VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INORMATIKOS FAKULTETAS
PROGRAMŲ SISTEMŲ KATEDRA

Daugiamačių duomenų klasifikavimo analizė

Classification Analysis of High-Dimensional Data

Bakalauro baigiamasis darbas

Atliko: Dainius Jocas (parašas)

Darbo vadovas: dr. Juozas Gordevičius (parašas)

Darbo recenzentas: prof. dr. Romas Baronas (parašas)

VILNIUS - 2012

SANTRAUKA

Šiame bakalauro baigiamajame darbe yra analizuojamas daugiamačių mažai mėginių turinčių biomedicininių duomenų klasifikavimas. Darbe pateikiama mašininio mokymosi teorijos, reikalingos atliekamai analizei, apžvalga. Daugiamačių duomenų analizė neapsieina be matų atrinkimo metodų taikymo, todėl darbe detaliai nagrinėjami pagrindiniai ir multikriteriniai matų atrinkimo metodai. Dirbant su daugiamačiais duomenims svarbus yra skaičiavimo laikas, todėl matų atrinkimo metodai lyginami spartos atžvilgiu. Daugiamačių duomenų klasifikavimas analizuotas pagal atrinktųjų matų įtaką klasifikavimo tikslumui bei pagal matų atrinkimo stabilumą. Analizės metu gauti duomenys rodo, kad nėra universaliai geriausios klasifikavimo strategijos skirtos daugiamačiams biomedicininiams duomenims. Todėl šis darbas yra įžanga į tolimesnius tyrimus.

Raktiniai žodžiai: daugiamačiai duomenys, mašininis mokymasis, klasifikavimas, matų atrinkimas.

SUMMARY

This bachelor thesis is dedicated to analysis of high-dimensional small-sample biomedical data. This thesis has overview of machine learning theory which is necessary to do desired analysis. First, because analysis of high-dimensional data is rarely done without feature selection, a special emphasis is put on basic and multicriterion feature selection methods. Second, when working with high-dimensional data computation time is always an issue, so, feature selection methods are compared according to their computational performance. Third, classification of high-dimensional data is analyzed by impact of selected features to the classification accuracy and robustness of feature selection. Analysis showed that there is no such thing like the best classification strategy for high-dimensional data biomedical data. That why this work is an introduction for the further studies.

Keywords: high-dimensional data, machine learning, classification, feature selection.

Turinys

ĮVADAS **5**

1. MAŠININIO MOKYMOSI APŽVALGA **8**

1.1. Mokymasis su mokytoju 8

1.1.1. Klasifikavimas 9

1.1.1.1. Klasifikatoriaus validavimas kryžminio patikrinimo metodu 9

1.1.1.2. Klasifikatoriaus validavimas įkelčių metodu 10

1.1.2. Regresinė analizė 11

1.1.3. Atraminių vektorių klasifikatoriai 11

1.1.3.1. Tiesiškai atskiriami duomenys 11

1.1.3.2. Tiesiškai neatskiriami duomenys 12

1.1.4. *Random Forest* klasifikatorius 13

1.2. Mokymasis be mokytojo 14

1.2.1. Klasterizavimas 14

1.3. Mokymosi su mokytoju ir mokymosi be mokytojo palyginimas 15

1.4. Kombinuotasis mokymasis 16

2. PAGRINDINIAI MATŲ ATRINKIMO METODAI **17**

2.1. *Fisher* įvertis 18

2.2. *Relief* metodas 18

2.3. Asimetrinis priklausomybės koeficientas 19

2.4. Absoliučių svorių SVM 19

2.5. Rekursyvus matų eliminavimas pagal SVM 20

3. STABILIŲ MATŲ ATRINKIMO METODAI **21**

3.1. Matų atrinkimo stabilumas 22

3.1.1. Stabilumo vertinimas 22

3.1.2. *Kuncheva* indexas 23

3.1.3. *Jaccard* indeksas 23

3.1.4. *Hamming* atstumas 23

3.2. Svoriais grįstas multikriterinis suliejimas 24

3.3. Reitingais grįstas multikriterinis suliejimas 25

3.4. Svoriais ir reitingais grįstas multikriterinis suliejimas 26

3.5. Multikriterinis rekursyvus matų eliminavimas 27
3.6. Konsensuso grupėmis grįstas stabilių matų atrinkimo metodas 28

4. EKSPERIMENTAI 30
4.1. Eksperimentuose naudoti duomenys 30
4.2. Metodologija 31
4.3. Matų atrinkimo metodų sparta 31
4.4. Klasifikavimo pagal atrinktus matus tikslumas 32
4.5. Matų atrinkimo stabilumas 34

