VILNIAUS UNIVERSITETAS

MATEMATIKOS IR INFORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ KATEDRA

PRAKTIKOS ATASKAITA

Praktiką atliko: **Dainius Jocas**
(studento vardas, pavardė) (parašas)

Programų sistemos, bakalauras, 4 kursas
(studijų programa, pakopa, kursas)

Praktikos institucija: Vilniaus universiteto matematikos ir informatikos institutas
(organizacijos pavadinimas)

Organizacijos praktikos vadovas: Mokslinis stažuotojas Dr. Juozas Gordevičius
(pareigos, vardas, pavardė)

Organizacijos praktikos vadovo įvertinimas: _____
(įvertinimas, parašas)

Universiteto praktikos vadovas: Dr. Juozas Gordevičius
(mokslo laipsnis, vardas, pavardė)

_____
(parašas)

Ataskaitos įteikimo data _____
Registracijos Nr. _____
Įvertinimas _____
(data, įvertinimas, parašas)

Vilnius, 2012

Turinys

ĮVADAS

Profesinei praktikai atlikti pasirinkau Vilniaus universiteto (VU) Matematikos ir informatikos instituto (MII) sistemų analizės skyrių dėl dviejų priežasčių. Pirma, norėjau pasinaudoti galimybe profesinės praktikos metu tęsti bakalauriniame darbe atliekamą tyrimą. Bakalauriniame darbe nagrinėjama biomedicininių, daug atributų turinčių (daugiamačių) duomenų suskirstymo į pageidaujamas kategorijas pagal vidinę duomenų struktūrą – klasifikavimo – problema.

Antroji mano pasirinkimo profesinę praktiką atlikti MII priežastis buvo ta, kad dirbdamas MII, turėsiu galimybę konsultuotis su daugiamačių duomenų analizės problematiką tiriančiais mokslininkais, nes norint pradėti spręsti bakalauriniame darbe iškeltą problemą reikia ir tais daugiamačiais duomenimis apibūdinamų procesų dalykinės srities, ir duomenų analizės, ir programavimo žinių. MII mokslininkų sukauptas žinių bagažas labai palengvino ir pagreitino mano susipažinimą su nagrinėjama problematika.

Profesinės praktikos tikslas – atlikti informatyvių matų (toliau *matų*) atrinkimo metodų palyginamąją analizę. Siekiant užsibrėžto tikslo profesinės praktikos metu buvo sprendžiami šie uždaviniai:

1. Susipažinti daugiamačių duomenų matų atrinkimo problematika bei moksline literatūra;
2. Suprogramuoti matų atrinkimo metodus: *Fisher*, *Relief*, asimetrinį priklausomybės koeficientą (angl. *asymetric dependency coefficient*, ADC), atraminių vektorių klasifikatoriumi (SVM) grįstą absoliučių svorių metodą (angl. *absolute weight support vector machines, AW-SVM*), svoriais grįstą multikriterinį suliejimą (angl. *score-based multicriterion fusion*), reitingais grįstą multikriterinį suliejimą (angl. *ranking-based multicriterion fusion*), svoriais ir reitingais grįstą multikriterinį rekursyvų matų eliminavimą[YM11b], konsensuso grupėmis grįsto stabilių matų atrinkimo metodą[LYD09] (angl. *consensous group stable feature selection*)
3. Palyginti suprogramuotų matų atrinkimo metodų skaičiavimo laiką, klasifikavimo tikslumą bei stabilumą.

Praktinės veiklos planas buvo sudarytas iš dviejų dalių:

1. suprogramuoti pasirinktus matų atrinkimo algoritmus;

2. palyginti suprogramuotus algoritmus skaičiavimo laiko, klasifikavimo tikslumo bei matų atrinkimo stabilumo atžvilgiais tarpusavyje.

Du penktadaliai numatyto profesinės praktikos laiko buvo skirta matų atrinkimo algoritmų programavimui, dar du penktadaliai buvo numatyti algoritmų palyginimui, o likęs laikas susipažinimui su dalykinės srities literatūra bei dalyvavimui MII rengiamuose seminaruose.

Profesinę praktiką pradėjau 2012 metų vasario 6 dieną. Ji truko 11 savaičių ir baigėsi 2012 metų balandžio 20 dieną. Ilgiau nei planuota užtruko matų atrinkimo metodų programavimo darbai, todėl teko sumažinti matų atrinkimo algoritmų lyginamųjų eksperimentų apimtis.

Likusi praktikos ataskaitos dalis yra organizuota taip: skyriuje 1 glaustai aprašau įstaigą, kurioje atlikau profesinę praktiką; skyriuje 2 aprašau praktikos veiklas ir praktikos užduotis; skyriuje 3 aprašau profesinės praktikos darbo rezultatus bei padarytas išvadas, praktikos darbo privalumus bei trūkumus, įgytas žinias bei patirtis, taip pat pateikiu pasiūlymų, kaip galima būtų geriau organizuoti darbo ir valdymo procesus praktikos atlikimo vietoje ir mokymą Vilniaus universitete.

1. VU MII

Vilniaus universiteto matematikos ir informatikos institutas (MII) nuo 2010 m. yra Vilniaus universiteto padalinys. Jame vykdomi tyrimai matematikos ir informatikos srityse. Instituto įkūrimo data laikoma 1965 m. spalio 1d., kai buvo panaikintas Lietuvos mokslų akademijos Fizikos ir technikos institutas ir įkurti trys nauji institutai, tarp kurių buvo Fizikos ir matematikos institutas, kuris laikomas MII pirmtaku.

Pagrindinė instituto veikla - moksliniai tyrimai ir eksperimentinė plėtra. Kitos veiklos sritys yra: mokslininkų ugdymas (doktorantūros studijos) – MII suteikta teisė ruošti matematikos, informatikos ir informatikos inžinerijos sričių mokslininkus; mokslo organizacinė veikla - konferencijos, seminarai, parodos, mokslinių knygų redagavimas; leidyba; mokymas, moksleivių ugdymas, švietimas. Mokslinė veikla sukoncentruota 12-oje mokslinių padalinių. Institute yra 5 matematikos krypties padaliniai, 7 informatikos bei informatikos inžinerijos padaliniai:

1. Atpažinimo procesų skyrius;
2. Atsitiktinių procesų skyrius;
3. Informatikos metodologijos skyrius;
4. Kompiuterinių tinklų laboratorija;
5. Matematinės logikos sektorius;
6. Programų sistemų inžinerijos skyrius;
7. Sistemų analizės skyrius (SAS);
8. SAS optimizavimo sektorius;
9. SAS operacijų tyrimo sektorius;
10. Skaičiavimo metodų skyrius (SMS);
11. SMS diferencialinių lygčių sektorius;
12. Tikimybių teorijos ir statistikos skyrius;

MII organizuoja moksleivių ugdymą: veikia jaunųjų programuotojų neakivaizdinė mokykla, rengiamos lietuvos moksleivių informatikos ir matematikos olimpiados, rengiamas

informacinių technologijų konkursas „Bebras“. MII yra vienas iš Lietuvos jaunųjų matematikų mokyklos steigėjų, jaunųjų matematikų konkurso „Kengūra“ rengėjas. Taip pat MII prisideda prie kompiuterijos naudotojų švietimo ir mokymo: dirba informatikos terminijos komisija, multimedijos centras humanitarams, palaikomas tinklalapis apie lietuviškų rašmenų naudojimą elektroninio pašto laiškuose.

MII leidybos skyriuje leidžiami periodiniai leidiniai: „Informatica“, „Informatics in Education“, „Lithuanian Mathematical Journal“, „Lietuvos matematikos rinkinys. LMD darbai“, „Mathematical Modelling and Analysis“, „Nonlinear Analysis. Modelling and Control“, „Olympiads in Informatics“. MII mokslininkai taip pat yra išleidę mokslinių bei mokslo populiarinimo knygų lietuvių ir anglų kalbomis, mokymo priemonių, interaktyvių kompaktinių diskų bei sukūrę įvairių internetinių informacinių sistemų (pvz. enciklopedinis kompiuterijos terminų žodynas).

MII man, kaip ir kiekvienam darbuotojui, parūpino: darbo vietą, prisijungimą prie vietinio tinklo, galimybę naudotis skaičiavimo ištekliais, galimybę su nuolaida pietauti vietinėje valgykloje. MII darbuotojai buvo draugiški, todėl apsipratau MII labai greitai. Jau nuo pat pirmosios profesinės praktikos dienos galėjau pradėti spręsti užsibrėžtus uždavinius.

