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 matematikos ir informatikos instituto (MII) sistemų analizės skyrių dėl keletos priežasčių. Visų 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.

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

Profesinės praktikos tikslas - atlikti dimensijų atrinkimo metodų palyginamąją analizę. Siekiant užsibrėžto tikslo profesinės praktikos metu buvo sprendžiami šie uždaviniai:

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

Praktinės veiklos planas buvo sudarytas iš dviejų dalių: suprogramuoti pasirinktus dimensijų atrinkimo algoritmus ir palyginti suprogramuotus algoritmus skaičiavimo laiko, klasifikavimo tikslumo bei dimensijų atrinkimo stabilumo atžvilgiais tarpusavyje. Du penktadaliai numatyto profesinės praktikos laiko buvo skirta dimensijų atrinkimo algoritmų prog-

ramavimui, dar du penktadaliai buvo numatyti algoritmų palyginimui, o likęs laikas susipažinimui su dalykinės srities literatūra bei dalyvavimui MII rengiamuose seminaruose.

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

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

1. ĮSTAIGOS APIBŪDINIMAS

Vilniaus universiteto matematikos ir informatikos institutas (MII) nuo 2010 m. yra Vilniaus universiteto padalinys užsiimantis tyrimais 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 skyrius atsakingas už visą eilę recenzuojamų periodinių leidinių: „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 taip pat yra išleidusi mokslinių bei mokslo populiarinimo knygų lietuvių ir anglų kalbomis, mokymo priemonių, interaktyvių kompaktinių diskų bei sukūrusi įvairių internetinių informacinių sistemų (pvz. enciklopedinis kompiuterijos terminų žodynas).

MII man, kaip ir kiekvienam darbuotojui, parūpino: darbo vietą, prisijungimo prie vietinio tinklo, galimybe naudotis skaičiavimo ištekliais, galimybe su nuolaida pietauti vietinėje valgykloje. MII darbuotojai buvo kolegiški, todėl apsipratimas MII įvyko labai greitai. Todėl jau nuo pat pirmosios profesinės praktikos dienos galėjau pradėti spręsti užsibrėžtus uždavinius.

2. PROFESINĖS PRAKTIKOS VEIKLOS APRAŠYMAS

Šiame skyriuje aprašysiu profesinės praktikos veiklas, kurias skirsčiau į užduotis. Profesinę praktiką sudarė trys užduotys: susipažinimas su daugiamačių duomenų dimensijų atrinkimo problematika, dimensijų atrinkimo metodų programavimas ir dimensijų atrinkimo metodų savybių tyrimas. Toliau šiame skyriuje aprašysiu kiekvieną užduotį atskirai.

2.1. Įvadas į profesinės praktikos metu nagrinėtą problematiką ir literatūros apžvalga

Nuolat vystosi technologijos skirtos gauti biomedicininius duomenis, pvz. genomo sekvenavimas [PLA09], o tai reiškia, kad didėja gaunamų duomenų detalumas. Detalumas reiškia, kad daugėja biomedicininius duomenis abibūdinančių faktorių arba matų skaičius. Duomenys, kur kiekvienas mėginys aprašomas dideliu kiekiu matų, yra vadinami daugiamačiais duomenimis.

Šiame darbe yra nagrinėjama biomedicinoje kaupiamų genetinių daugiamačių duomenų analizės specifika. Šie duomenys yra ypatingi tuo, kad jie įprastai turi šimtus kartų daugiau matų nei mėginių. Santykinai mažas mėginių skaičius turimas, nes mėginio gavimo kaina yra aukšta. Biomedicininių duomenų analizę apsunkina ir tai, kad matavimai, kuriais tie duomenys gaunami, įneša atsitiktinių duomenų - triukšmo. Triukšmas matavimo metu gali atsirasti dėl įvairių priežasčių, pvz. netinkamai paruoštų cheminių preparatų. Kai duomenys yra triukšmingi, didėja tikimybė duomenyse rasti atsitiktinių priklausomybių. Tai yra viena priežasčių, kodėl biomedicininių duomenų analizės procesas yra sudėtingas.