REZULTATAI IR IŠVADOS 37

LITERATŪRA 38

SĄVOKŲ APIBRĖŽIMAI 42

ĮVADAS

Šiame darbe yra nagrinėjama biomedicinoje kaupiamų genetinių daugiamačių duomenų analizės specifika. Duomenys, kurių kiekvienas mėginys aprašomas dideliu skaičiumi matų, yra vadinami daugiamačiais duomenimis. Biomedicininiai duomenys yra specifiški tuo, kad jie turi šimtus kartų daugiau matų nei mėginių. Kadangi mėginio gavimo kaina yra aukšta, turimas mažas mėginių skaičius [PLA09]. 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. Triukšminguose duomenyse, didėjant juos apibūdinančių matų skaičiui, didėja tikimybė, kad bus rasta atsitiktinių priklausomybių. Tai yra pagrindinė priežastis, kodėl biomedicininių duomenų analizės procesas yra sudėtingas.

Problemos apibrėžimas

Biomedicininių duomenų klasifikavimo užduotis yra atskirti sveikųjų pacientų mėginius nuo sergančiųjų. Klasifikavimu siekiama nustatyti, kurie matai, veikdami drauge, geriausiai paaiškina skirtumą tarp ligos paveiktų ir nepaveiktų 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 (arba klases) pagal jų matus [Fis36]. Sukonstruotos funkcijos yra vadinamos klasifikatoriais, o jų konstravimo algoritmai – klasifikavimo algoritmais. Klasifikatoriai paruošiami su turimais mėginiais – treniravimosi duomenimis – ir informacija apie jų būklę (sveikas ar sergantis). Klasifikatoriaus ruošimo procesas yra vadinamas apmokymu. Klasifikatoriai yra validuojami su testavimo duomenimis, o naudojami nustatant nematytų mėginių būklę.

Dirbant su biomediciniais duomenimis dažniausiai turime tik kelias dešimtis mėginių, todėl, norint geriau įvertinti klasifikatoriaus tikslumą, yra naudojami pakartotinio mėginių poaibio atrinkimo (angl. *resampling*) metodai: kryžminio patikrinimo (angl. *cross-validation*) arba įkelčių (angl. *bootstrap*[1]). Šių metodų naudojimas su duomenimis, kurių tikrasis pasiskirstymas nėra žinomas, padeda įvertinti klasifikavimo rezultatų variabilumą

[1]Terminas *bootstrap* „įkelties“ prasme pradėtas naudoti dar Rudolfo Ericho Raspės knygoje „Barono Miunchauzeno nuotykiai“(1785), kurioje Baronas Minchauzenas užkėlė save ant arklio tempdamas į viršų savo batų raištelius (angl. *bootstraps*).

(angl. *variance*) ir sisteminį nuokrypį (angl. *bias*).

Dėl „daugiamatiškumo prakeiksmo“ (angl. *the curse of dimentionality*) didėjant matų 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 (angl. *feature selection*) metodai [GE03]. 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 su jais 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.

Kadangi biomedicininiuose duomenyse reikšmingų matų kiekis tiriamai problemai yra nedidelis, todėl tyrėjams norint geriau suprasti nagrinėjamus biomediciininius duomenis yra svarbu orientuotis į mažesnį matų poaibį, kuris yra svarbus nagrinėjamai problemai. Tokioje situacijoje tampa svarbu, kaip varijuoja atrenkamų matų aibė, kai matų atrinkimas vykdomas su vis kitu mėginių poaibiu. Matai, kurie keičiant mėginių, naudojamų matų atrinkime, poaibį yra vėl ir vėl atrenkami, yra vadinami stabiliais [DK82]. Parametras, parodantis kaip stabiliai yra atrenkami matai, yra vadinamas stabilumu *(angl. robustness)*. Tačiau skirtingi matų atrinkimo metodai tiems patiems mėginiams gali atrinkti skirtingus matus. Taip pat, suskaidžius duomenis į persidengiančius poaibius ir atrinkus tą patį kiekį matų tuo pačiu metodu, gaunamas skirtingas matų poaibis. Matų aibės sumažinimas paspartina biomediciininių duomenų tyrimus – tyrėjams reikia atlikti bandymus su mažesniu mėginių skaičiumi, kuriant mediciininius diagnostikos įrankius, naudojamų matų kiekis įtakoja įrankio kainą. Todėl stabilių matų atrinkimas dirbant su biomedicininiais duomenimis yra svarbus.