2. PROFESINĖS PRAKTIKOS VEIKLA

Profesinę praktiką sudarė trys užduotys:

1. Susipažinti su matų atrinkimo daugiamačiuose duomenyse problematika bei moksline literatūra;
2. Suprogramuoti matų atrinkimo metodus;
3. Ištirti matų atrinkimo metodų savybes.

Toliau aprašau kiekvieną užduotį atskirai.

2.1. Matų atrinkimas daugiamačių duomenų klasifikavimui

Savo bakalauriniame darbe nagrinėju biomedicinoje kaupiamų genetinių daugiamačių duomenų analizės specifiką. Šie duomenys yra specifiški tuo, kad jie turi šimtus kartų daugiau matų nei mėginių. Kadangi mėginio gavimo kaina yra aukšta, turimas mažas mėginių skaičius. Biomedicininių duomenų analizę apsunkina ir tai, kad matavimai, kuriais tie duomenys gaunami, yra triukšmingi. Triukšmas matavimo metu atsiranda dėl cheminių reakcijų netikslumo, tiriamo organizmo sudėtingumo. Kai duomenys yra triukšmingi ir didėja juos apibūdinančių matų skaičius, didėja tikimybė duomenyse rasti atsitiktinių priklausomybių. Tai yra pagrindinė priežastis, kodėl biomedicininių duomenų analizės procesas yra sudėtingas.

Biomedicininių duomenų klasifikavimo užduotis yra atskirti sveikų pacientų mėginius nuo sergančiųjų. Klasifikavimu siekiama nustatyti, kurie matai, veikdami drauge, geriausiai paaiškina skirtumą tarp ligos paveiktų ir sveikų mėginių. Labiausiai ligą paaiškinančių matų nustatymas galėtų palengvinti tiriamų ligų diagnozės ir gydymo metodų kūrimą. Klasifikavimu yra vadinamas duomenų analizės procesas, kurio metu yra sukonstruojama funkcija, atskirianti duomenis į grupes pagal jų matus [Fis36]. Sukonstruotos funkcijos yra vadinamos klasifikatoriais, o jų konstravimo algoritmai – klasifikavimo algoritmais. Klasifikatoriai paruošiami naudojant turimus mėginius – treniravimosi duomenis – ir informaciją apie jų būklę (sveikas ar sergantis). Klasifikatoriaus ruošimo procesas yra vadinamas apmokymu. Klasifikatoriai paprastai naudojami nustatant naujų, dar nematytų, mėginių būklę.

Dėl „daugiamatiškumo prakeiksmo“ (angl. *the curse of dimentionality*) – didėjant ma-

tų kiekiui mėginiai pasidaro panašūs, todėl bandymas juos klasifikuoti tolygus spėliojimui [Bel66]. Biomediciniinių duomenų kontekste galima daryti prielaidą, kad ne visi matai yra susiję su tiriama problema, pvz. gaubtinės žarnos vėžiu, dėl to, kad duomenys yra daugiamačiai. Paprastai nagrinėjamai problemai svarbus yra mažas, palyginus su visu, matų kiekis. Todėl biomediciniinių duomenų daugiamatiškumui sumažinti yra naudojami informatyviausių matų atrinkimo metodai [GE03] (angl. *feature selection*). Pagal tai, kaip susiję su klasifikatoriumi, matų atrinkimo metodai skirstomi į tris kategorijas [SAVdP08]: filtruojantys (angl. *filter*), prisitaikantys (angl. *wrapper*) ir įterptiniai (angl. *embedded*) metodai. Filtruojančiais metodais pirmiausia yra atrenkami informatyviausi matai, o tada apmokomas klasifikatorius. Prisitaikančiųjų metodų atveju, pirma, apmokomas klasifikatorius su visais matais, antra, parenkamas matų poaibis ir apmokomas klasifikatorius, tada po daugkartinio matų aibių įvertinimo pagal klasifikavimo rezultatus yra nusprendžiama, kuris matų poaibis yra labiausiai tinkamas klasifikavimui. Įterptinių metodų atveju matų atrinkimo procesas yra neatsiejamas nuo klasifikavimo proceso – pats klasifikatorius įvertina matus.

Matų atrinkimas yra svarbi biomediciniinių duomenų apdorojimo (angl. *preprocessing*) etapo dalis. Naudojant matų atrinkimo metodus, galima kovoti su daugiamatiškumo prakeiksmu matų skaičių priartinant prie mėginių skaičiaus. Todėl svarbu yra pasirinkti geriausiai tinkančią matų atrinkimo strategiją. Kadangi ir pačių matų atrinkimo metodų veikimas priklauso nuo konkrečių duomenų, tai metodo pasirinkimas yra sudėtinga užduotis.

Naudodami matų atrinkimo metodus, biomediciniinius duomenis tiriantys mokslininkai susiduria su atrinktųjų matų aibės stabilumo problema – atrenkant matus pagal kitą mėginių poaibį, gaunamas kitas matų poaibis. Matų atrinkimo nestabilumas yra sąlygotas šių veiksnių:

1. Duomenys yra triukšmingi ir kai kurie matai gali būti palaikyti informatyviais vien dėl atsitiktinių priežasčių;

2. Daugiamačiuose duomenyse tikėtina, kad dalis matų koreliuoja tarpusavyje, todėl, kuris iš koreliuojančių matų bus pasirinktas, priklauso nuo to, kuriuos mėginius pasirinksime klasifikatoriaus apmokymui;

3. Kiekvienas matų atrinkimo algoritmas daro skirtingas prielaidas apie tai, kurie matai yra informatyvūs.

Galime daryti išvadą, kad skirtingi metodai tiems patiems duomenims gali atrinkti skirtingus matus. Taip pat, suskaidžius turimus duomenis į atskiras persidengiančias aibes ir atrinkus

tą patį kiekį matų tuo pačiu metodu, gaunamos skirtingos matų aibės. Be to, kuo triukšmingesni duomenys, kuo mažiau turima mėginių ir kuo daugiau yra matų, tuo ryškesnė yra ši problema [LYD09].

Matų atrinkimo stabilumo problemą pirma siūlyta spręsti surandant matų grupių tankio centrus ir naudoti matus, kurie artimiausi tiems centrams [YDL08]. Pasiūlytas grupių tankių algoritmas užtrunka $O(\lambda n^2 m)$ laiko, kur n yra matų kiekis, o m – mėginių skaičius. Vėliau Loscalzo ir kt. pasiūlė mokymo duomenis skaidyti poaibiais ir kiekviename poaibyje ieškoti tankių matų grupių, o tada imti sprendimą balsavimo principu [LYD09]. Nors šie metodai siūlo stabilų matų atrinkimą, tačiau jų panaudojamumą daugiamačiuose duomenyse riboja skaičiavimo sudėtingumas.

Yang ir Mao pasiūlė reitinguoti matus remiantis keletos matų atrinkimo metodų rezultatais [YM11a]. Galutinis matų reitingų sąrašas gaunamas, kai po kiekvieno matų atrinkimo yra išmetama vienas žemiausią reitingą turintis matas iš matų aibės, ir matų atrinkimas yra kartojamas tol, kol nebelieka matų. Tačiau ši matų atrinkimo strategija yra ribota, nes matų atrinkimo metodų kiekis yra ribotas ir skirtingų metodų dažnai negalima atlikti išskirstytų skaičiavimų aplinkoje. Tai riboja šio metodo pritaikomumą daugiamačių duomenų analizėje.

Praktikos metų išstudijavau esamus stabilių matų atrinkimo metodus nustačiau, kad jie tik šiek tiek padidina matų atrinkimo stabilumą, bet problemos iš esmės neišsprendžia.

2.2. Suprogramuoti matų atrinkimo algoritmai

Profesinės praktikos metu suprogramavau populiariausius matų atrinkimo metodus. Taip pat programavau ir matų atrinkimo stabilumą didinančius metodus. Toliau šiame poskyryje aprašau šiuos metodus.

2.2.1. *Fisher* įvertis

Fisher įvertis vertina individualius matus pagal matų klasių atskiriamąją galią. Mato įvertis yra sudarytas iš tarpklasinio skirtumo santykio su vidiniu klasės pasiskirstymu:

$$FR(j) = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2}, \tag{1}$$

kur, j – yra mato indeksas, μ_{jc} – mato j reikšmių vidurkis klasėje c, σ^2_{jc} – mato j reikšmių standartinis nuokrypis klasėje c, kur $c = 1, 2$. Kuo didesnis yra *Fisher* įvertis, tuo geriau ts matas atskiria klases. Nors ir paprastas, šis metodas neįvertina matų tarpusavio sąveikų.

