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 yra 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 turimas. Biomediciniinių 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 biomediciniinių duomenų analizės procesas yra sudėtingas.

Biomediciniinių 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 atrenkamos 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 išvadas, 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 klasės, 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 klasės (angl. *nearest-Hit*), x_M – artimiausias x kaimynas iš kitos klasės(angl. *nearest-Miss*), $diff(j, x, x')$ – j-ojo mato reikšmių skirtumas tarp laisvai pasirinkto objekto x ir atitinkamo 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 maximali ir minimali j-ojo matų reikšmės. „Susietumo“ reikšmių atnaujinimas yra vykdomas n kartų ir kuo didesnė galutinė reikšmė, tuo svarbesnis matas. Aprašyta algoritmo versija yra skirta dviejų klasių atvejui, tačiau yra ir multiklasinis algoritmo variantas.

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 metodas (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 – dimensijų kiekis, w_j – j-osios dimensijos svoris, x_j – j-osios dimensijos kintamasis, b_0 – konstanta. Dimensijos absoliutus[1] svoris w_j gali būti panaudotas dimensijų reitingavimui. Pastebėtina, kad svorių nustatymas yra atliekamas tik vieną kartą[2].

2.3. Suprogramuotų dimensijų atrinkimo algoritmų palyginimas

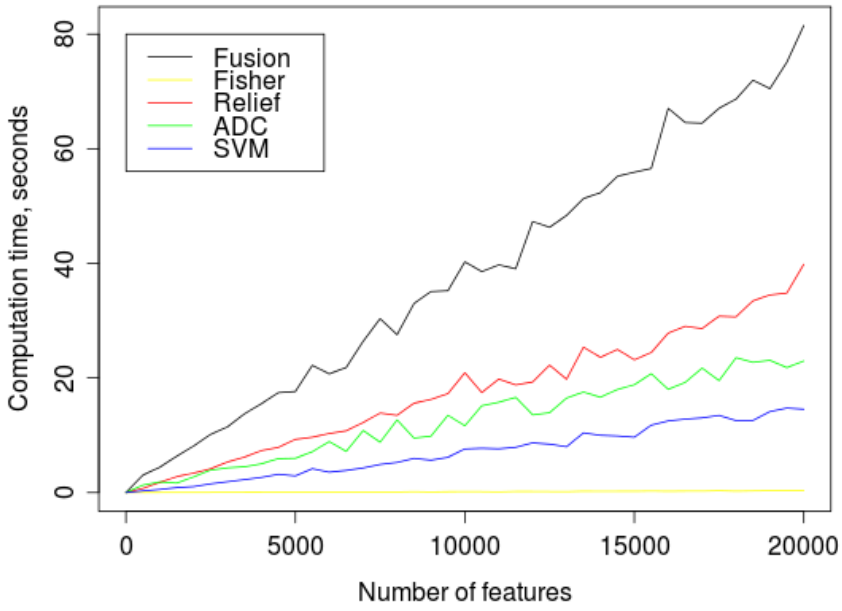
2.3.1. Dimensijų atrinkimo algoritmų skaičiavimo laikas

Eksperimentai buvo atlikti su AltarA duomenų rinkiniu, kompiuteryje naudojant tik vieną procesoriaus branduolį, bet 2 GB RAM atminties.

1 pavaizduotas skaičiavimo laikas nuo mėginius apibūdinančių dimensijų skaičiaus.

[1] Svorį reikia imti absoliutaus dydžio, nes neigiamas svoris implikuoja priklausomybę vienai klasei, o teigiamas kitai klasei.

[2] SVM-RFE dimensijų atrinkimo metodas svorius nustato daug kartų.