Klasifikavimu [Fis36] yra vadinamas duomenų analizės procesas, kai duomenys suskirstomi į grupes pagal tam tikrus jų požymius. Algoritmai arba funkcijos, kurios turimus duomenis priskiria iš anksto žinomoms grupėms - atlieka klasifikavimą - yra vadinami klasifikatoriais. Klasifikatoriai paruošiami naudojant turimus mėginius - treniravimosi duomenis - ir informaciją apie jų būklę (sveikas ar sergantis). Klasifikatoriaus ruošimo procesas yra vadinamas apmokymu. Apmokyti klasifikatoriai paprastai naudojami nustatant naujų, dar nematytų, mėginių būklę. Biomedicininių duomenų klasifikavimo tipinė 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.

Biomediciniinių duomenų kontekste galima daryti prielaidą, kad ne visi matai yra susiję su tiriama problema, pvz. gaubtinės žarnos vėžiu, dėl tokių faktorių, kaip triukšmas duomenyse. Paprastai nagrinėjamai problemai svarbus yra mažas, palyginus su visu, matų kiekis. Ši biomediciniinių duomenų ypatybė veda prie „daugiamatiškumo prakeiksmo“ (angl. *the curse of dimentionality*) [Bel66] - didėjant matų kiekiui mėginiai pasidaro panašūs, todėl bandymas juos klasifikuoti tolygus spėliojimui. Todėl biomediciniinių duomenų daugiamatiškumui sumažinti yra naudojami informatyviausių dimensijų atrinkimo metodai [GE03] (angl. *feature selection*). Pagal tai, kaip susiję su klasifikatoriumi, dimensijų atrinkimo metodai skirstomi į tris kategorijas [SAVdP08]: filtruojantys (angl. *filter*), prisitaikantys (angl. *wrapper*) ir įterptiniai (angl. *embedded*) metodai. Filtruojančiai metodais pirmiausia yra atrenkamos informatyviausios dimensijos, o tada apmokomas klasifikatorius. Prisitaikančiųjų metodų atveju, pirma, apmokomas klasifikatorius su visomis dimensijomis, antra, parenkamas dimensijų poaibis ir apmokomas klasifikatorius, tada po daugkartinio dimensijų aibių įvertinimo pagal klasifikavimo rezultatus yra nusprendžiama, kuris dimensijų poaibis yra labiausiai tinkamas klasifikavimui. Įterptinių metodų atveju dimensijų atrinkimo procesas yra neatsiejamas nuo klasifikavimo proceso - pats klasifikatorius įvertina dimensijas.

Dimensijų atrinkimas yra svarbi biomediciniinių duomenų apdorojimo (angl. *preprocessing*) etapo dalis. Naudojant dimensijų atrinkimo metodus galima kovoti su daugiamatiškumo prakeiksmu dimensijų skaičių priartinant prie mėginių skaičiaus. Todėl svarbu yra pasirinkti geriausiai tinkančią dimensijų atrinkimo strategiją. Kadangi ir patys dimensijų atrinkimo metodai, turi savo ypatybių, pvz. algoritmo sudėtingumas, tai pačių dimensijų atrinkimo metodų pasirinkimas tampa sudėtinga užduotimi.

Naudodami dimensijų atrinkimo metodus, biomediciniinius duomenis tiriantys mokslininkai susiduria su atrinktųjų dimensijų aibės stabilumo problema - atrenkant dimensijas pagal kitą mėginių poaibį, gaunamas kitas dimensijų poaibis. Dimensijų atrinkimo nestabilumas yra sąlygotas šių veiksnių:

1. Duomenys yra triukšmingi ir kai kurios dimensijos gali būti palaikytos informatyviomis grynai dėl atsitiktinių priežasčių;

2. Daugiamačiuose duomenyse tikėtina, kad dalis dimensijų koreliuoja, todėl, kuri iš ko-

reliuojančių dimensijų bus pasirinkta, priklauso nuo to, kuriuos mėginius pasirinksime klasifikatoriaus apmokymui;

3. Kiekvienas dimensijų atrinkimo algoritmas daro skirtingas prielaidas apie tai, kurios dimensijos yra informatyvios.

Galime daryti išvadas, kad skirtingi metodai tiems patiems duomenims gali atrinkti skirtingas dimensijas. Taip pat, suskaidžius turimus duomenis į atskiras persidengiančias aibes ir atrinkus tą patį kiekį dimensijų tuo pačiu metodu, gaunamos skirtingos dimensijų aibės. Be to, kuo triukšmingesni duomenys, kuo mažiau turima mėginių ir kuo daugiau yra dimensijų, tuo ryškesnė yra ši problema [LYD09].

