A_{ij} = E_{ij} \cdot \rho_i \cdot \chi_j + \epsilon_{ij}$, kes ϵ_{ij}. Esimene faktor on peidetud baastase, mida tähistatakse E_{ij}. Eeldatakse, et maatriksi E elemendid on igas plokis konstantsed: $E_{ij} = E_{i'j'}$, kui vaatlused i ja i' ning tunnused j ja j' kuuluvad samasse klastrisse. Teine faktor, mis on tähistatud ρ_i, iseloomustab vaatluse i väärtuste taset üle kõigi

tunnuste. Viimane faktor, mis on tähistatud χ_j, iseloomustab tunnuste j väärtuste taset üle kõigi vaatluste.

Oletame, et maatriks A: $n \times m$ on faktorite E_{ij}, ρ_i ja χ_j ligikaudne korrutis. Biklasterdamise eesmärk on, leida maatriks E, eeldades, et A on antud. Võttes suvaliselt kaks vaatlust i ja k, mis kuuluvad sarnaste vaatluste alamhulka, siis keskmiselt erineb nende väärtus ρ_i/ρ_k korda. Järelikult kui normaliseerida maatriksi A read i ja k, siis keskmiselt peaks olema need read võrdsed. Kahe vaatluse väärtuste sarnasus on veel ilmsem, kui võtta väärtuste keskmised üle kõikide sarnaste tunnuste. Olgu R diagonaalmaatriks, mille element r_i on maatriksi A rea i summa, ehk $R = diag(A \cdot 1_m)$, kus 1_m tähistab $m \times 1$ vektorit, mille kõik elemendid on 1. Samuti olgu $u = (u_1, u_2, \ldots, u_m)$ tunnuste klassifitseerimisvektor, kus u on konstantne üle kõikide sarnaste tunnuste. Kui järjestada tunnused nii, et kõik ühte klassi kuuluvad tunnused on kõige ees, siis on järjestatud ka vektor u. Sellest tulenevalt $v = R^{-1}Au$ on ligikaudne vaatluste klassifitseerimisvektor, mille elemendid on konstantsed üle sarnaste vaatluste. Korrutades maatriksit A vasakult maatriksiga R^{-1}, normeeritakse maatriksi A read. Korrutades tulemust omakorda vektoriga u, saadakse väärtuste keskmiste kaalutud summa üle kõikide tunnuste. Kui leidub peidetud struktuur, siis kõikide sarnaste vaatluste väärtuste keskmiste kaalutud summa on sama.

Sarnaselt saab leida tunnuste struktuuri. Olgu maatriks C diagonaalmaatriks, mille elemendid on maatriksi A veergude summad, $C = diag(1_n^T \cdot A)$. Korrutades $C^{-1}A^Tv$, saame iga tunnuse j kohta väärtuste keskmiste kaalutud summa.

Järgnevalt korrutades maatriksit $C^{-1}A^TR^{-1}A$ tunnuste klassifitseerimisvektoriga u, saadakse samuti tunnuste klassifitseerimisvektori. Tähistame seda maatriksit M_1. Maatriks M_1 on positiivselt poolmääratud, leiduvad on ainult reaalarvulised mittenullilised omaväärtused ning maatriksi domineeriv omavektor on $(1/\sqrt{m})1_m$, millele vastav omaväärtus on 1. Eeldame ka, et $rank(E) = min(n_c, n_r)$, kus n_c on tunnuste klasside arv ning n_r on vaatluste klasside arv. Sellest tulenevalt saame öelda, et eksisteerib vähemalt üks vektor, mis rahuldab võrdust

$$M_1u = \lambda u \tag{1.5}$$

Üks sellistest vektoritest on triviaalne vektor $(1/\sqrt{m})1_m$. Samuti leidub ka vähemalt üks vaatluste klassifitseerimisvektor v, mis rahuldab võrdust

$$M_2v = \lambda v \tag{1.6}$$

, kus $M_2 = R^{-1}AC^{-1}A^T$. Siinkohal tuleb tähele panna, et maatriksitel M_1 ja M_2 on samad omaväärtused, kuna varem on defineeritud, et $v = R^{-1}Au$.
Vastavad omaväärtused saab leida, kui leida singulaarväärtuste lahutus maatriksil $\hat{A} = R^{-1/2}AC^{1/2}$, võrdust

$$\hat{A}^T\hat{A}w = C^{-1/2}A^TR^{-1}AC^{-1/2}w = \lambda w \tag{1.7}$$

mida kasutatakse maatriksi $\hat{A}$ singulaarväärtuste leidmiseks, on ekvivalentne võrrandile (1.5), kus $u = C^{-1/2}w$. Sarnaselt on ka võrdus

$$\hat{A}\hat{A}^Tz = R^{-1/2}AC^{-1}A^TR^{-1/2}z = \lambda z \tag{1.8}$$

ekvivalentne võrrandile (1.6), kus $v = R^{-1/2}z$.
Maatriksite omavektorid ongi otsitavad tunnuste ja vaatluste klassifitseerimisvektorid u ja v.

1.2.2 Jaccardi indeks

Jaccardi indeks on kahe hulga sarnasust näitav suurus. Indeks leiab kasutust erinevates valdkondades, nagu informatsiooni taastamine, andmekaeve ja masinõpe. Jaccardi indeks mõõdab kahe lõpliku hulga suhtelist ühisosa. (Kosub 2016) Indeks on defineeritud järgnevalt

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} \tag{1.9}$$

Valemis on hulga norm selles hulgas olevate elementide arv. Jaccardi indeksid kasutatakse klasterdamise puhul kahe klastri saranasuse mõõtmiseks, kus kahe täpselt sama klastri puhul on indeks 1 ning kahe täiesti erineva klastri puhul on indeks 0.

1.3 Logistiline regressioon

Järgnev peatükk põhineb Ene Kääriku konspektile (Käärik 2013).
Binaarse uuritava tunnuse puhul kasutatakse logistilist mudelit, millega hinnatakse šansi logaritmi. Sündmuse esinemise šanss on defineeritud kui $\Pi = \frac{\pi}{1-\pi}$, kus $\pi = \mathbf{P}(Y = 1)$, ehk sündmuse esinemise tõenäosus. Logistiline mudel on defineeritud kui

$$ln\frac{\pi}{1-\pi} = \beta_0 + \beta_1X_1 + \cdots + \beta_kX_k \tag{1.10}$$

Antud juhul $X_1, X_2, \ldots, X_k$ on mudeli argumenttunnused ning $\beta_0, \beta_1, \ldots, \beta_k$ on mudeli parameetrid.

Šansside suhe on defineeritud kui vaatluse i ja j šansside jagatis

$$OR = \frac{\Pi_i}{\Pi_j} = \frac{\frac{\pi_i}{1-\pi_i}}{\frac{\pi_j}{1-\pi_j}} \tag{1.11}$$

Šansside suhte usaldusvahemiku leidmiseks kasutatakse asümptootilist χ^2-jaotust. Kui usaldusvahemikku jääb arv 1, siis võib öelda, et tegemist on statistiliselt ebaolulise šansside suhtega olulisuse nivool α.