Matų atrinkimas yra svarbi biomediciininių 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 tinkamą matų atrinkimo strategiją. Kadangi pačių matų atrinkimo metodų veikimas priklauso nuo

konkrečių duomenų, taip pat matų atrinkimo algoritmą reikia derinti ir prie sprendžiamo uždavinio, tai paties matų atrinkimo metodo pasirinkimas yra sudėtinga užduotis.

Matų atrinkimo stabilumo problemą Yang ir Mao [YM11] siūlė spręsti reitinguojant matus remiantis keliai matų atrinkimo metodais. Galutinis matų reitingo 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 matų atrinkimo metodų kiekis yra ribotas ir skirtingų metodų dažnai negalima vykdyti išskirstytų skaičiavimų aplinkoje. Tai riboja šio metodo pritaikomumą daugiamačių duomenų analizėje.

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

Remiantis Yang, Mao bei Loscalzo darbuose pateiktomis įžvalgomis, bus stengiamasi pasiūlyti tyrimų kryptis, kurios padėtų sukurti metodus, skirtus spręsti stabilių matų atrinkimo problemą. Idėja yra sugrupuoti matus pagal greitą klasterizavimo algoritmą, išrinkti reprezentatyvius matus, transformuoti matų erdvę ir joje vykdyti matų atrinkimą remiantis keletu matų atrinkimo metodų.

Darbo tikslas yra išanalizuoti daugiamačių duomenų klasifikavimo ypatybes. Jam yra keliamos šios užduotys:

1. Susipažinti su pažangiausiais klasifikavimo ir matų atrinkimo metodais;
2. Atlikti matų atrinkimo metodų palyginamuosius eksperimentus;
3. Pasiūlyti kryptis, kaip dabartiniai metodai gali būti patobulinti.

Darbo struktūra

Tolimesnė darbo struktūra tokia: skyriuje nr.1 apžvelgiama su atliekamu tyrimu susijusi mašininio mokymosi teorinė medžiaga; skyriuje nr.2 gilinamasi į pagrindinius matų atrinkimo metodus; skyriuje nr.3 apžvelgiamos matų atrinkimo stabilumą didinančios matų atrinkimo strategijos; skyriuje nr.4 aprašyti atlikti eksperimentai bei jų rezultatai.

1. MAŠININIO MOKYMOSI APŽVALGA

Mašininis mokymasis (angl. *machine learning*) yra dirbtinio intelekto šaka, kurios tyrėjai siekia įgalinti kompiuterius tobulinti savo elgseną (mokytis) empirinių duomenų atžvilgiu [DHS01]. Pagal tai, kokie yra turimi empiriniai duomenys, mašininis mokymasis yra skirstomas į mokymąsi su mokytoju (angl. *supervised learning*) ir mokymąsi be mokytojo (angl. *unsupervised learning*).

Biomedicinos kontekste mašininiu mokymusi siekiama atrasti dėsningumus turimuose duomenyse ir išmokti juos panaudoti. Biomedicininius duomenis sudaro mėginių aibė apibūdinama daugeliu matų. Mėginiai priklauso kelioms grupėms (klasėms), pvz., sergančių ir kontrolinių pacientų grupėms. Mašininio mokymosi su mokytoju tikslas – išmokti funkciją atskiriančią mėginius į grupes ir pritaikyti ją dar nematytiems mėginiams. Mokymosi be mokytojo tikslas – nustatyti mėginių grupes pagal turimus duomenis.

Toliau šiame skyriuje apžvelgiami mašininio mokymosi teorijos pagrindai, kurie yra būtini analizuojant daugiamačius biomedicininius duomenis: mokymasis su mokytoju, mokymasis be mokytojo, atraminių vektorių klasifikatoriai, *Random Forest* klasifikatorius.