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

Yang ir Mao pasiūlė reitinguoti dimensijas remiantis keletos dimensijų atrinkimo metodų rezultatais [YM11a]. Galutinis dimensijų reitingų sąrašas gaunamas, kai po kiekvieno dimensijų atrinkimo yra išmetama viena žemiausią reitingą turinti dimensija iš dimensijų aibės, ir dimensijų atrinkimas yra kartojamas tol, kol nebelieka dimensijų. Tačiau dimensijų atrinkimo metodų kiekis yra ribotas ir skirtingų metodų dažnai negalima atlikti paraleliai. Tai riboja šio metodo pritaikomumą daugiamačių duomenų analizėje.

Didinti dimensijų atrinkimo stabilumui metodų yra, tačiau visi jie turi savų niuansų. Todėl tolesni dimensijų atrinkimo metodų tyrimai turi prasmę.

2.2. Suprogramuoti dimensijų atrinkimo algoritmai

Profesinės praktikos metu suprogramavau populiariausius dimensijų atrinkimo metodus. Taip pat programavau ir dimensijų atrinkimo stabilumą didinančius metodus. Toliau šiame skyrelyje aprašysiu suprogramuotus metodus.

2.2.1. *Fisher* įvertis

Fisher įvertis vertina individualias dimensijas pagal dimensijos klasių atskiriamąją galią. Dimensijos 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 dimensijos indeksas, μ_{jc} - dimensijos j reikšmių vidurkis klasėje c, σ_{jc}^2 - dimensijos j reikšmių standartinis nuokrypis klasėje c, kur $c = 1, 2$. Kuo didesnis yra *Fisher* įvertis, tuo geriau ta dimensija atskiria klases.

2.2.2. *Relief* metodas

Relief metodas iteratyviai skaičiuoja dimensijų „susietumą“. Pradžioje „susietumas“ visoms dimensijoms yra lygus nuliui. Kiekvienoje iteracijoje atsitiktinai[1] pasirenkamas objektas iš mėginių aibės, surandami artimiausi kaimynai iš tos pačios ir kitos klasės, ir atnaujinamos visų dimensijų „susietumo“ reikšmės. Dimensijos įvertis yra vidurkis visų objektų atstumų iki artimiausių kaimynų iš tos pačios ir kitos klasės:

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

kur $W(j)$ - j-osios dimensijos „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-osios dimensijos 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-osios dimensijos reikšmės. „Susietumo“ reikšmių atnaujinimas yra vykdomas n kartų ir kuo didesnė galutinė reikšmė, tuo svarbesnė dimensija. Pastebėtina, kad aprašyta algoritma versija yra skirta dirbti su dviejų klasių atveju, tačiau yra ir multiklasinis algoritmo variantas.

[1] Pastebėtina, kad dėl atsitiktinumo faktoriaus klasifikavimo ir dimensijų atrinkimo stabilumo rezultatai varijuoja.

2.2.3. Asimetrinis priklausomybės koeficientas

Asimetrinis priklausomybės koeficientas (ADC) yra dimensijų reitingavimo motodas, kuris matuoja klasės Y etiketės (angl. *label*) tikimybinę priklausomybę j-ąjai dimensijai, naudodamas informacijos prieaugį [Ken83] (angl. information gain):

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

kur $H(Y)$ - klasės Y entropija [Sha01], o $MI(Y, X_j)$ - yra bendrumo informacija [Sha01] (angl. mutual information) tarp klasės etiketės Y ir j-osios dimensijos

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

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

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

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

Kuo didesni ADC įverčiai, tuo dimensija yra svarbesnė, nes turi daugiau informacijos apie mėginių klases.