1.4 Spearmani korrelatsioonikordaja

Speramani astakkorrelatsioonikordaja on mitteparameetriline suurus korrelatsiooni mõõtmiseks. Olgu meil kaks hulka $X = (X_1, X_2, \ldots, X_n)$ ning $Y = (Y_1, Y_2, \ldots, Y_n)$. Olgu R_{X_i} hulga X elemendi i astak ning R_{Y_i} hulga Y elemendi i astak. (Dodge 2008) Siis on Spearmani korrelatsioonikordaja defineeritud kui

$$\rho = 1 - \frac{6 \cdot \sum_{i=1}^{n} (R_{X_i} - R_{Y_i})^2}{n(n^2 - 1)} \tag{1.12}$$

2. TMR-metaboloomika andmestiku analüüs

2.1 TMR-metaboloomika andmestik

Tegemist on kaheosalise andmestikuga, mis mõlemad pärinevad Eesti Geenivaramust.

Andmestiku esimene osa koosneb 10840 vaatlusest ning 225 tunnusest, kus iga vaatluse puhul on tegemist eraldi inimesega. Andmed on kogutud Eestis aastatel 2002 kuni 2011 vabatahtlikelt üle Eesti. Kõik vabatahtlikud olid vanuses 18 – 103. Kõik tunnused on leitud tuumamagnetresonantsspektroskoopiaga, ehk TMR-spektroskoopiaga. Tunnuste puhul on tegu metaboliidide, lipiidide ja lipoproteiinidega. Lipiidide puhul on tunnustes kasutatud lühendeid, mis tähendavad

- VLDL - *Very low density lipid*, ehk väga madala tihedusega lipiid
- LDL - *Low density lipid*, ehk madala tihedusega lipiid
- IDL - *Intermediate density lipid* ehk keskmise tihedusega lipiid
- HDL - *High density lipid* ehk suure tihedusega lipiid

Teises andmestikus on 10802 vaatlust ning 16 tunnust, kus iga vaatluse puhul on jällegi tegemist eraldi inimesega. Andmed on kogutud samadelt inimestelt, mis esimese andmestiku puhul ning tulevad Eesti Geenivaramuga liitumisel täidetud küsimustikust.

Mõlemat klasterdamismeetodit rakendati metaboliitide andmestiku peal. Kuna erinevate metaboliitide väärtuste vahemikud on erinevad, siis enne meetodite rakendamist normeeriti kõik veerutunnused nii, et iga tunnuse keskväärtus oleks 0 ning standardhälve oleks 1. Seejärel liideti esialgsele andmestikule teine andmestik ning kombineeritud andmestikku kasutati analüüsi tegemiseks.

2.2 Kasutatud tarkvara

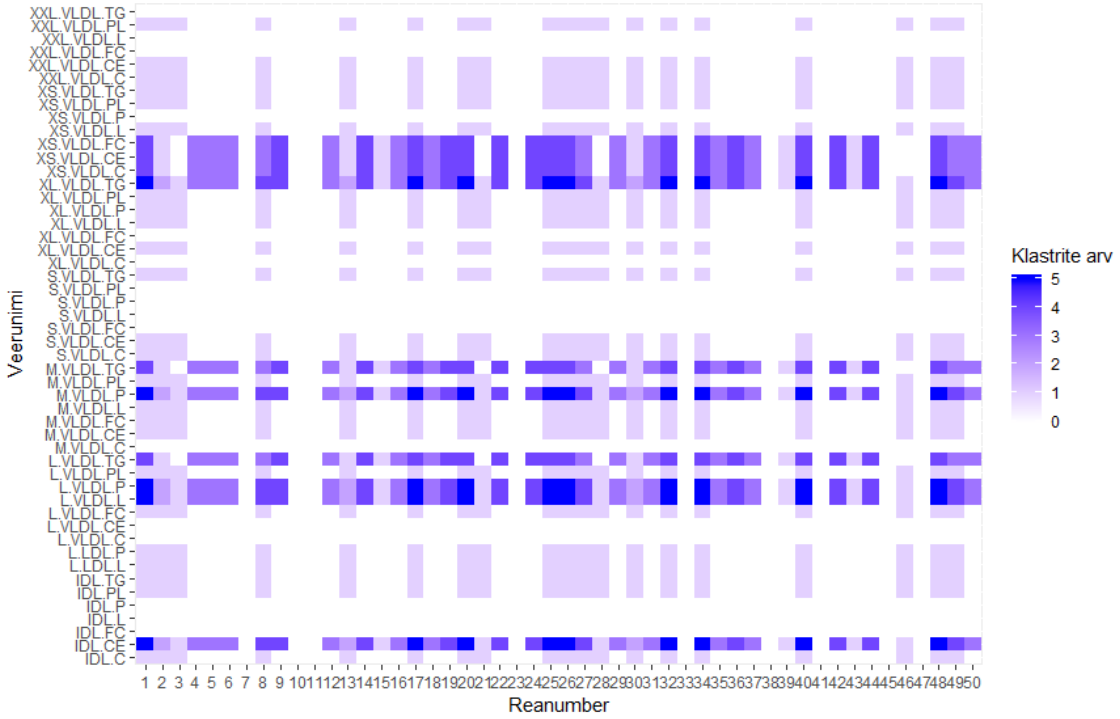
Antud töö raames kasutati statistikapaketti R, versiooni 3.5.1 (R Core Team 2018).
Biklasterdamiseks kasutati paketti *biclust* (Kaiser et al. 2018).
Hierarhiliseks klasterdamiseks kasutati R baaskäsklust *hclust* ning k-keskmiste klasterdamiseks R baaskäsklukst *kmeans* (R Core Team 2018).
Kõikide jooniste tegemiseks kasutati paketti *ggplot*2 (Wickham 2016).

2.3 Tulemused

2.3.1 TMR-metaboloomika andmestiku spektraalne biklasterdamine

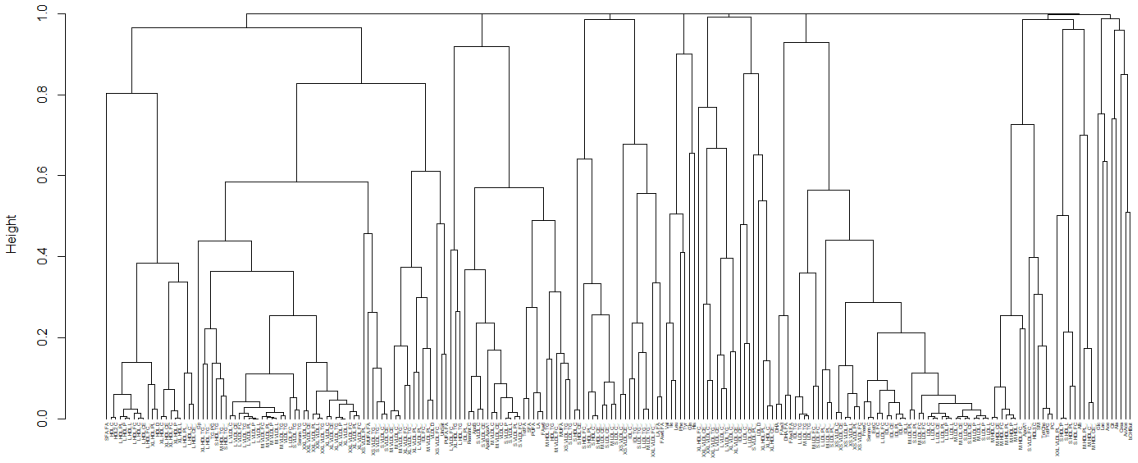
Esimesena kasutati andmestikul spektraalse biklasterdamise algoritmi.
Selleks, et paremini aru saada millised on spektraalse biklasterdamise tulemused, koostati alamandmestik, mis koosnes esialgse andmestiku esimesest viiekümnest reast ja viiekümnest veerust. Alamandmestiku peal spektraalset biklasterdamist kasutades saadi tulemuseks kuus erinevat klastrit. Kuna mitmes erinevas klastris võivad esineda samad vaatlused ja tunnused, siis on joonisel (2.1) välja toodud, mitu korda mingi vaatlus ja tunnus esines mõnes klastris. Jooniselt on näha, et andmestikus leiduvad mõned tunnused, mis esinevad klastrites rohkem kui teised, näiteks $XS.VLDL.TG$, ning mõned tunnused, mida ei ole üheskis klastris, näiteks $IDLP$. Sarnast mustri on näha ka ridade puhul, et näiteks vaatlus number 10 ei esine üheski klastris, kuid vaatlus number 25 on mitmes erinevas klastris.