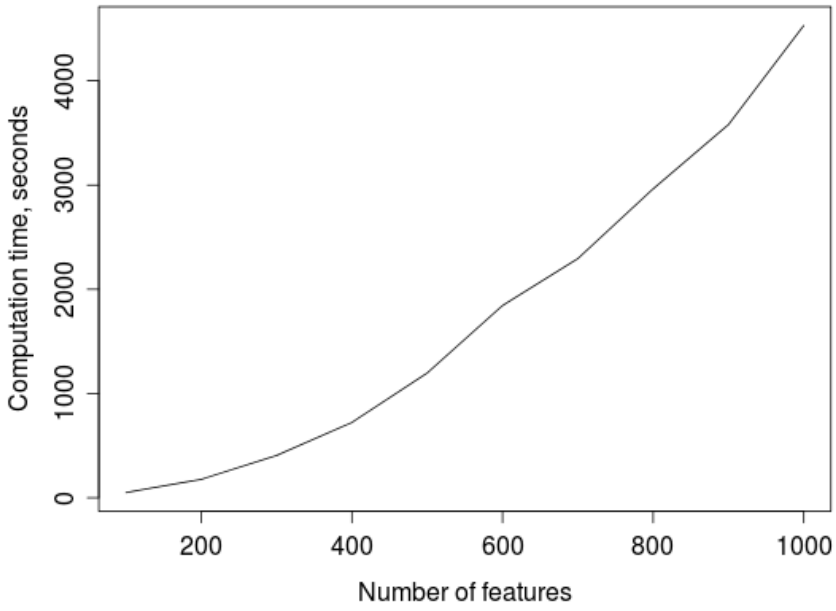


1 pav.: Pagrindinių dimensijų atrinkimo metodų skaičiavimo laikas.

Pats sparčiausias dimensijų atrinkimo metodas yra *Fisher* įvertis. Lėčiausias multikriterinis dimensijų atrinkimo *Fusion* metodas. 2 pavaizduotas konsensuso grupėmis grįsto dimensijų atrinkimo algoritmo skaičiavimo laiko priklausomybė nuo mėginius apibūdinančių dimensijų kiekio. Algoritmo sudėtingumas laiko atžvilgiu yra kvadratinis. Jei lyginsime su dimensijų reitingavimo algoritmais, tai šis algoritmas yra daug kartų lėtesnis.

Pagal gautus laiko priklausomybės nuo dimensijų kiekio grafikus galime daryti išvadą, kad CGS algoritmas daugiamačių duomenų dimensijų atrinkimui nėra tinkamas, nes darbo laikas yra per didelis.

2.3.2. Klasifikavimo tikslumas



2 pav.: Konsensuso grupėmis grįtas dimensijų atrinkimo metodo skaičiavimo laikas.

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 vieni matų atrinkimo metodai pvz. AW-SVM, greitai atrenka matus, kurie pagerina klasifikavimo tikslumą, tačiau atrinktas matų poaibis nėra stabilus; arba konsensuso grupėmis grįstas matų atrinkimas yra labai lėtas, tačiau jis atrenka stabilų matų poaibį, kuris pagerina klasifikavimo procesą. Todėl vienareikšmiškai teigti, kuris matų atrinkimo algoritmas yra absoliučiai geriausias, negalima – matų atrinkimo algoritmą visada reikia pasirinkti atsižvelgiant į pačius tiriamus duomenis ir tiriamai problemai keliamus uždavinius.

Didžiausias praktikos darbo privalumas yra tai, kad visą darbo dieną galima skirti konkrečių uždavinių įgyvendinimui ir planuoti darbų atlikimo laiką. To pasiekti galima, nes atvažiavus į profesinės praktikos vietą ryte iki pat vakaro nereikia gaišti laiko kelionėms mieste. Be to, atliekant profesinę praktiką yra proga pasisemti patirties iš kolegų.

Dirbdamas su VU MII mokslininkais įgijau daugiamačiais duomenimis apibūdinamų

procesų, duomenų analizės žinių bei pagerinau programavimo įgūdžius. Daugiamačiais duomenimis apibūdinamų procesų žinių sėmiausi MII rengiamų seminarų metu. Duomenų analizės žinių sėmiausi iš mokslinės literatūros, bei konsultacijų su kolegomis. Profesinės praktikos metu reikėjo programuoti 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 jaučiau duomenų analizės ir statistikos žinių stygių – nagrinėjama problematika reikalauja gilesnio žinojimo.