2.2.2. *Relief* metodas

Relief metodas iteratyviai skaičiuoja matų „susietumą“. Pradžioje „susietumas“ visiems matams yra lygus nuliui. Kiekvienoje iteracijoje atsitiktinai pasirenkamas mėginys iš mėginių aibės, surandami artimiausi kaimynai iš tos pačios ir kitos grupių, ir atnaujinamos visų matų „susietumo“ reikšmės. Dėl atsitiktinumo faktoriaus klasifikavimo ir matų atrinkimo stabilumo rezultatai naudojant šį metodą varijuoja. Mato įvertis yra vidurkis visų objektų atstumų skirtumų iki artimiausių kaimynų iš kitos ir tos pačios klasių:

$$W(j) = W(j) - \frac{diff(j, x, x_H)}{n} + \frac{diff(i, x, x_M)}{n}, \tag{2}$$

kur $W(j)$ – j-ojo mato „susietumo“ įvertis, n – mėginių aibės dydis, x – atsitiktinai pasirinktas mėginys, x_H - artimiausias x kaimynas iš tos pačios grupės (angl. *nearest-Hit*), x_M – artimiausias x kaimynas iš kitos grupės (angl. *nearest-Miss*), $diff(j, x, x')$ – j-ojo mato reikšmių skirtumas tarp atsitiktinai pasirinkto objekto x ir atitinkamo jo kaimyno, kur skirtumą į intervalą $[0, 1]$ normalizuojanti funkcija yra:

$$diff(j, x, x') = \frac{|x_j - x'_j|}{x_{j_{max}} - x_{i_{min}}}, \tag{3}$$

kur $x_{j_{max}}$ ir $x_{j_{min}}$ yra maksimali ir minimali j-ojo matų reikšmės. „Susietumo“ reikšmių atnaujinimas yra vykdomas n kartų ir kuo didesnė galutinė reikšmė, tuo svarbesnis matas. Šis algoritmas atsižvelgia į matų tarpusavio priklausomybes, nes mėginio artimiausias kaimynas yra ieškomas pagal visus mėginį apibūdinančius matus. Aprašyta algoritmo versija yra skirta dviejų klasių atvejui, tačiau yra ir multiklasinis algoritmo variantas [RSK03].

2.2.3. Asimetrinis priklausomybės koeficientas

Asimetrinis priklausomybės koeficientas (angl. *asymetric dependency coefficient*, ADC) yra matų reitingavimo motodas, kuris matuoja mėginio grupės tikimybinę priklausomybę j-

ajam matui, naudodamas informacijos prieaugį (angl. *information gain*) [Ken83]:

$$ADC(Y, j) = \frac{MI(Y, X_j)}{H(Y)}, \tag{4}$$

kur $H(Y)$ – klasės Y entropija (angl. *entropy*), o $MI(Y, X_j)$ – yra tarpusavio informacija [Sha01] (angl. *mutual information*) tarp mėginio grupės Y ir j-ojo mato.

$$H(Y) = -\sum_{y} p(Y = y) log p(Y = y), \tag{5}$$

$$H(X_j) = -\sum_{x} p(X_j = x) log p(X_j = x), \tag{6}$$

$$MI(Y, X_j) = H(Y) + H(X_j) - H(Y, X_j), \tag{7}$$

$$H(Y, X_j) = -\sum_{y, x_j} p(y, x_j) log p(y, x_j), \tag{8}$$

Kuo didesni ADC įverčiai, tuo matas yra svarbesnis, nes turi daugiau informacijos apie mėginio priklausomybę grupei.

2.2.4. Absoliučių svorių SVM

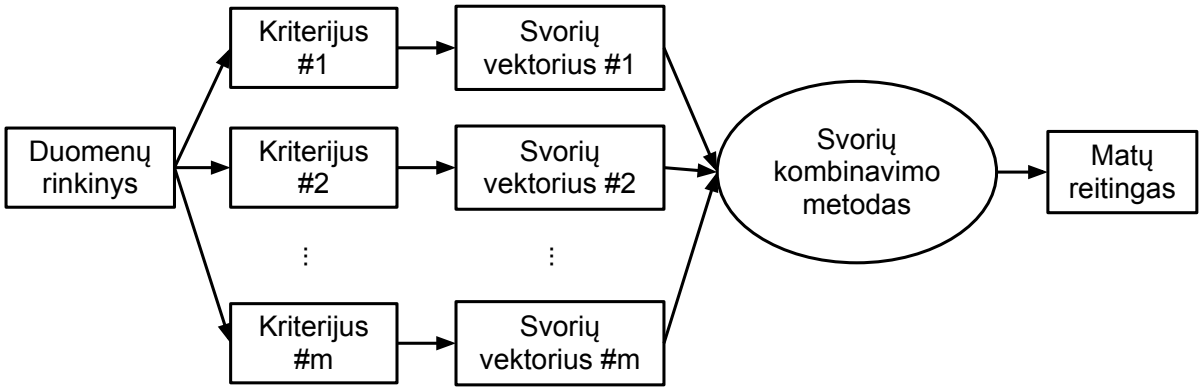
Atraminių vektorių klasifikatorius (SVM) yra vienas populiariausių klasifikavimo algortimų, nes jis gerai susidoroja su daugiamačiais duomenimis [GWBV02]. Yra keletas bazinių SVM variantų [Vap00], bet šiame darbe naudosime tiesinį SVM, nes jis demonstruoja gerus rezultatus analizuojant genų ekspresijos duomenimis. Tiesinis SVM yra hiperplokštuma apibrėžta kaip:

$$\sum_{j=1}^{p} w_j x_j + b_0 = 0, \tag{9}$$

kur p – matų kiekis, w_j – j-ojo mato svoris, x_j – j-ojo mato kintamasis, b_0 – konstanta. Mato absoliutus svoris w_j gali būti panaudotas matų reitingavimui. Svorį reikia imti absoliutaus dydžio, nes neigiamas svoris implikuoja priklausomybę vienai grupei, o teigiamas kitai grupei. Pastebėtina, kad svorių nustatymas yra atliekamas tik vieną kartą (SVM-RFE matų atrinkimo metodas svorius matams nustato daug kartų).

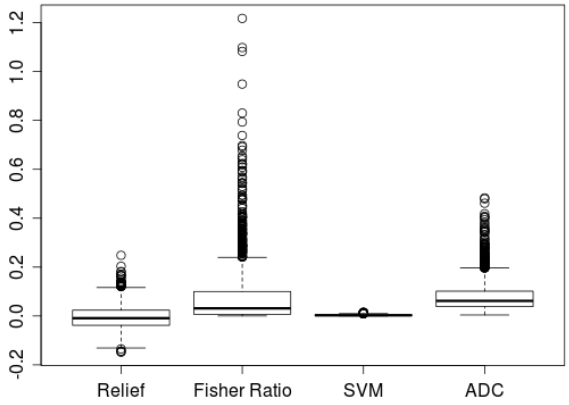
2.2.5. Svoriais grįstas multikriterinis suliejimas

Svoriais grįsto multikriterinio matų atrinkimo suliejimo pagal svorius algoritmo pirmajame žingsnyje kiekvienas bazinis metodas priskiria duomenų rinkinio matams svorius,

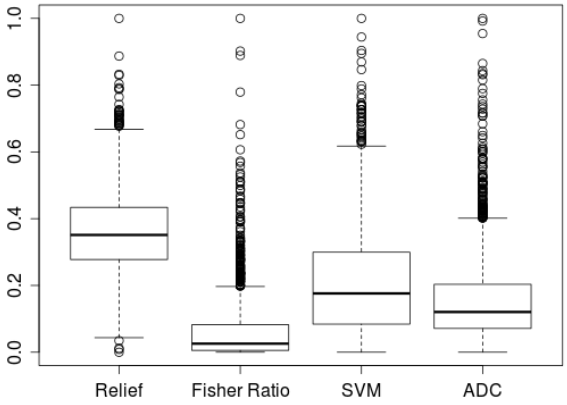


1 pav.: Svoriais grįstas multikriterinis suliejimas.

tada tie svoriai yra kombinuojami į vieną sutarties (angl. *consensus*) svorių vektorių, kurio pagrindu yra gaunami matų reitingai. Algoritmas yra pavaizduotas 1 pav.



2 pav.: Pavienių matų atrinkimo metodų nenormalizuotas svorių pasiskirstymas.