Normeeritud andmestiku peal jooksutati seejärel algoritmi ning tulemuseks saadi 259 klastrit, kuid sellest vaadatakse esimest 24 klastrit, mis kuuluvad 7 erineva veeruklastri alla. Nagu ka alamandmestiku spektraalse biklasterdamise puhul, siis leidub ka selle andmestiku puhul ridu ja veerge, mis kuuluvad rohkem kui ühte klastrisse. Kõik spektraalse biklasterdamise tulemusel saadud klastrid algavad nimetusega *bic* ning kõik veeruklastrid algavad nimetusega *bklast*. Ühte veeruklastrisse kuuluvad kõik klastrid, millel on täpselt samad tunnused ning millel ei esine ühtegi korduvat rida. Näiteks kuulub veeruklastrisse *bklast*1 klastrid *bic*1, *bic*2 ning *bic*3.



Joonis 2.1: Alamandmestiku spektraalse biklasterdamise tulemus

2.3.2 TMR-metaboloomika andmestiku hierarhiline ning k-keskmiste klasterdamine



Joonis 2.2: Veergude hierarhilise klasterdamise tulemusel saadud dendrogramm

Teisena kasutati normaliseeritud andmestiku veergude hierarhilist klasterdamist. Selleks leiti kõikide tunnuste omavahelised Spearmani korrelatsioonikordajad. Korrelatsioonikordajatest seejärel koostati korrelatsioonimaatriks. Saadud korrelatsioonimaatriksi elementidest võeti absoluutväärtus ning lahutati arvust 1. Tulemuseks oli maatriks, mille elemendid olid intervallis $[0, 1]$, kus 0 tähistas tugevat seost ning 1 seose puudumist. Saadud maatriksi elemente kasutati kauguse mõõdikuna hierarhilises klasterdamises. Hierarhilises klasterdamises kasutati *Complete linkage* ehk maksimaalset klastrite vahelist erinevust. Tulemuseks saadi joonisel 2.2 olev dendrogramm.

Seejärel võeti dendrogrammi alusel kaheksa kõige tugevamalt korreleeritud veergude gruppi ning igas grupis tehti k-keskmiste klasterdamist, kus $k = 3$. Tulemusena saadi kokku 24 erinevat klastrit. Kõik hierarhilise ja seejärel k-keskmiste klasterdamise tulemusel saadud klastrid algavad nimetusega *hic* ning kõik veeruklastrid algavad nimetusega *hklast*. Nii hierarhiline k-keskmiste klasterdamine kui ka hierarhiline klasterdamine jagavad kõik tunnused ja vaatlused lõikumatuteks hulkadeks. See tähendab, et jällegi kuuluvad ühte veeruklastrisse kõik klastrid, millel on täpselt samad tunnused ning millel ei esine ühtegi korduvat rida. Näiteks veeruklastrisse *hklast*1 kuuluvad klastrid *hic*1, *hic*2 ja *hic*3.

2.3.3 Klasterdamise tulemuste analüüs

Tabelis 2.1 on välja toodud nii spektraalse biklasterdamise kui ka hierarhilise ja k-keskmiste klasterdamise tulemusel saadud klastrid. Tabelist on näha, et klastrid tulevad enamasti sama suurusega mõlema meetodi puhul. Samas on on näha, et kuigi kesmiselt tulevad klastrid ridade arvu poolest sarnase suurusega, siis hierarhilise ja k-keskmiste klasterdamise puhul on ridade arvu varieeruvus suurem. Veergude puhul on näha, et hierarhilise ja k-keskmiste klasterdamise puhul on palju suuremaid ja väiksemaid veerge ning biklasterdamise puhul on jällegi veergude suurus stabiilsem.

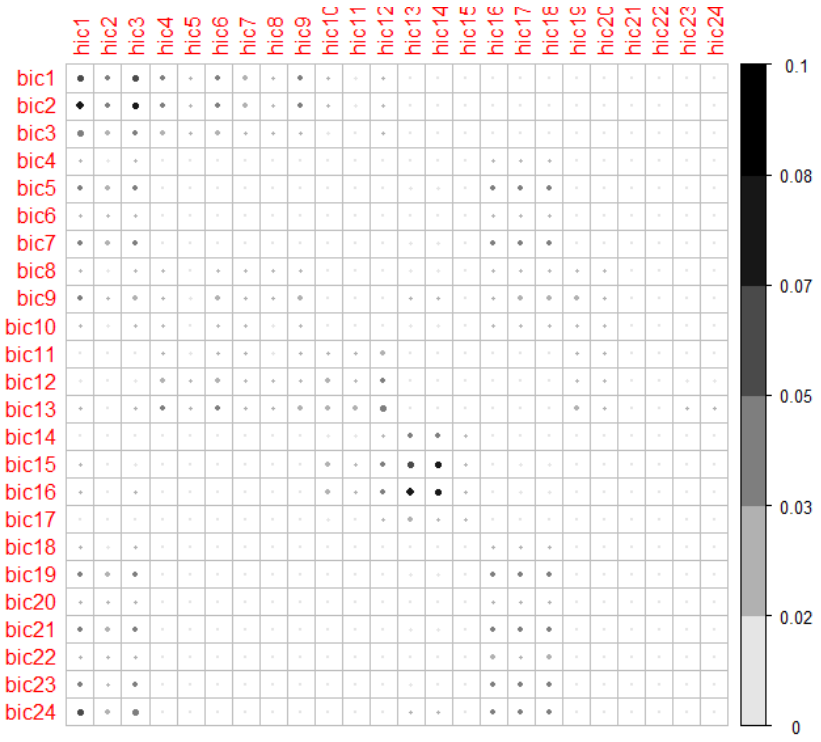
Kahe klasterdamismeetodi klastrite elementide võrdluseks vaadatakse, millised elemendid on klastritel ühised. Selleks kasutatakse antud töö raames Jaccardi indeksit, mis leiab suhte kahe klastrite ühisosa ning ühendi vahel.

Tabel 2.1: Kahe klasterdamismeetodi tulemusel saadud klastrid.