2.2.4. Absoliučių svorių SVM

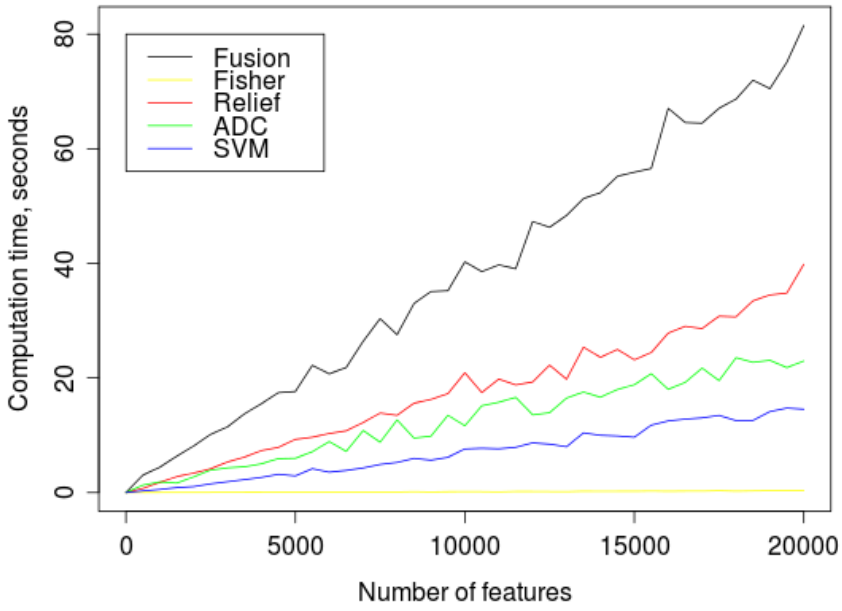
Atraminių vektorių metodas (SVM) yra vienas populiariausių klasifikavimo algortimų, nes jis gerai susidoroja su daugiamačiais duomenimis [GWBV02]. Yra keletas bazinių SVM variantų [Vap00], bet šiame darbe naudosime tiesinį SVM, nes jis demonstruoja gerus rezultatus analizuojant genų ekspresijos duomenimis. 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[2] svoris w_j gali būti panaudotas dimensijų reitingavimui. Pastebėtina, kad svorių nustatymas yra atliekamas tik vieną kartą[3].

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

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



1 pav.: Pagrindinių dimensijų atrinkimo metodų skaičiavimo laikas.

2.3. Suprogramuotų dimensijų atrinkimo algoritmų palyginimas

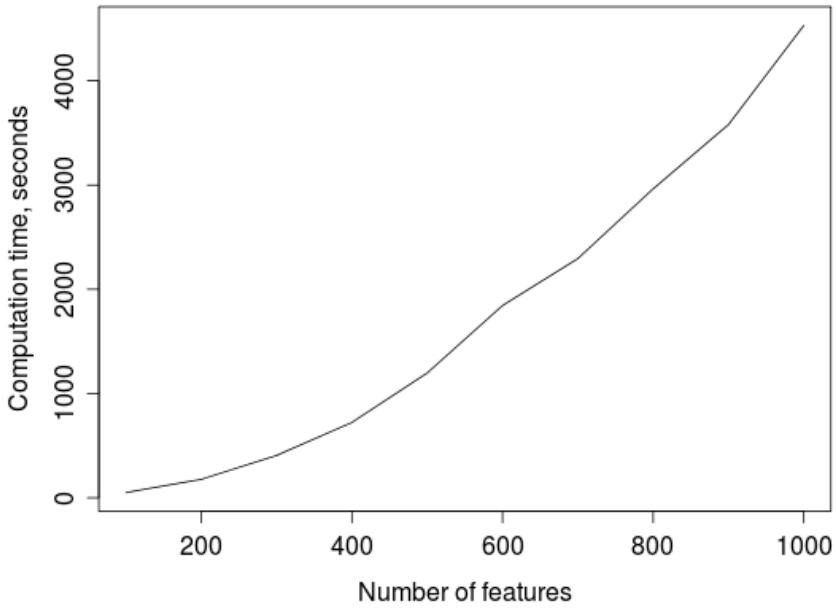
2.3.1. Dimensijų atrinkimo algoritmų skaičiavimo laikas

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. 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

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.

[PLA09] E. Pettersson, J. Lundeberg, and A. Ahmadian. Generations of sequencing technologies. *Genomics*, 93(2):105–111, 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.