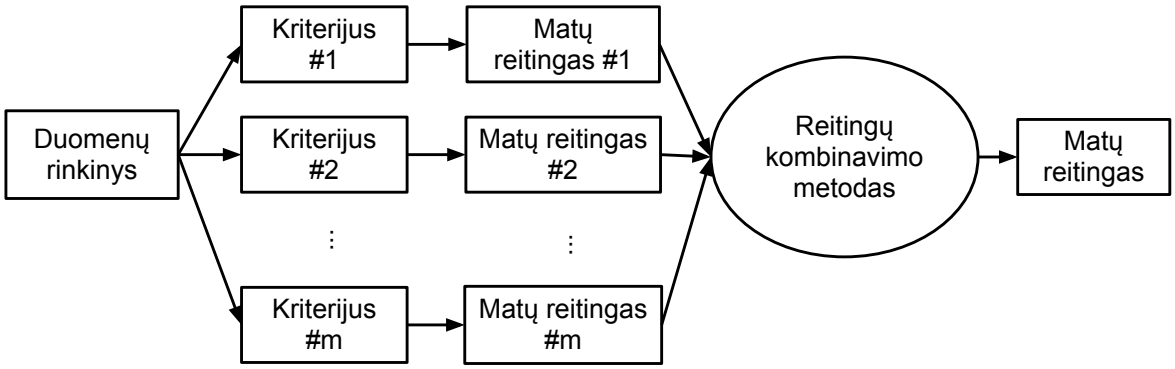


3 pav.: Pavienių matų atrinkimo metodų normalizuotas svorių pasiskirstymas.

Suliejant svorius svarbu yra užtikrinti, kad svoriai, gauti naudojant skirtingus bazinius kriterijus, būtų palyginami. Todėl svorių normalizavimas turi būti atliekamas prieš svorių kombinavimą. Kitu atveju matų įvertinimai bus nepalyginami. Paveikslėlyje 2 pav. nenormalizuotų pavienių matų vertinimo metodų skiriasi netgi suteiktų svorių intervalai. Paveikslėlyje 3 pav. matome, kad net ir normalizavus svorius gana stipriai skiriasi svorių kvartiliai – tai reikia atkreipti dėmesį interpretuojant galutinius matų vertinimo rezultatus. Šiame darbe svoriai yra normalizuoti intervale $[0, 1]$ pagal formulę:

$$u'_i = \frac{u_i - u_{i_{min}}}{u_{i_{max}} - u_{i_{min}}}, \tag{10}$$

kur u_i - matų svorių vektorius pagal i kriterijų, $u_{i_{min}}$ - minimali u_i svorių vektoriaus reikšmė,



4 pav.: Reitingais grįstas multikriterinis suliejimas.

$u_{i_{max}}$ - maksimali u_i svorių vektoriaus reikšmė, u'_i - normalizuotų svorių vektorius.

Sutarties svorių vektorius u yra vidurkis normalizuotų svorių vektorių:

$$u = \frac{1}{m} \sum_{i=1}^{m} u'_i, \tag{11}$$

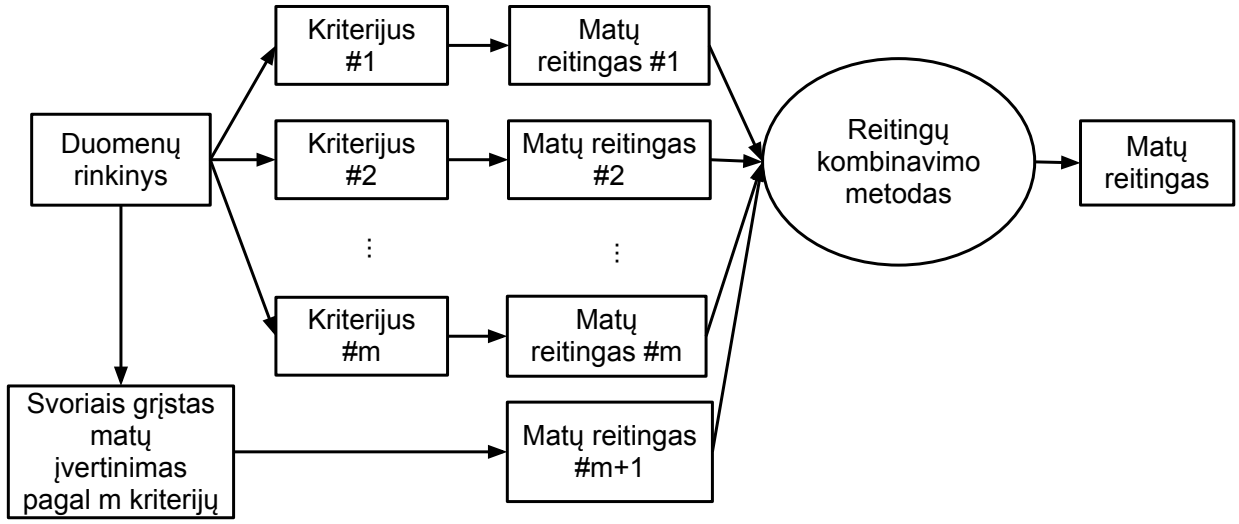
kur m yra bazinių kriterijų skaičius. Reikia paminėti, kad didesnė svorio reikšmė reiškia, kad dimensija yra reikšmingesnė klasifikavimui.

2.2.6. Reitingais grįstas multikriterinis suliejimas

Reitingais grįsto multikriterinio suliejimo pagal reitingus metodas gauna mėginių aibę aprašančių matų reitingą, pagal keletą bazinių matų reitingavimo kriterijų. Algoritmo pirmajame žingsnyje keletas matų atrinkimo kriterijų grąžina matų reitingus, paskui tie reitingai yra kombinuojami į vieną bendra matų reitingą. Algoritmas yra pavaizduotas 4 pav. Suliejimo pagal reitingus metodas nereikalauja matų atrinkimo metodų rezultatų normalizavimo, nes galima dimensijoms priskirtus reitingus iškart kombinuoti. Skirtingai nei suliejimo pagal svorius algoritme, baziniai matų atrinkimo kriterijai turi gražinti matų reitingus, o ne svorius.

Matų reitingų kombinavimui yra keletas metodų[DKNS01], tačiau paprastumo dėlei šiame darbe naudosiu Borda balsavimą[1] (angl. *Borda count*). Tarkime, kad turime m basuotojų ir p kandidatų aibę. Tada Borda balsavimo metodas kiekvienam i-ajam balsuotojui sukuria balsų vektorių v_i tokiu būdu: geriausiai įvertintam kandidatui suteikiama p taškų, antrajam kandidatui $p-1$, ir t.t. Galutiniai taškai yra gaunami sudedant visų balsuotojų

[1]Dar žinomas kaip „Pažymių metodas“. Jis buvo pasiūlytas prancūzų matematiko ir fiziko Jean-Charles de Borda 1770 metais.



5 pav.: Svoriais ir reitingais grįstas multikriterinis suliejimas.

taškus

$$v = \sum_{i=1}^{m} v_i, \tag{12}$$

kur v yra suminių taškų vektorius, o iš jo galime gauti ir galutinius matų reitingus.

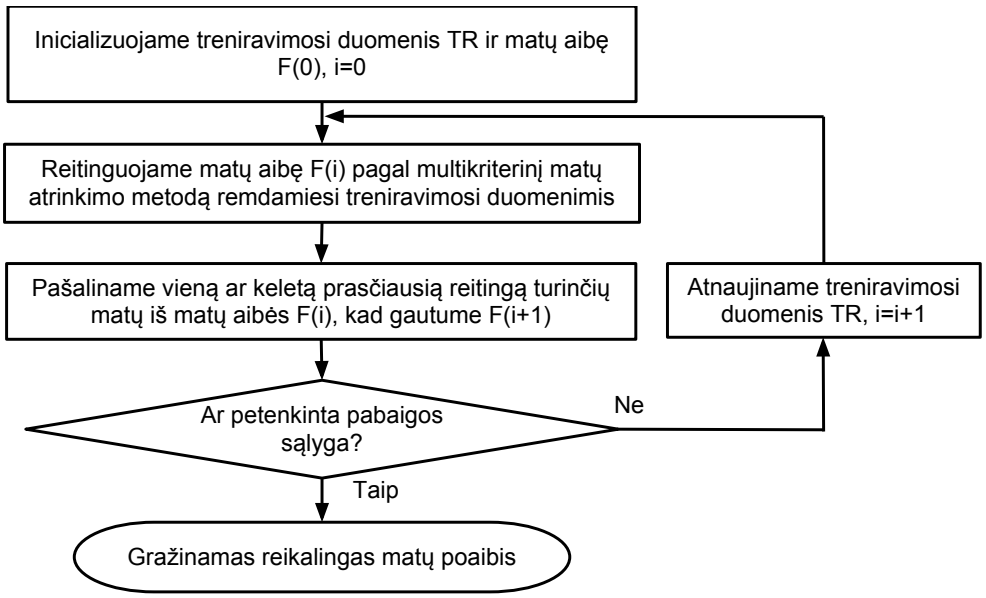
2.2.7. Svoriais ir reitingais grįstas multikriterinis suliejimas

Svoriais ir reitingais grįsto multikriterinio suliejimo metodas nuo reitingais grįsto multikriterinio suliejimo metodo skiriasi tuo, kad kaip dar vienas reitingas yra panaudojamas svoriais grįsto multikriterinio matų atrinkimo metu gautas reitingas. Multikriterinio matų įverčių ir pagal svorius, ir pagal reitingus metodas vyksta trimis žingsniais:

1. Gauname matų reitingus pagal m pavienių matų atrinkimo motodų;
2. Suliejame matų įverčius pagal svorius, taip gauname vieną matų reitingą;
3. Reitinguojame matus pagal visus turimus $m + 1$ pavienius reitingus.