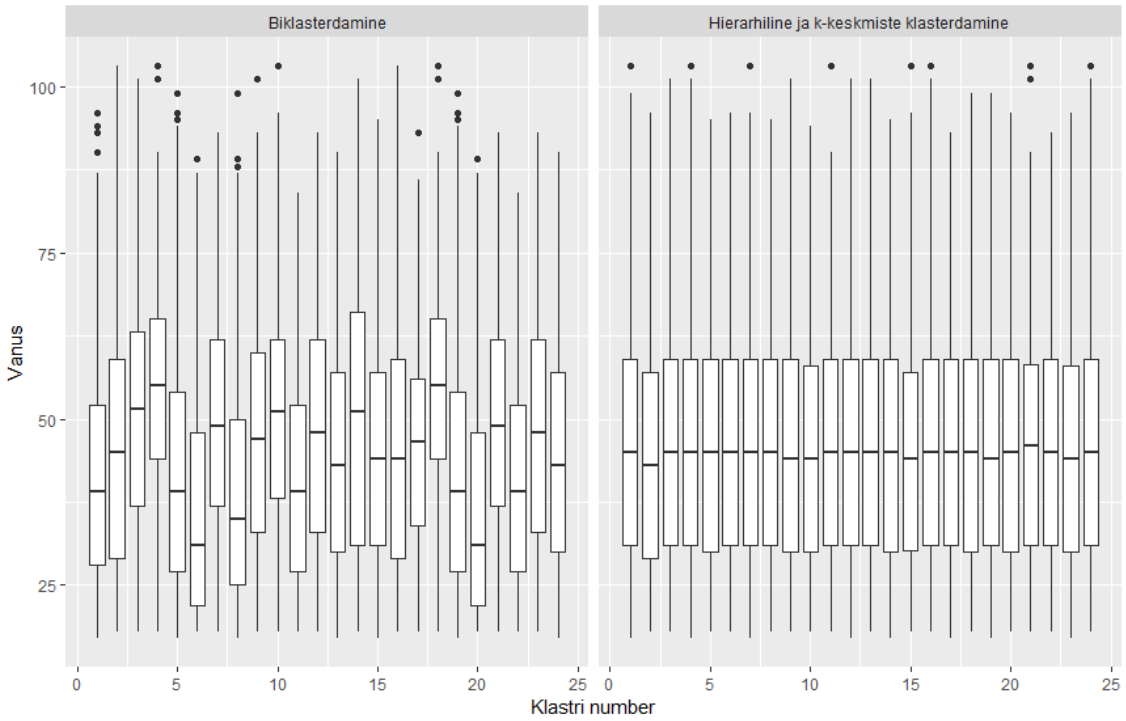
Biklasterdamine			Hierarhiline klasterdamine		
Klastri nimetus	Ridu	Veerge	Ridu	Veerge	Klastri nimetus
Bklast1 bic1	3579	48	5120	75	Hklast1 hic1
Bklast1 bic2	4129	48	1418	75	Hklast1 hic2
Bklast1 bic3	2410	48	4302	75	Hklast1 hic3
Bklast2 bic4	1687	22	4188	27	Hklast2 hic4
Bklast2 bic5	3915	22	1903	27	Hklast2 hic5
Bklast2 bic6	1605	22	4749	27	Hklast2 hic6
Bklast2 bic7	3633	22	3655	47	Hklast3 hic7
Bklast3 bic8	2930	36	2365	47	Hklast3 hic8
Bklast3 bic9	4368	36	4820	47	Hklast3 hic9
Bklast3 bic10	2807	36	2882	23	Hklast4 hic10
Bklast4 bic11	2257	35	2515	23	Hklast4 hic11
Bklast4 bic12	3238	35	5443	23	Hklast4 hic12
Bklast4 bic13	4505	35	4924	20	Hklast5 hic13
Bklast5 bic14	1481	23	4677	20	Hklast5 hic14
Bklast5 bic15	3976	23	1239	20	Hklast5 hic15
Bklast5 bic16	4172	23	3278	17	Hklast6 hic16
Bklast5 bic17	1211	23	3626	17	Hklast6 hic17
Bklast6 bic18	1687	22	3936	17	Hklast6 hic18
Bklast6 bic19	3915	22	5948	8	Hklast7 hic19
Bklast6 bic20	1605	22	4086	8	Hklast7 hic20
Bklast6 bic21	3633	22	806	8	Hklast7 hic21
Bklast7 bic22	2257	22	1684	7	Hklast8 hic22
Bklast7 bic23	3238	22	4648	7	Hklast8 hic23
Bklast7 bic24	4505	22	4508	7	Hklast8 hic24

Indeks arvutati paariviisi kõikide biklasterdamise ning hierarhilise ja k-keskmiste klasterdamise tulemusel saadud klastrite vahel kasutades valemit (1.9), kus $|A \cup B|$

on klastri A ja klastri B ühisosa elementide arv. Seejärel koostati indeksitest maatriks. Joonisel 2.3 on näha arvutuse tulemused. Jooniselt on näha, et kõige tugevamalt on seotud biklastrid 15-16 ning hierarhilise ning k-keskmiste klasterdamise klastrid 13-45. Tegemist on väga nõrkade seostega, keskmiselt seosekordajaga $0,083$. Veel on näha, et leidub ka seos hierarhilise ja k-keskmiste klasterdamise klastrite 1 ja 3 ning biklasterdamise klastrite 1, 2 ning 24 vahel. Needki seosed on väga nõrgad, keskmiselt seosekordajaga $0,075$. Ülejäänud klastrite vahel on näha seoseid, kuid need on väga nõrgad, alla $0,05$, ning ei vääri mainimist.