Mokymą universitete siūlyčiau gerinti peržiūrint studijų tvarkaraštį. Profesinė praktika yra naudinga studijų procesui, bet ji prasideda iškart po labai įtemptos sesijos, kuri 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 praktika, kurios pradžioje studentas tiesiog natūraliai (ir pagrįstai) norėtų pailsėti. O, savo ruožtu, praktikoje reikia skubėti spręsti praktikai išsikeltus uždavinius, nes laiko gali, ir greičiausiai pritrūks. Žinoma, laiko pritrūks, jei profesinė praktika atliekama ne nuolatinėje darbovietėje. Atliekant praktiką nuolatinėje darbovietėje laiko nepritrūks, nes darbai dažniausiai yra 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 trijų savaičių atostogos; rudens semestro pabaiga perkeliama į sausio pabaigą; tada mėnuo sesijai; savaitė ar dvi atostogų; praktika ir bakalaurinis darbas arba pavasario semestras; atsiskaitymai už pavasario semestrą.

Literatūra

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

[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.

[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.

[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.

[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.

[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.

STUDENTO PROFESINĖS PRAKTIKOS ĮVERTINIMO ANKETA

Studento vardas, pavardė	Dainius Jocas
Organizacijos pavadinimas	Vilniaus Universiteto Matematikos ir informatikos institutas
Informacija apie organizacijos praktikos vadovą	
Vardas Pavardė	Juozas Gordevičius
Pareigos	Mokslininkas stažuotojas
Telefonas	861060239

1-20 teiginiuose prašome pažymėti atitinkamą skaičių pagal pateiktą vertinimo skalę:

labai silpnai – 1 ---------- 5 – labai gerai

Įvertinkite studento teorinių žinių ir praktinių gebėjimų lygį praktikos pradžioje ir pabaigoje

1.	Studento teorinių žinių lygis praktikos pradžioje	1	2	3	4	5
2.	Studento praktinių gebėjimų lygis praktikos pradžioje	1	2	3	4	5
3.	Studento teorinių žinių lygis praktikos pabaigoje	1	2	3	4	5
4.	Studento praktinių gebėjimų lygis praktikos pabaigoje	1	2	3	4	5

Įvertinkite bendrųjų studento gebėjimų lygį praktikos metu

5.	Praktikoje taiko reikiamas žinias, įgūdžius ir metodus	1	2	3	4	5
6.	Įvardija esamas problemas ir siūlo jų sprendimo būdus	1	2	3	4	5
7.	Stebi ir įvertina procesų/veiklos efektyvumą	1	2	3	4	5
8.	Prisiima atsakomybę už atliekamas užduotis	1	2	3	4	5
9.	Efektyviai paskirsto laiką	1	2	3	4	5

Įvertinkite studento asmenines savybes praktikos metu

10.	Punktualumas	1	2	3	4	5
11.	Motyvacija	1	2	3	4	5
12.	Imlumas praktiniam mokymui	1	2	3	4	5
13.	Darbštumas	1	2	3	4	5
14.	Patikimumas	1	2	3	4	5
15.	Bendravimas su darbuotojais ir klientais	1	2	3	4	5

Įvertinkite studento praktikos temos ir praktikos metu įvykdytų uždavinių reikšmę

16.	Sudėtingumas	1	2	3	4	5
17.	Nauda organizacijai	1	2	3	4	5
18.	Svarbumas versle	1	2	3	4	5
19.	Novatoriškumas nacionaliniame lygmenyje	1	2	3	4	5
20.	Novatoriškumas tarptautiniame lygmenyje	1	2	3	4	5

Nurodykite, kokių žinių ir gebėjimų studentui labiausiai trūko atliekant praktines užduotis

Organizacijos praktikos vadovo parašas _____

Data _____