Algoritmas yra pavaizduotas 5 pav.

Kadangi yra suliejami keli mažai koreliuojantys matų reitingavimo metodai, yra pasiekiamas didesnis matų atrinkimo stabilumas, kai varijuoja treniravimosi duomenų poaibis (angl. *subsampling*) [YM11a].



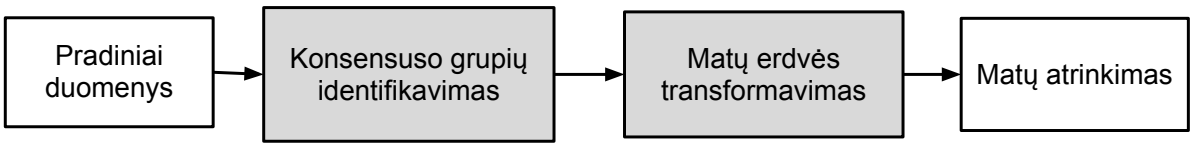
6 pav.: Multikriterinio rekursyvaus matų eliminavimo algoritmas.

2.2.8. Multikriterinis rekursyvus matų eliminavimas

Jei matų atrinkimo tikslas yra pagerinti klasifikavimo rezultatus, tai taikymas multikriterinių matų atrinkimo metodų nebūtinai duos pageidaujamą rezultatą, nes yra pastebėta, kad vien matų reitingavimas nebūtinai suranda geriausią matų poaibį. Tam, kad būtų surastas geriausias matų poaibis reikia kombinuoti multikriterinį matų reitingavimą su paieškos strategija. Rekursyvus matų eliminavimas yra dažnai naudojama paieškos strategija matų atrinkimui. Todėl yra kombinuojamas multikriterinis matų reitingavimas ir rekursyvus matų eliminavimas.

Multikriterinis rekursyvus matų eliminavimas[YM11a] susideda iš dviejų dalių: keletos matų atrinkimo kriterijų suliejimo ir pagal svorius, ir pagal reitingus, ir rekursyvaus matų eliminavimo. Algoritmas pavaizduotas 6 pav. Standartinis rekursyvus matų eliminavimas, kai vienos iteracijos metu yra eliminuojamas vienas matas, padidina algoritmo sudėtingumą. Todėl genų ekspresijos duomenims prasmingiau yra eliminuoti keletą matų kiekvienoje iteracijoje.

Nors SVM-RFE (angl. *Support Vector Machines – Recursive Feature Elimination*) matų atrinkimo algoritmas ir yra labai populiarus, tačiau yra žinoma, kad jam trūksta stabilumo [GWBV02]. Todėl kombinuodami didesnį stabilumą turintį multikriterinį matų atrinkimą su rekursyvaus matų eliminavimo paieškos strategija, gauname stabilesnį matų atrinkimo algoritmą.



7 pav.: Konsensuso grupėmis grįstas stabilių matų atrinkimas.

2.3. Konsensuso grupėmis grįstas stabilių matų atrinkimo metodas

Konsensuso grupėmis grįstas stabilių matų atrinkimo metodas(angl. *Consensus Group Stable feature selection*, CGS), pirma, identifikuoja panašių matų grupes, antra, pagal surastas grupes transformuoja matų aibę, trečia, transformuotoje matų aibėje atlieka matų atrinkimą [LYD09]. Schematiškai šis algoritmas pavaizduotas 7 pav.

CGS metodo pagrindinė dalis yra panašių matų identifikavimas. Šio uždavinio sprendimui naudojamas *Dense Group Finder* (DGF) algoritmas. DGF aprašytas algoritme nr. 1. CGS algoritme agal matai pagal DGF algoritmą yra sugrupuojami keletą kartų. Po pakartotinio grupavimo yra ieškoma stabilių grupių – jei matas buvo sugrupuotas į konkrečią grupę daugiau nei pusėje grupavimų, tai matas ir priklausys tai konsensuso grupei. Matų aibės transformavimas vyksta iš kiekvienos konsensuso grupės išrenkant reprezentatyviausią matą – konkretų matą esantį arčiausiai konsensuso grupės vidurkio. Išrinktieji reprezentatyviausieji matai ir sudaro transformuotą matų aibę. Transformuotoje matų aibėje vykdomas matų antrinkimas koriuo nors matų atrinkimo metodu Φ, pavyzdžiui, *Relief* matų atrinkimo metodu.

Algoritmas nr. 1 DGF – *Dense Group Finder*

Įeitis: duomenys $D = \{x_i\}_{i=1}^n$, branduolio plotis h
Išeitis: tankios matų grupės $G_1, G_1, ..., G_L$
for $i = 1$ **to** n **do**
  Inicializuojame $j = 1, y_{i,j} = x_i$
  repeat
    Suskaičiuoti tankio centrą $y_{i,j+1}$ pagal (13)
  until konverguoja
  Nustatyti tankio centrą $y_{i,c} = y_{i,j+1}$ (Nustatyti piką p_i kaip $y_{i,c}$)
  Sulieti piką p_i su artimiausiais pikais, jei atstumai tarp jų $< h$
end for
Iš kiekvieno unikalaus piko p_r, pridėkime x_i į G_r, jei $||p_r - x_i|| < h$

$$y_{i,j+1} = \frac{\sum_{i=1}^{n} x_i K(\frac{y_j - x_i}{h})}{\sum_{i=1}^{n} K(\frac{y_j - x_i}{h})} j = 1, 2, ... \quad (13)$$

kur $K(x)$ – *kernel* funkcija, h – *kernel* plotis, y – tankio centras.

Algoritmas nr. 2 Konsensuso grupėmis grįstas stabilių matų atrinkimas

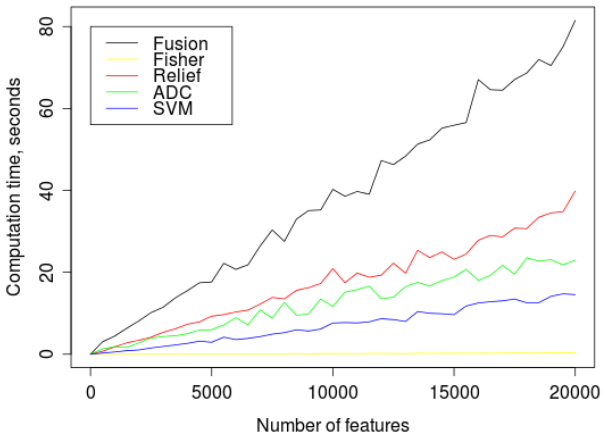
Įeitis: mėginių aibė D, iteracijų skaičius t, matų atrinkimo metodas Φ
Išeitis: atrinktos konsensuso matų grupės $CG_1, CG_1, ..., CG_k$
// Konsensuso grupių identifikavimas
for $i = 1$ **to** n **do**
    Parinkti mėginių poaibį D_i iš D
    Gauti panašių matų grupes pagal $DGF(D_i, h)$
end for
for kiekvienai matų porai X_i ir $X_j \in D$ **do**
    Nustatyti $W_{i,j}$ = dažnis, kai X_i ir X_j yra toje pačioje grupėje $/t$
end for
Sudaryti konsensuso grupes $CG_1, CG_1, ..., CG_L$ atliekant hierarchinį klasterizavimą visiems matams pagal $W_{i,j}$
//Matų atrinkimas grįstas konsensuso grupėmis
for $i = 1$ **to** l **do**
    Parinkti reprezentatyvų matą X_i iš CG_i
    Įvertinti mato informatyvumą $\Phi(X_i)$
end for
Reitinguoti konsensuso grupes $CG_1, CG_1, ..., CG_L$ pagal $\Phi(X_i)$
Pasirinkti k matų, turinčių geriausią reitingą