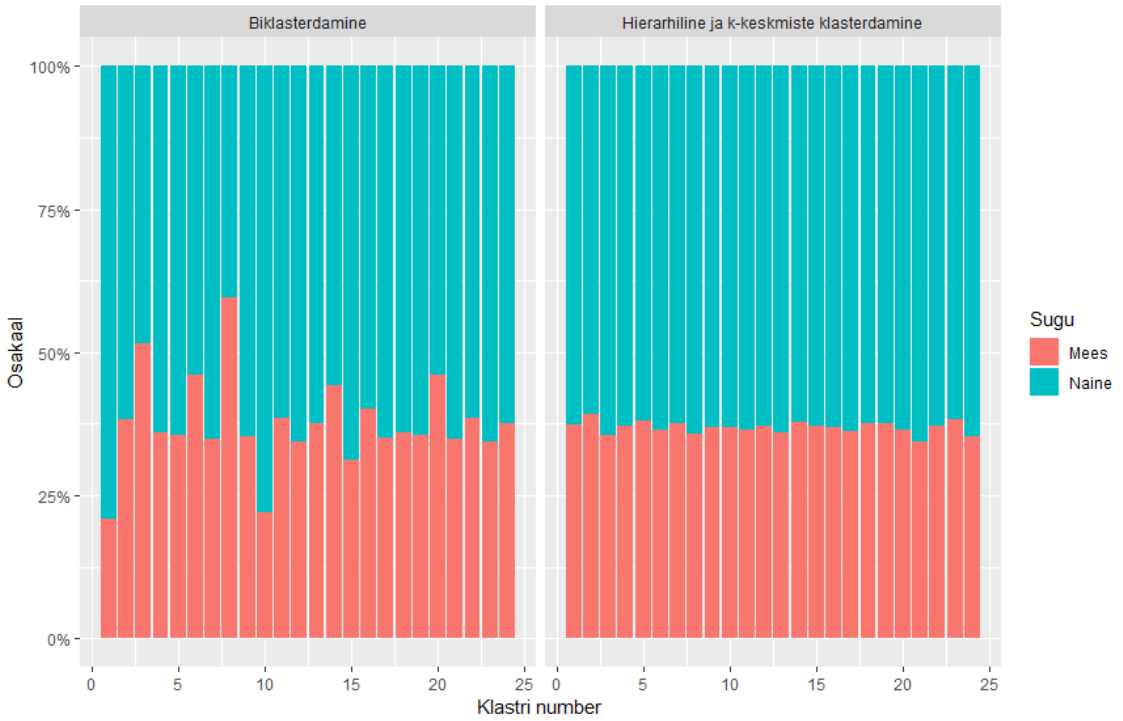


Joonis 2.3: Kahe klasterdamismeetodi klastrite vahelised Jaccardi indeksid



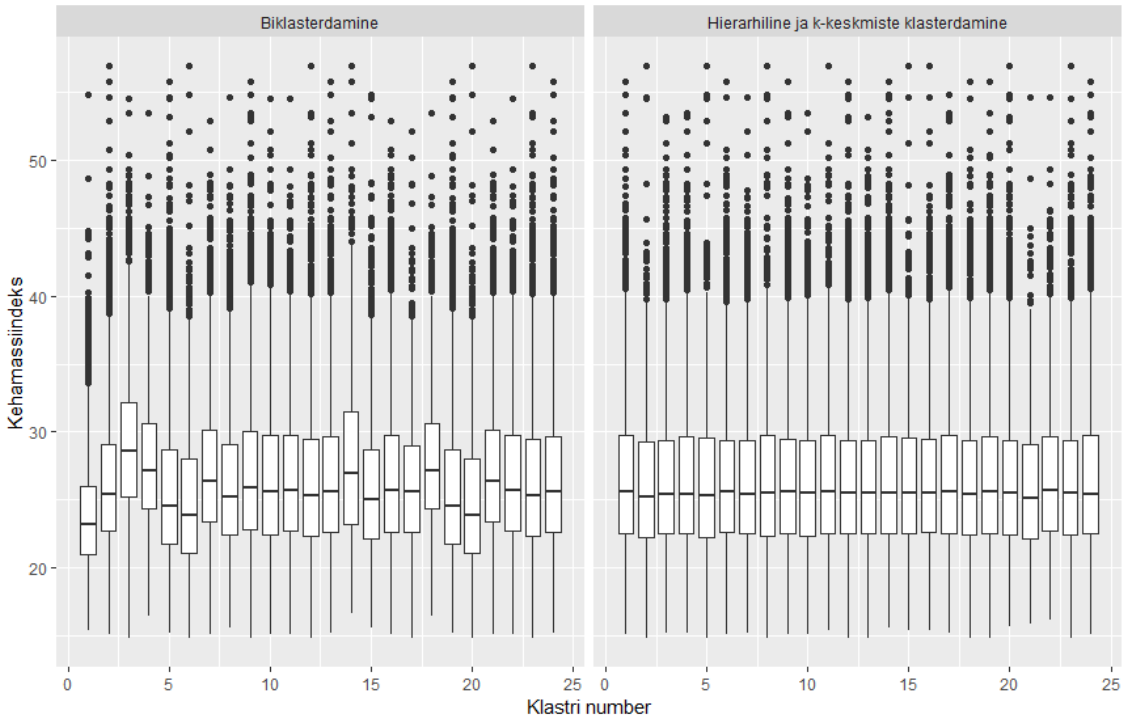
Joonis 2.4: Vanuse jagunemine kahe erineva klasterdamismeetodi rakendamise tulemusel saadud klastrites

Jooniselt 2.4 on näha kuidas jaotuvad erinevates klastrites inimeste vanused. Hierarhilise ja k-keskmiste klasterdamise puhul on näha, et kõikide klastrite keskmised vanused on väga lähestikku. Enamikel klastritel on täpselt sama keskmine vanus. Jooniselt ilmneb, et enamasti esimese ja kolmanda kvartiili piirid on võrdsed. Biklasterdamise klastrite puhul on näha, kuidas klastrid on vanuste suhtes väga erinevad. Näiteks on klastrite 4, 7 ja 18 keskmine vanus üsnagi kõrgel, ligikaudu 55. Samas on näha, et näiteks klastritel 8 ja 20 on keskmised vanused pigem madalad. Antud klastrite puhul on esimese ja kolmanda kvartiili vahe klastriti väga erinev. Näiteks klastrite 4, 7 ja 18 puhul on esimese ja kolmanda kvartiili vahe väike. Võrdluseks on näiteks klastri 14 esimese ja kolmanda kvartiili vahe üsna suur.



Joonis 2.5: Soo jagunemine kahe erineva klasterdamismeetodi rakendamise tulemusel saadud klastrites

Jooniselt 2.5 on näha sooline jaotuvus klastrites. Hierarhilise ja k-keskmiste klasterdamise puhul on näha, et keskmiselt on igas klastris meeste osakaal umbes 37%. Biklasterdamise klastrite puhul on näha, et näiteks klastritel 1 ja 10 on meeste osakaal teistest väiksem, umbes 24%. Kuid näiteks klastrite 3 ja 8 puhul on näha, et mõlemal klastril on meeste osakaal üle 50%.



Joonis 2.6: Kehamassiindeksi jagunemine kahe erineva klasterdamismeetodi rakendamise tulemusel saadud klastrites

Jooniselt 2.6 on näha kehamassiindeksite jaotus klastrites. Jällegi on näha, et hierarhilise ja k-keskmiste klasterdamise klastrite puhul on keskmised kehamassiindeksid väga sarnased. Samuti on näha, et ka esimese ja kolmanda kvartiili vahe on kõikidel klastritel sarnane. Biklasterdamise klastrite puhul erinevad enamasti klastrite keskmised, kuid on ka väga sarnaseid klastreid. Kõige madalama kehamassiindeksiga klaster on klaster 1, samuti on sellel ka nähtavalt kõige väiksem esimese ja kolmanda kvartiili vahe. Kõige kõrgema keskmise kehamassiindeksiga klaster on klaster 3.

2.3.4 Logistilise regressiooni mudeli rakendamine

Järgnevalt uuriti, kas klastrid annavad informatsiooni inimese suremuse kohta. Uuritavaks tunnuseks võeti *surem*5, mis on binaarne tunnus ning näitab, kas proovi andnud inimene on surnud kuni viie aasta jooksul pärast proovi andmist. Kuna igas veeruklastris olevad klastrid on üksteist välistavad, siis moodustati nendest faktor-

tunnused.

Suremuse tunnuse uurimiseks kasutati logistilist regressiooni ning argumentideks võeti peale klastri tunnuste veel sugu, kehamassiindeks ning vanus. Algsesse mudelis olid kõik tunnused sees. Logistilise regressiooni mudelist oli näha, et ainukesed olulised, olulisuse nivool $\alpha = 0,05$ tunnused olid sugu, vanus, $hklast2$, $bklast4$ ning $bklast5$. Seejärel eemaldati mudelist kõik ebaolulised tunnused ning koostati uus mudel. Antud mudelis osutusid kõik tunnused oluliseks, välja arvatud tunnuse $bklast5$ tase $b17$. Seejärel, koostades mudeli kohta tõepärasuhte testi, on näha, et kõik tunnused tervikuna on mudeli suhtes olulised.