2.4. Suprogramuotų matų atrinkimo algoritmų palyginimas

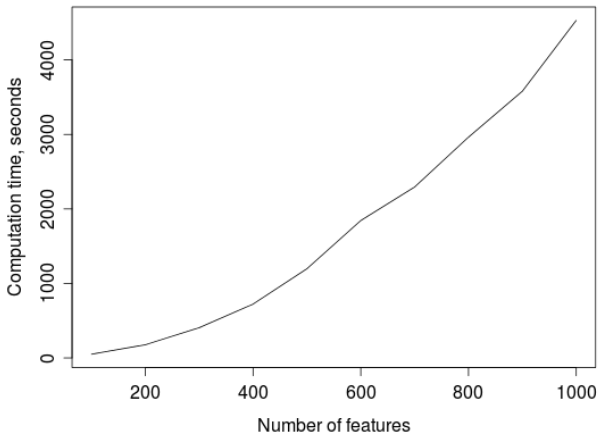
Matų atrinkimo metodus palyginau trim aspektais: darbo laiko, klasifikavimo pagal atrinktus matus tikslumo, bei matų atrinkimo stabilumo. Toliau aprašysiu eksperimentuose gautus rezultatus atskirai.

2.4.1. Matų atrinkimo metodų darbo laikas

Matų atrinkimo metodų darbo laikas buvo palygintas naudojant vieną biomedicininių duomenų rinkinį - AltarA [Ins]. Skaičiavimai buvo atlikti kompiuteryje naudojant vieną procesoriaus branduolį veikiantį 2.66 GHz, bei 2 GB RAM atminties. 8 pav. ir 9 pav. pavaizduota matų atrinkimo metodo darbo laiko priklausomybė nuo mėginius apibūdinančių matų skaičiaus. Matų atrinkimo metodų darbo laikas yra atvaizduotas dviem grafikais, nes atliekant eksperimentų rezultatai parodė, kad CGS matų atrinkimo metodas yra apie 1000 kartų lėtesnis už kitus suprogramuotus matų atrinkimo metodus, todėl viename grafike



8 pav.: Pagrindini matų atrinkimo metodų darbo laikas.

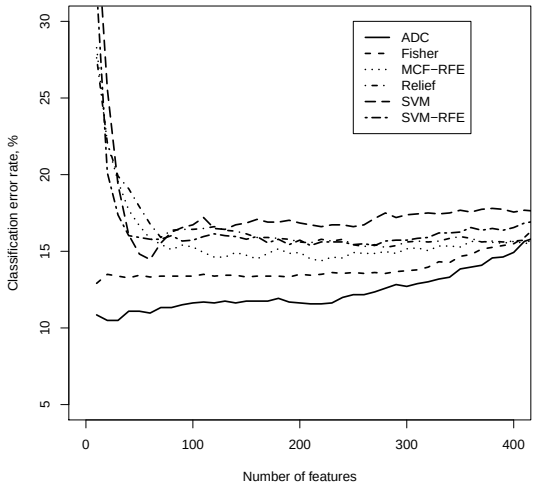


9 pav.: Konsensuso grupėmis grįsto matų atrinkimo metodo darbo laikas.

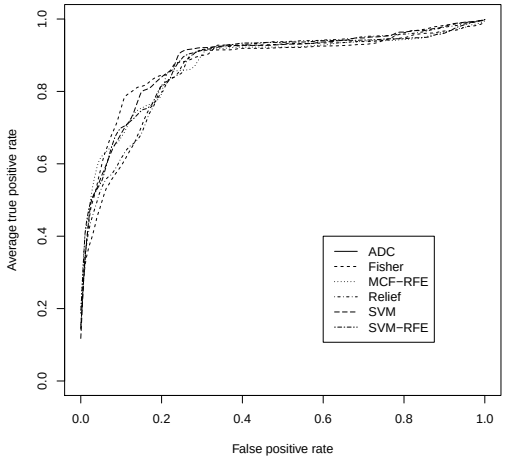
neįmanoma atvaizduoti visų turimų matų atrinkimo metodų. Pagal 8 pav. galime daryti išvadą, kad sparčiausias matų atrinkimo metodas yra *Fisher* įvertis. Pagal gautus matų darbo laiko priklausomybės nuo matų kiekio grafikus galime daryti išvadą, kad CGS algoritmas daugiamačių duomenų matų atrinkimui nėra tinkamas, nes darbo laikas yra per ilgas.

2.4.2. Klasifikavimo pagal atrinktus matus tikslumas

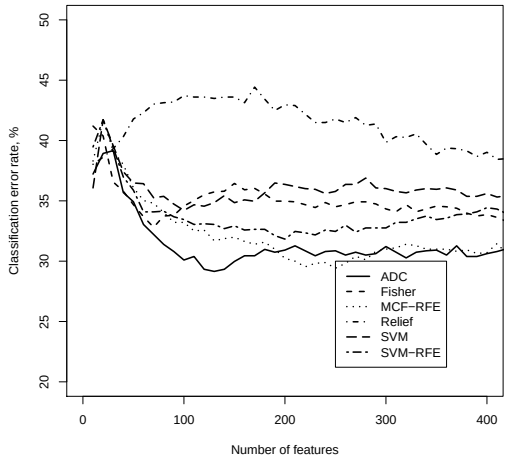
Matų atrinkimo metodų įtaką klasifikavimo klasifikavimo tikslumui buvo matuojama naudojant tris biomediciniinių duomenų rinkinius: Gaubtinės žarnos auglio (angl. Colon) [ABN+99], Centrinės nervų sistemos (CNS) [PTG+02], prostatos [SFR+02]. Klasifikavimui buvo naudojami tiesiniai atraminių vektorių klasifikatoriai (SVM) [Vap00], su parametru $C = 0.01$, kurį nustačiau empiriškai. Keičiant parametrus keičiasi ir klasifikavimo tikslumas. Klasifikatoriui apmokyti buvo naudojama 90% atsitiktinai parinktų mėginių iš duomenų rinkinio. Likusiais 10% mėginių buvo testuojamas klasifikatorius. Klasifikatorius buvo testuojamas po 300 kartų su įvairiu matų skaičiumi: nuo 10 iki 500. Klasifikavimo tikslumas pavaizduotas dviejų tipų grafikais: vidutinio klaidų procento priklausomybės nuo atrinktų dimensijų skaičiaus, bei ROC kreivėmis, kurios buvo gautos pagal duomenis gautus klasifikuojant su tiek atrinktų dimensijų su kiek klasifikavimo tikslumas buvo pats geriausias [GS66].



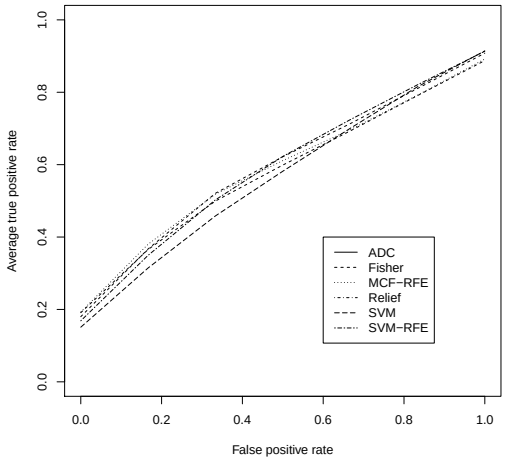
10 pav.: Gaubtinės žarnos auglio meginių klasifikatorių tikslumas.



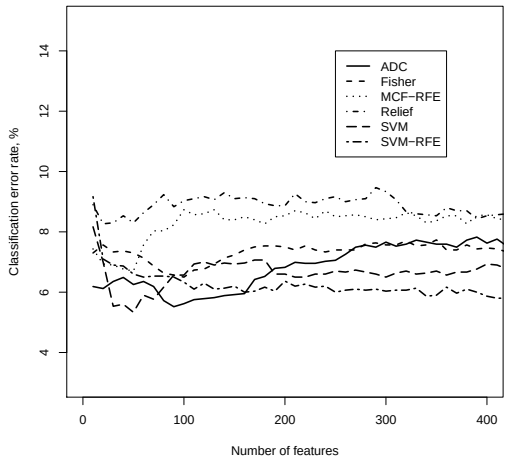
11 pav.: Gaubtinės žarnos auglio mėginių klasifikatorių ROC kreivės.