Logistilise regressiooni mudeli tulemused

Tabel 2.2: Logistilise regressiooni mudeli statistiliselt olulised tunnused

Tunnus	Tase	Šansside suhe	95% usaldusintervall	Olulisuse tõenäosus
Hklast2	hic4	-	-	$0,0017290$
	hic5	0,70	(0,50 ; 0,97)	
	hic6	0,65	(0,51 ; 0,83)	
Bklast4	bic11	-	-	$8,032 \cdot 10^{-6}$
	bic12	2,45	(1,61 ; 3,71)	
	bic13	1,67	(1,09 ; 2,55)	
Bklast5	bic14	-	-	$0,0009613$
	bic15	0,51	(0,36 ; 0,71)	
	bic16	0,72	(0,53 ; 0,97)	
	bic17	0,77	(0,50 ; 1,18)	
Sugu	Mees	-	-	$3,942 \cdot 10^{-9}$
	Naine	0,50	(0,40 ; 0,63)	
Vanus		1,09	(1,08 ; 1,10)	$< 2,2 \cdot 10^{-16}$

Tabelis 2.2 on näha kõikide oluliste tunnuste tasemete šansside suhted ning 95% usaldusintervallid. Kuna ainus tase, millel on usaldusintervallis sees 1, on tunnuse $bklast5$ tase $bic17$, siis see on ka ainus tase, mis on statistiliselt ebaoluline. Kuid tun-

nus tervikuna on mudelis ikka tähtis. On näha, et tasemed $hic5$ ja $hic6$ vähendavad viie aasta jooksul suremuse šanssi, kui võrrelda neid tasemega $hic4$. Samuti saab öelda, et tasemed $bic15$ ja $bic16$ vähendavad viie aasta jooksul suremuse šanssi, kui võrrelda neid tasemega $bic14$. Vastupidiselt tasemed $bic12$ ja $bic13$ suurendavad viie aasta jooksul suremuse šanssi, kui võrrelda neid tasemega $bic11$. Mudelist ilmneb, et naistel on väiksem šanss surra viie aasta jooksul. Samuti ka vanuse kasvades suureneb šanss surra viie aasta jooksul.

Mudelis oluliste tunnuste analüüsimine

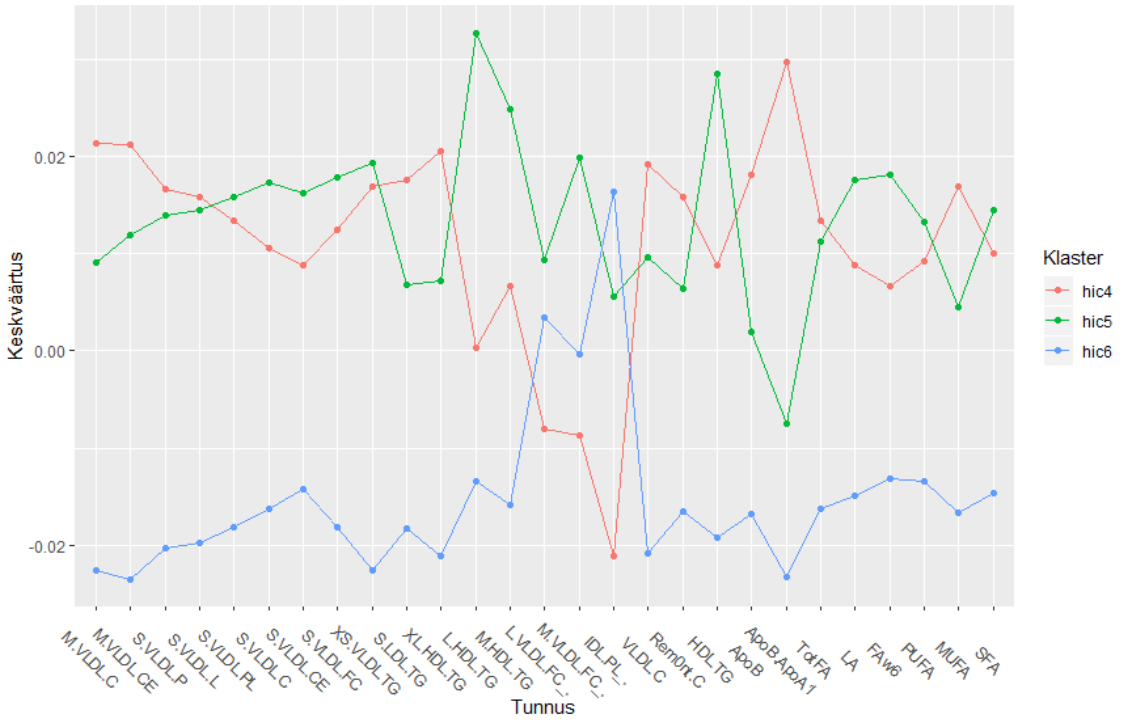
Tabelis (2.3) on välja toodud logistilise regressioni mudelis statistiliselt oluliste veeruklastrite ($hklast2$, $bklast4$ ja $bklast5$) tunnused. Tabelist tuleb välja, et kuigi leidub erinevates veeruklastrites samu tunnuseid, nagu näiteks $M.HDL.FC$ või $S.VLDL.C$, siis enamikud tunnused on klastritel unikaalsed, nagu näiteks LA või Gp.

Järgnevalt leiti iga klastri tunnuste keskväärtused ning kanti need graafikule. Kuna algne andmestik on normeeritud nii, et keskväärtus oleks 0 ning standardhälve 1, siis peaks ka klastrites olema keskmiselt iga tunnus keskväärtusega 0 ning standardhälbega 1.

Joonisel 2.7 on välja toodud veeruklastri $hklast2$ iga tunnuse keskväärtus erinevates klastrites. On näha, et leiduvad mõned tunnused igas klastris, mille keskväärtused on 0, kuid enamasti on need nullist erinevad. Samas erinevad enamike tunnuste klastrisisesed keskväärtused nullist vähem kui $0,02$ kahe võrra. Tegemist on väga väikse erinevusega. Jooniselt samuti ilmneb, et enamasti klastrid $hic4$ ja $hic6$ peegeldavad üksteist. Ehk, kui klastris $hic4$ on mõne tunnuse keskväärtus suur, siis klastris $hic6$ on sama tunnuse keskväärtus vastupidise märgiga umbes sama suur.

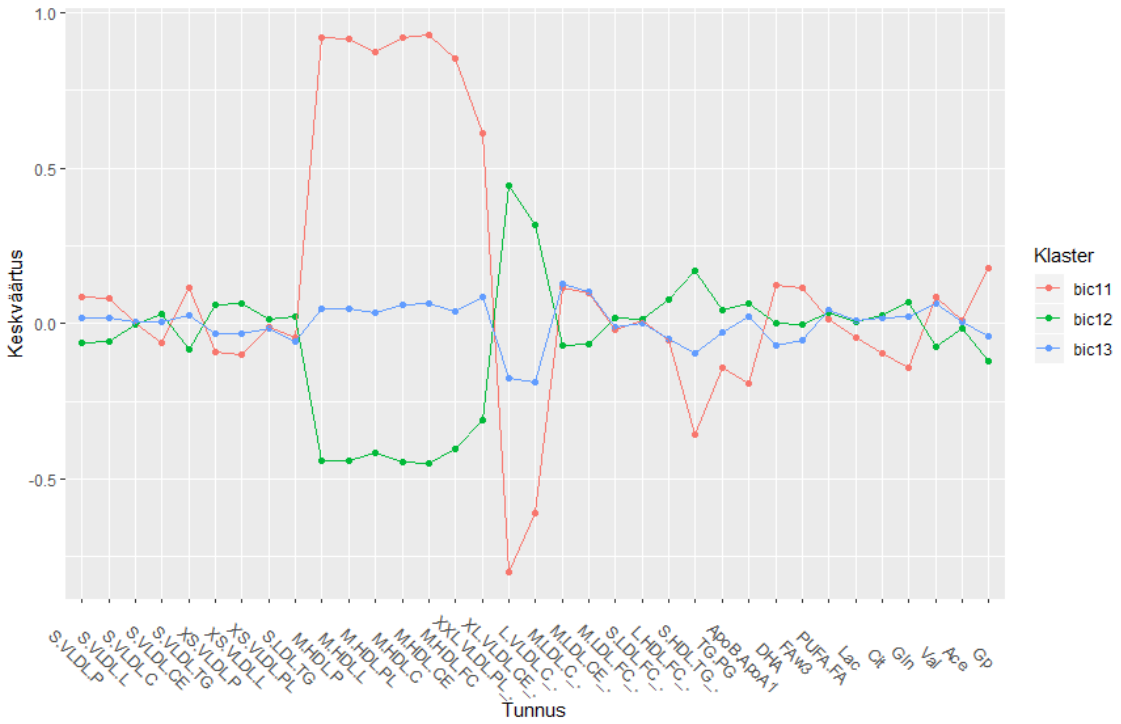
Tabel 2.3: Logistilise regressiooni mudeli statistiliselt oluliste klastrite tunnused