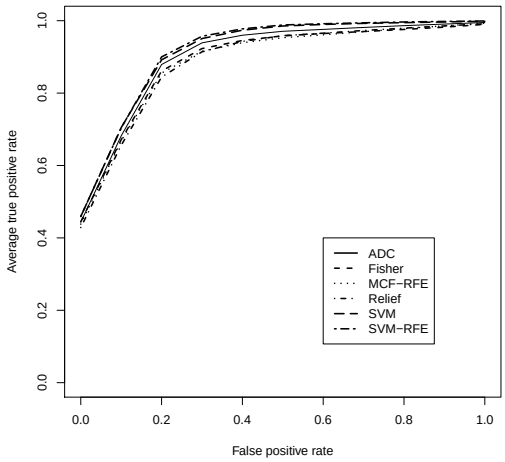
12 pav.: Centrinės nervų sistemos meginių klasifikatorių tikslumas.



13 pav.: Centrinės nervų sistemos mėginių klasifikatorių ROC kreivės.



14 pav.: Prostatos meginių klasifikatorių tikslumas.



15 pav.: Prostatos mėginių klasifikatorių ROC kreivės.

10 pav. matome, kad gaubtinės žarnos auglio duomenų rinkinio matus geriausiai atrenka ADC metodas. Tik šiek tiek prasčiau pasirodo *Fisher* įvertis. Blogiausiai su gaubtinės žarnos auglio mėginiais susidoroja absoliučių svorių SVM matų atrinkimo metodas.

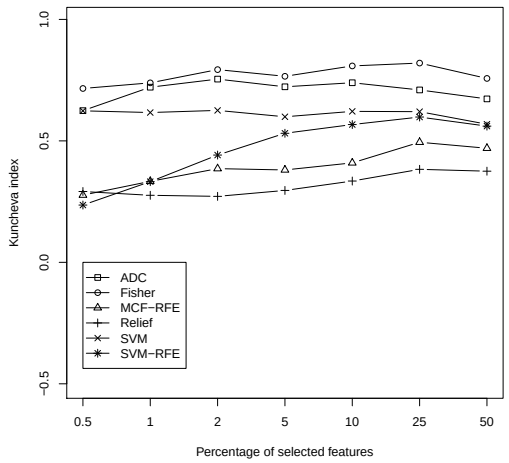
Centrinės nervų sistemos duomenų rinkinys yra sunkiai klasifikuojamas, nes vidutinis klaidų skaičius yra apie 35%, kai, pvz. gautinės žarnos auglio duomenų rinkinio vidutinis klaidų skaičius yra tik 15%. 12 pav. matome, kad šiam duomenų rinkiniui vidutiniškai geriausiai matus atrenka ADC ir multikriterinio rekursyvaus dimensijų eliminavimo metodai. Prasčiausiai pasirodo *Relief* metodas.

14 pav. matome, kad prostatos duomenų rinkinio matus klasifikavimui geriausiai atrenka ADC absoliučių svorių SVM metodas. Prasčiausiai matus atrenka *Relief*.

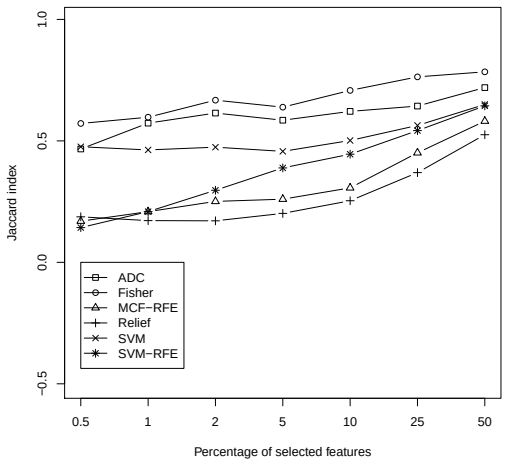
Apibendrindamas gautus klasifikavimo tikslumo matavimo rezultatus, galiu teigti, kad nėra vieno absoliučiai geriausio matų atrinkimo metodo. Reikia eksperimentuoti, kad būtų rastas konkrečiai problemai geriausiai tinkantis matų atrinkimo metodas. Tačiau rezultatai parodė, kad matų atrinkimas svariai prisideda prie geresnio klasifikatoriaus sukūrimo.

2.4.3. Matų atrinkimo stabilumas

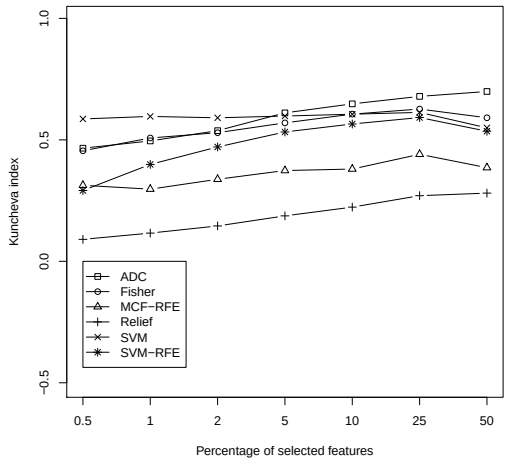
Matų atrinkimo stabilumas buvo tiriamas naudojant tuos pačius biomedicininių duomenų rinkinius kaip ir tiriant klasifikavimo pagal atrinktus matus tikslumą. Matų atrinkimo stabilumas buvo matuojamas pagal *Kuncheva* ir *Jaccard* indeksus. Stabilumas pats savaime nėra svarbus, jis turi būti matuojamas atsižvelgiant į klasifikavimo tikslumą. Todėl šio skyrelio grafikus reikia nagrinėti atsižvelgiant į skyrelio, kuriame buvo nagrinėtas klasifikavimo pagal atrinktus matus tikslumas.



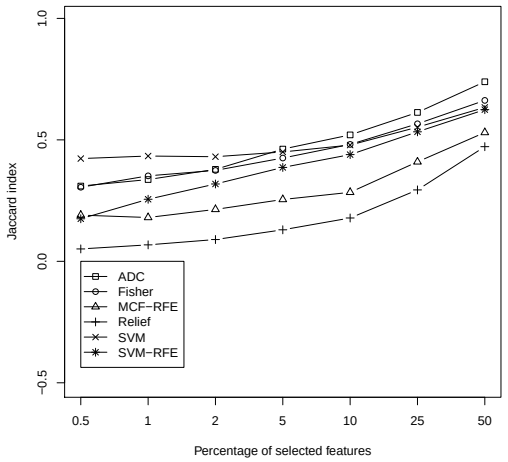
16 pav.: Matų atrinkimo gaubtinės žarnos auglio mėginiams stabilumo grafikas pagal Kuncheva indeksą.



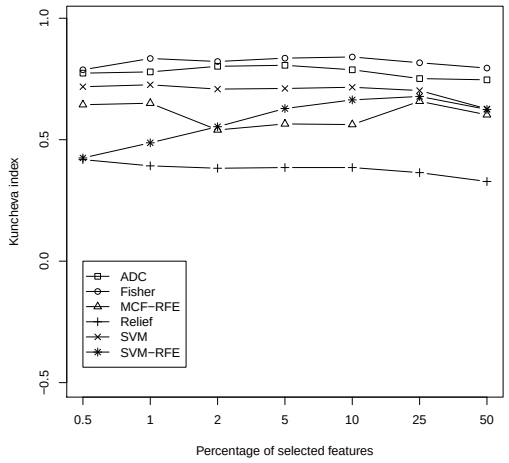
17 pav.: Matų atrinkimo gaubtinės žarnos auglio mėginiams stabilumo grafikas pagal Jaccard indeksą.



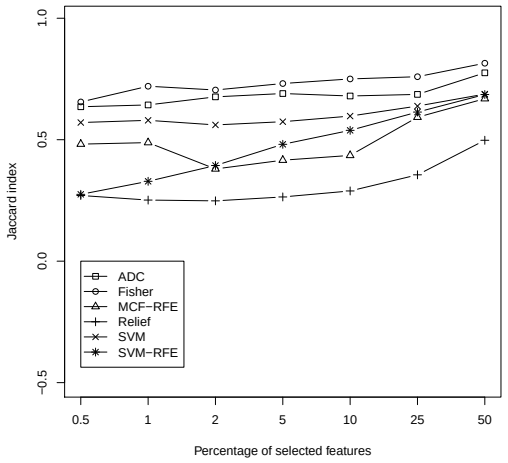
18 pav.: Matų atrinkimo CNS mėginiams stabilumo grafikas pagal Kuncheva indeksą.



19 pav.: Matų atrinkimo CNS mėginiams stabilumo grafikas pagal Jaccard indeksą.