Hklast2		Bklast4		Bklast5	
M.VLDL.C	M.VLDL.CE	S.VLDL.P	S.VLDL.L	IDL.TG	S.HDL.P
S.VLDL.P	S.VLDL.PE	S.VLDL.C	S.VLDL.CE	S.HDL.L	S.HDL.C
S.VLDL.C	XS.VLDL.TG	S.VLDL.TG	XS.VLDL.P	S.HDL.CE	XXL.VLDL.FC
S.LDL.TG	XL.HDL.TG	XS.VLDL.L	XS.VLDL.PL	XL.VLDL.TG	S.VLDL.PL
L.HDL.TG	M.HDL.TG	S.LDL.TG	M.HDL.P	S.VLDL.FC	XS.VLDL.FC
L.VLDL.FC	M.VLDL.FC	M.HDL.L	M.HDL.PL	XS.VLDL.TG	IDL.PL
IDL.PL	VLD.LC	M.HDL.C	M.HDL.CE	IDL.TG	L.LDL.TG
Rem0nt.C	HDL.TG	M.HDL.FC	XXL.VLDL.PL	M.LDL.TG	XL.HDL.PL
ApoB	ApoB.ApoA1	XL.VLDL.CE	L.VLDL.C	L.HDL.PL	M.HDL.PL
TotFA	LA	M.LDL.C	M.LDL.CE	M.HDL.FC	S.HDL.C
FAw6	PUFA	M.LDL.FC	S.HDL.TG	S.HDL.CE	MUFA.FA
MUFA	SFA	TG.PG	ApoB.ApoA1	Crea	
		DHA	FAw3		
		PUFA.FA	Lac		
		Cit	Gln		
		Val	Ace		
		Gp			



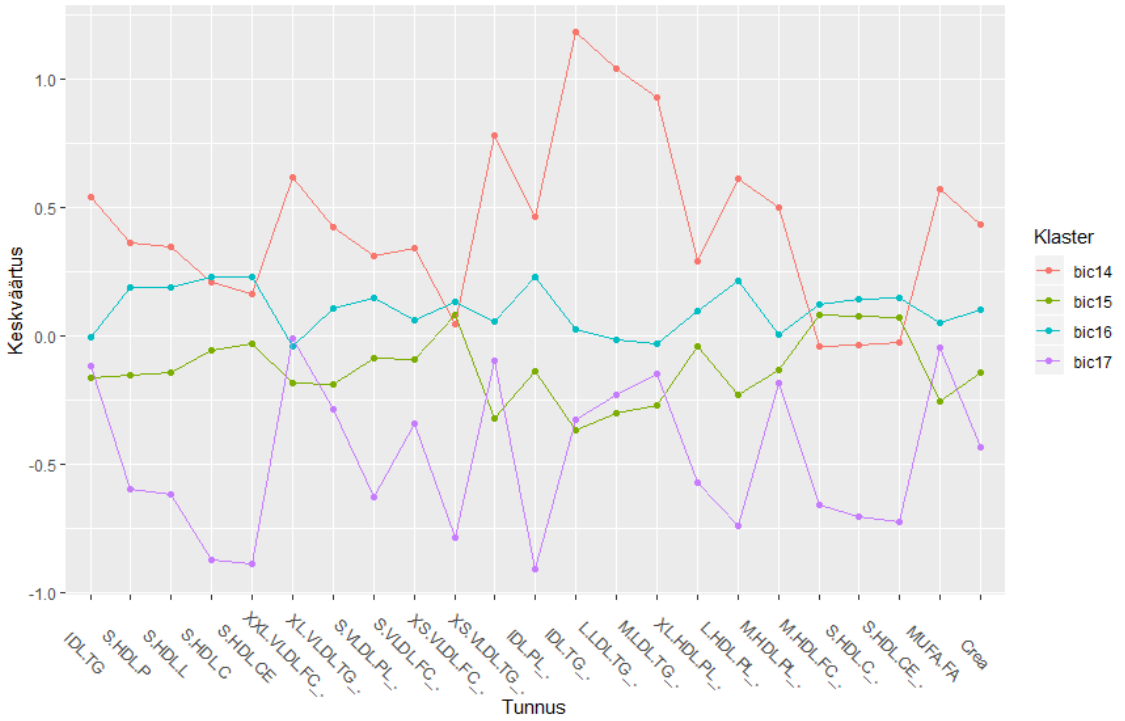
Joonis 2.7: Veeruklastri hklast2 tunnuste keskväärtused normeeritud andmestikus

Joonisel 2.8 on välja toodud veeruklastri *bklast*4 iga tunnuse keskväärtus erinevates klastrites. On näha, et klastritel *bic*11 ja *bic*12 leiduvad tunnused, mille keskväärtus erineb andmestiku keskmisest väga palju. Klastri *bic*11 maksimaalne erinevus on peaaegu 1 ning klastri *bic*12 väärtus on peaaegu $0,5$. On näha ka, et iga klastri enamike tunnuste keskväärtus erineb üldisest keskväärtusest vähemalt $0,1$ võrra. Samuti on näha, et klastrid *bic*11 ja *bic*12 peegeldavad üksteist. Ehk, kui klastris *bic*11 on mõne tunnuse keskväärtus suur, siis klastris *bic*12 on sama tunnuse keskväärtus vastupidise märgiga samuti suur. Tundub, nagu enamasti oleks klastri *bic*11 tunnuse keskväärtus absoluutväärtuselt kaks korda suurem, kui klastri *bic*12 sama tunnuse keskväärtus.



Joonis 2.8: Veeruklastri bklast4 tunnuste keskväärtused normeeritud andmestikus

Joonisel 2.9 on välja toodud veeruklastri *bklast*5 iga tunnuse keskväärtus kõigis neljas erinevas klastris. On näha, et enamasti erinevad klastrite tunnuste keskväärtused nullist vähemalt $0,1$ võrra klastrites *bic*15 ja *bic*16. Klastrite *bic*14 ja *bic*17 puhul erinevad tunnuste keskväärtused nullist enamasti vähemalt $0,25$. Samuti, kui eelnevate klastrite puhul on klastrid peegeldanud, siis veeruklastri *bklast*5 puhul tundub, et klastrite *bic*14 ja *bic*17 tunnuste keskväärtused on nihkes. Ehk iga klastri *bic*14 ja *bic*17 tunnuse keskväärtuste vahe on sarnane iga tunnuse puhul. Samuti tundub, et klastrite *bic*15 ja *bic*16 tunnuste keskväärtused on nihkes. Kõige rohkem erinevad üldisest keskväärtusest klastrid *bic*14 ja *bic*17. Klastri *bic*14 maksimaalne erinevus nullist on peaaegu $1,25$ ning klastri *bic*17 maksimaalne erinevus on peaaegu 1.



Joonis 2.9: Veeruklastri bklast5 tunnuste keskväärtused normeeritud andmestikus

Kokkuvõte

Käesoleva töö eesmärgiks oli võrrelda kahe erineva meetodiga saadud biklastreid ning välja selgitada, kas tulemuseks saadud biklastrid kirjeldavad konkreetse inimese suremust. Esimeseks uuritavaks meetodiks oli hierarhilise ja k-keskmiste klasterdamise kombinatsioon, kus esmalt rakendati andmestikule tunnuste hierarhilist klasterdamist ning seejärel veeruklastrites rakendati vaatlustele k-keskmiste klasterdamist. Teiseks uuritavaks meetodiks oli spektraalne biklasterdamine, mis kasutas maatriksi omaväärtusi ning omavektoreid klastrite leidmiseks.

Mõlema meetodi rakendamine andmestiku peal oli edukas. Hierarhilise klasterdamise tulemusel saadi kaheksa veeruklastrit ning iga veeruklaster jagati k-keskmiste meetodit rakendades veel omakorda kolmeks klastriks. Selle tulemusel saadi kokku 24 klastrit, mis jagunesid kaheksa veeruklastri alla. Spektraalse biklasterdamise tulemusel saadi kokku 259 klastrit, millest valiti välja esimesed 24 klastrit, mis kuulusid omakorda seitsme veeruklastri alla.

Seejärel uuriti, kuidas jagunevad mõlema meetodi klastrites sugu, vanus ning kehamassiindeks. Ilmnes, et nii soo, vanuse kui ka kehamassiindeksi jagunemine hierarhilise ja k-keskmiste klasterdamise klastrites oli ühtlane kõigis klastrites. Spektraalse biklasterdamise klastrites erinesid andmed klastriti. Vanuse puhul olid keskmised väga erinevad ning samuti varieerusid rohkem vahed esimese ja kolmanda kvartiili vahel. Kehamassiindeksi puhul oli samuti erinevusi, kuid need olid väiksemad klastrite vahel, kui vanuse puhul. Sooline jagunemine oli samuti klastrite puhul vägagi erinev.

Logistilise regressiooni mudeliga uuriti, kuidas mõjutavad klastrid inimese šanssi surra järgmise viie aasta jooksul. Selleks koostati binaarne tunnus, mis näitas, kas inimene on viimase viie aasta jooksul surnud. Seejärel koostati mudel, milles tunnusteks olid kõik veeruklastrid, mille faktoriteks olid nende all olevad klastrid. Samuti lisati mudelisse sugu, vanus ning kehamassiindeks. Ainsateks statistiliselt olulisteks tunnusteks jäid kaks spektraalse biklasterdamise veeruklastrit, $bklast4$ ja $bklast5$, ning üks hierarhilise ja k-kesmkiste veeruklaster, $hklast2$, samuti ka vanus ning sugu. Kahel veeruklastril olid kõik tasemed, ehk klastrid, olulised ning kolmandal oli üks tase ebaoluline.

Viimaks uuriti mudelis statistiliselt oluliste klastrite tunnuste keskväärtusi. Oli näha, et hierarhilise ja k-keskmiste klasterdamise veeruklastri $hklast2$ klastrite tunnuste keskväärtused olid lähedal kogu andmestiku keskmisele, maksimaalse erinevusega 0.035. Spektraalse biklasterdamise klastrite tunnuste erinevused andmestiku keskmisest olid suuremad. Veeruklastris $bklast4$ oli näha suuremat tunnuste keskväärtuse varieeruvust andmestiku keskmisest, maksimaalse erinevusega 0.9. Samuti erines ka veeruklastris $bklast5$ tunnuste keskväärtus üldise andmestiku keskväärtusest, maksimaalse erinevusega 1.2.

Seega võib öelda, et uuritud kahest meetodist annab paremaid klastreid spektraalne biklasterdamine. Selle meetodi rakendamisel saadud klastrid erinesid sooliselt, vanuseliselt ning kehamassiindeksi poolest üksteisest palju rohkem kui hierarhilise ja k-keskmiste klasterdamise puhul. Samuti oli näha, et spektraalse biklasterdamise klastrid kirjeldasid rohkem ka inimese suremust. Uurides lõpuks ka klastri tunnuste keskväärtusi oli näha, et need erinesid andmestiku keskväärtustest palju rohkem, kui hierarhilise ja k-keskmiste klasterdamise klastrid.

Kindlasti vajaks antud teema tulevikus lähemalt uurimist, kuna käesoleva töö raames uuriti väga väheseid klastreid ning k-keskmiste klasterdamise puhul oli fikseeritud klastrite arv.

Kasutatud kirjandus

Dodge, Yadolah (2008). *The Concise Encyclopedia of Statistics*. Springer, l. 502–505.

James, Gareth et al. (2013). *An Introduction to Statistical Learning with Applications in R*. 6. väljaanne. Springer, l. 385–399.

Kaiser, Sebastian et al. (2018). *biclust: BiCluster Algorithms*. R package version 2.0.1. URL: `https://CRAN.R-project.org/package=biclust`.

Kluger, Yuval et al. (2003). "Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions". *Genome Research* 10.11, l. 703–716.

Kosub, Sven (2016). *A note on the triangle inequality for the Jaccard distance*.

Käärik, Ene (2013). *Loengukonspekt aines Andmeanalüüs II*, l. 106–111.

Madeira, Sara C. ja Arlindo L. Oliveira (2004). "Biclustering Algorithms for Biological Data Analysis: A Survey". *IEEE Transactions on Computational Biology and Bioinformatics* 2004.2, l. 24–45.

R Core Team (2018). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: `https://www.R-project.org/`.

Wickham, Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. ISBN: 978-3-319-24277-4. URL: `http://ggplot2.org`.

Lihtlitsents lõputöö reprodutseerimiseks ja üldsusele kättesaadavaks tegemiseks

Mina, Villem Lassmann,

1. annan Tartu Ülikoolile tasuta loa (lihtlitsentsi) minu loodud teose „Kahe klasterdamismeetodi võrdlus TÜ Eesti Geenivaramu metaboloomika andmestiku näitel“, mille juhendaja on Krista Fischer, reprodutseerimiseks eesmärgiga seda säilitada, sealhulgas lisada digitaalarhiivi DSpace kuni autoriõiguse kehtivuse lõppemiseni.

2. Annan Tartu Ülikoolile loa teha punktis 1 nimetatud teos üldsusele kättesaadavaks Tartu Ülikooli veebikeskkonna, sealhulgas digitaalarhiivi DSpace kaudu Creative Commonsi litsentsiga CC BY NC ND 3.0, mis lubab autorile viidates teost reprodutseerida, levitada ja üldsusele suunata ning keelab luua tuletatud teost ja kasutada teost ärieesmärgil, kuni autoriõiguse kehtivuse lõppemiseni.

3. Olen teadlik, et punktides 1 ja 2 nimetatud õigused jäävad alles ka autorile.

4. Kinnitan, et lihtlitsentsi andmisega ei riku ma teiste isikute intellektuaalomandi ega isikuandmete kaitse õigusaktidest tulenevaid õigusi.

Villem Lassmann
08.05.2019