20 pav.: Matų atrinkimo prostatos mėginiams stabilumo grafikas pagal Kuncheva indeksą.



21 pav.: Matų atrinkimo prostatos mėginiams stabilumo grafikas pagal Jaccard indeksą.

Pagal 16 pav. ir 17 pav. matome, kad gaubtinės žarnos auglio duomenų rinkinio matus stabiliausiai atrenka *Fisher* įvertis. Mažiausiai stabiliai matus atrenka *Relief* metodas.

Pagal 18 pav. ir 19 pav. matome, kad centrinės nervų sistemos duomenų rinkinio matus stabiliausiai atrenka absoliučių svorių SVM matų atrinkimo metodas. Mažiausiai stabiliai matus atrenka *Relief* metodas.

Pagal 20 pav. ir 21 pav. matome, kad prostatos duomenų rinkinio matus stabiliausiai atrenka *Fisher* įvertis. Mažiausiai stabiliai matus atrenka *Relief* metodas.

Apibendrindamas matų atrinkimo stabilumo matavimus galiu sakyti, kad matų atrinkimo stabilumas priklauso ne tik nuo matų atrinkimo metodo, bet ir nuo duomenų rinkinio, kurio matai yra atrinkinėjami. Lengvai klasifikuojamo prostatos duomenų rinkinio matų atrinkimo stabilumas vidutiniškai yra didesnis nei sunkiai klasifikuojamo CNS duomenų rinkinio. Eksperimentų rezultatai rodo, kad *Relief* matų atrinkimo metodas yra nestabiliaus iš tirtųjų. Gana geru stabilumu pasižymi ADC metodas bei *Fisher* įvertis.

3. REZULTATAI, IŠVADOS IR PASIŪLYMAI

Profesinės praktikos metu susipažinau su matų atrinkimo daugiamačiuose duomenyse problematiką nagrinėjančia literatūra, suprogramavau pagrindinius matų atrinkimo algoritmus bei atlikau suprogramuotų algoritmų palyginamąją analizę. Įvykdęs profesinei praktikai keltus uždavinius, galiu tvirtinti, kad matų atrinkimas daugiamačiuose duomenyse yra sudėtinga problema, nes vienareikšmiškai teigti, kuris matų atrinkimo algoritmas yra absoliučiai geriausias, negalima – matų atrinkimo algoritmą visada reikia pasirinkti atsižvelgiant į tiriamus duomenis ir sprendžiamai problemai keliamus uždavinius.

Didžiausias praktikos darbo privalumas yra tai, kad visą darbo dieną galima skirti konkrečių uždavinių įgyvendinimui ir planuoti darbų atlikimo laiką. Tai pasiekiama turint asmeninę darbo vietą. Be to, atliekant profesinę praktiką yra proga pasisemti patirties iš vyresniųjų kolegų.

Dirbdamas su VU MII mokslininkais įgijau daugiamačiais duomenimis apibūdinamų procesų žinių – dalyvavau MII seminaruose. Duomenų analizės žinių sėmiausi iš mokslinės literatūros, bei konsultacijų su kolegomis. Taip pat profesinės praktikos metu pagerinau programavimo įgūdžius, nes tyrimus atlikti reikėjonaudojantis statistinei analizei skirta programavimo kalba *R*.

Džiugina tai, kad universitete įgytų bendrųjų programavimo įgūdžių pakako atliekant profesinę praktiką. Reikėjo tik išmokti dirbti su nauja programavimo kalba. Tačiau praktikos metu jaučiau duomenų analizės ir statistikos žinių stygių – nagrinėta problematika reikalauja gilesnio žinojimo.

Mokymą universitete siūlyčiau gerinti peržiūrint studijų tvarkaraštį. Profesinė praktika yra naudinga studijų procesui, bet ji prasideda iškart po įtempto sesijos mėnesio, kuris prasideda dar prieš šv. Kalėdas, o prieš šv. Kalėdas yra semestro pabaiga, kurios metu reikia užbaigti semestro darbus, parašyti visus kontrolinius, bei pasirūpinti Kalėdinėmis dovanomis artimiesiems. Kitaip tariant, po poros sunkių mėnesių (gruodis, sausis), prasideda profesinė praktika, kurios pradžioje studentas tiesiog natūraliai (ir pagrįstai) norėtų pailsėti. O nuo pat pirmųjų profesinės praktikos dienų reikia skubėti spręsti išsikeltus uždavinius, nes laiko gali, ir greičiausiai pritrūks.

Žinoma, šansai didesni, kad laiko pritrūks, jei profesinė praktika atliekama ne nuolatinėje darbovietėje. Atliekant praktiką nuolatinėje darbovietėje laiko neturėtų pritrūkti, nes

nuolatinėje darbovietėje darbai planuojami ilgesniam nei 11 savaičių periodui. Tačiau darbas nuolatinėje darbovietėje studijų metų nėra tai, ką turėtų daryti studentas – studentas turi studijuoti universitete.

Mano siūlomas studijų tvarkaraštis būtų toks: rudens semestras prasideda mėnesiu vėliau nei įprasta - spalio pirmą ar rugsėjo paskutinę savaitę; šventiniu laikotarpiu (šv. Kalėdos, Naujieji metai) studentams yra duodamos dviejų ar triejų savaičių atostogos; rudens semestro pabaiga perkeliama į sausio pabaigą; tada mėnuo sesijai; pora savaičių atostogų; praktika ir bakalaurinis darbas arba pavasario semestras; atsiskaitymai už pavasario semestrą. Pagal mano siūlomą tvarkaraštį vasaros atostogos sutrumpėtų iki mėnesio, bet tai tikrai nebūtų bėda, nes, pragmatiškai žiūrint, du mėnesiai yra per mažai pradėti rimtai dirbti, bet du mėnesiai yra per daug laiko tiesiog nieko neveikti. Studijų tvarkaraščio peržiūrėjimas pagerintų visą studijų procesą.

Literatūra

[ABN+99] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745, 1999.

[Bel66] R.E. Bellman. *Adaptive control processes: A guided tour*. Princeton University Press, 1966.

[DKNS01] C. Dwork, R. Kumar, M. Naor, and D. Sivakumar. Rank aggregation revisited. In *Proceedings of WWW10*, pages 613–622, 2001.

[Fis36] R.A. Fisher. The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, 7(2):179–188, 1936.

[GE03] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.

[GS66] D.M. Green and J.A. Swets. *Signal detection theory and psychophysics*, volume 1974. Wiley New York, 1966.

[GWBV02] I. Guyon, J. Weston, S. Barnhill, and V. Vapnik. Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1):389–422, 2002.

[Ins] Stanley Medical Research Institute. Online genomics database. [žiūrėta 2012-04-03]. Prieiga per internetą: <www.stanleygenomics.org>.

[Ken83] J.T. Kent. Information gain and a general measure of correlation. *Biometrika*, 70(1):163–173, 1983.

[LYD09] S. Loscalzo, L. Yu, and C. Ding. Consensus group stable feature selection. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 567–576. ACM, 2009.

[PTG+02] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, et al. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[RSK03] Marko Robnik-Sikonja and Igor Kononenko. Theoretical and empirical analysis of relieff and rrelieff. *Machine Learning*, 53(1-2):23–69, 2003.

[SAVdP08] Y. Saeys, T. Abeel, and Y. Van de Peer. Robust feature selection using ensemble feature selection techniques. *Machine Learning and Knowledge Discovery in Databases*, pages 313–325, 2008.

[SFR+02] D. Singh, P.G. Febbo, K. Ross, D.G. Jackson, J. Manola, C. Ladd, P. Tamayo, A.A. Renshaw, A.V. D'Amico, J.P. Richie, et al. Gene expression correlates of clinical prostate cancer behavior. *Cancer cell*, 1(2):203–209, 2002.

[Sha01] C. E. Shannon. A mathematical theory of communication. *SIGMOBILE Mob. Comput. Commun. Rev.*, 5(1):3–55, January 2001.

[Vap00] V.N. Vapnik. *The nature of statistical learning theory*. Statistics for engineering and information science. Springer, 2000.

[YDL08] L. Yu, C. Ding, and S. Loscalzo. Stable feature selection via dense feature groups. In *Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 803–811. ACM, 2008.

[YM11a] F. Yang and KZ Mao. Robust feature selection for microarray data based on multicriterion fusion. *IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB)*, 8(4):1080–1092, 2011.

[YM11b] Feng Yang and K.Z. Mao. Robust feature selection for microarray data based on multicriterion fusion. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 8(4):1080 –1092, july-aug. 2011.