1.1. Mokymasis su mokytoju

Žmonės mokosi iš patirties, tačiau, skirtingai nei žmonės, kompiuteriai patirties neturi, todėl kompiuteriai turi mokytis iš patyrimą apibūdinančių duomenų – mokymosi duomenų (angl. *training data*). Mokymosi su mokytoju tikslas yra sukonstruoti funkciją, kuri galėtų būti naudojama nuspėti testavimo duomenų (angl. *testing data*) charakteristikų reikšmes pagal mokymosi duomenis. Mokytojo vaidmenį atlieka mokymosi duomenys, kurių spėjamų charakteristikų reikšmės yra iš anksto žinomos. Pagal tai, kokias charakteristikas bandoma nuspėti mokymasis su mokytoju yra skirstomas į dvi rūšis:

1. Klasifikavimas (angl. *classification*) – pagal mokymosi duomenų nepriklausomus kintamuosius bandoma nuspėti kokybinius (kategorinės reikšmės) priklausomus kintamuosius.

2. Regresinė analizė (angl. *regression*) – pagal mokymosi duomenų nepriklausomus kintamuosius bandoma nuspėti kiekybinius (tolydinės reikšmės) priklausomus kintamuosius.

1.1.1. Klasifikavimas

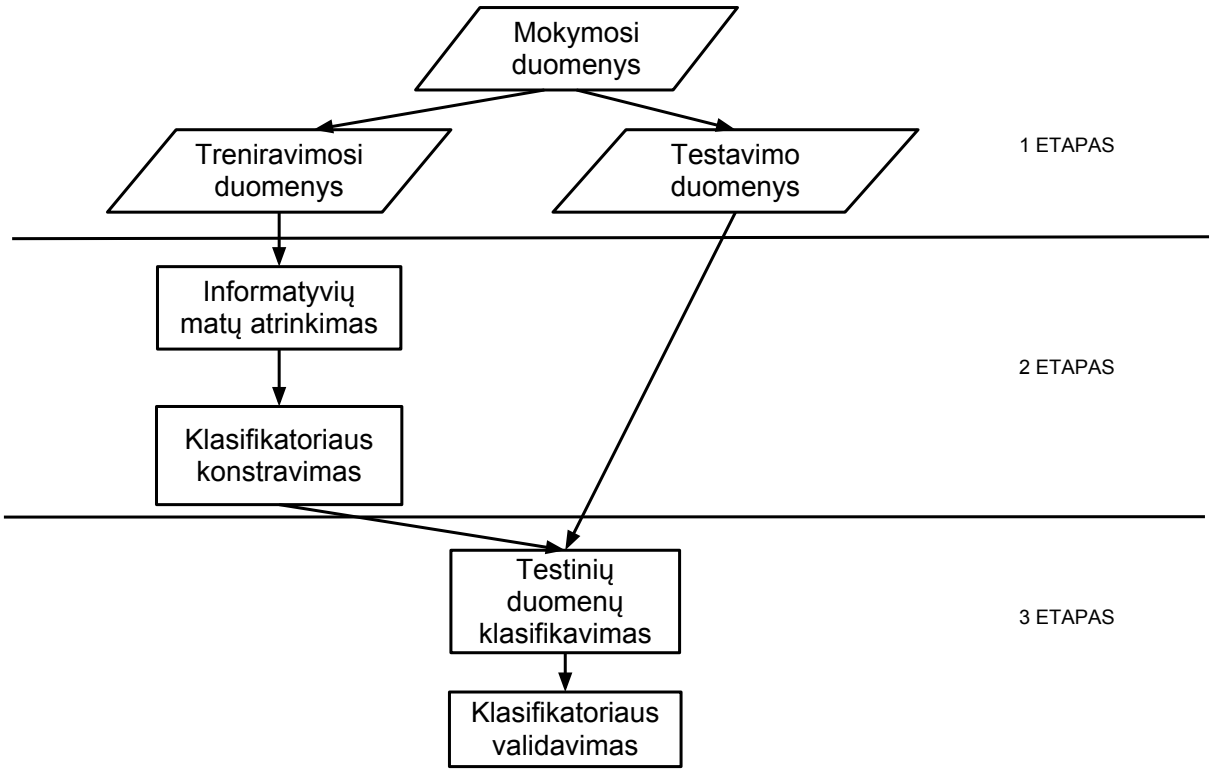
Mašinininio mokymosi kontekste klasifikavimu yra vadinama problema, kai naudojantis mokymosi duomenis reikia išmokti nustatyti, kuriai klasei priklauso mėginys. Klasifikavimo procesas trimis etapais pavaizduoti 1 pav. srautų diagramoje. Klasifikavimo etapai:

1. Visa mokymosi duomenų aibė yra atsitiktinai padalinama į dvi nepersidengiančias aibes: treniravimosi duomenys (pvz., 90% visų mokymosi duomenų) ir testavimo duomenys(pvz., likę 10%);
2. Pagal turimus treniravimosi duomenis yra atrenkami informatyviausi matai, pvz., naudojant *Relief* metodą. Remiantis atrinktaisiais matais yra konstruojamas klasifikatorius - funkcija, pagal kurią nematyti mėginiai bus priskirti vienai iš klasių.
3. Sukonstruotu klasifikatoriumi testavimo duomenys yra suskirstomi į klases. Pagal tai, kiek mėginių klasifikatorius priskyrė teisingoms klasėms yra vertinama klasifikatoriaus tikslumas. Klasifikatorius validuojamas naudojant tokius metodus kaip kryžminis patikrinimas (angl. *cross validation*)

Dirbant su biomedicininiais duomenimis tipinė užduotis yra pagal paciento mėginį apibūdinančius matus sukonstruoti klasifikatorių, kuris bandys nuspėti, kuriai pacientų klasei – sergančiųjų ar sveikųjų – priklauso tiriamasis mėginys. Klasifikavimą galima vertinti pagal:

- Klasifikavimo tikslumą (angl. *accuracy*) – santykį tarp teisingai suklasifikuotų mėginių ir visų mėginių;
- Klasifikavimo nuostolius (angl. *error rate*) – santykį tarp neteisingai suklasifikuotų mėginių ir visų mėginių;
- ROC (angl. *receiver operating characteristic*, ROC) kreivę – grafikas, rodantis klasifikatoriaus jautrumą ir specifiškumą. Grafiko abscisių ašyje atidedamos klasifikatoriaus specifiškumo (angl. *false positive*) reikšmės, o ordinačių ašyje jautrumo (angl. *true positive*) reikšmės.

1.1.1.1. Klasifikatoriaus validavimas kryžminio patikrinimo metodu Naudojant kryžminio patikrinimo (angl. *cross-validation*) metodą, daug kartų sudaromos skirtingos treniravimosi ir testavimo mėginių imtys. Taikant atskirą šio metodo variantą, kryžminį patikrinimą paliekant vieną mėginį (angl. *leave-one-out cross-validation*), iš mokymosi duo-



1 pav.: Klasifikavimo srautų diagrama su paaiškinimais.

menų išimamas vienas (testavimo) mėginys, o su likusiais apmokomas klasifikatorius, kuris klasifikuoja išbrauktąjį mėginį. Procesas tęsiamas tol, kol suklasifikuojami visi mėginiai. Kitais kryžminio patikrinimo metodo variantais iš treniravimosi mėginių yra išimama po keletą mėginių. Pagal tai, kiek testavimo mėginių klasifikatorius priskyrė klaidingai kategorijai, yra nustatoma vidutinė klaidingo klasifikavimo tikimybė. Šiuo metodu gauti įverčiai pasižymi dideliu klasifikavimo rezultatų variabilumu [BND04].

1.1.1.2. Klasifikatoriaus validavimas įkelčių metodu Naudojant įkelčių metodą, iš N dydžio mėginių aibės yra paimama tokio pačio dydžio atsitiktinių mėginių imtis su pasikartojimais, kuri vadinama įkelties treniravimosi imtimi. Į šią imtį nepaimti mėginiai yra priskiriami testavimo imčiai. Naudojant įkelties treniravimosi mėginių imtį yra apmokomas klasifikatorius, kuris klasifikuoja testavimo imtį. Procesą kartojant gaunama klasifikavimo nuostolių įverčių imtis. Šios imties vidurkis yra vidutinis klasifikavimo nuostolio įvertis. Dažniausiai naudojamas „0.623 įkelčių“ (angl. *0.623*[2] *bootstrap*) metodas. Šiuo metodu gautas vidutinio klasifikavimo nuostolio įvertis pasižymi mažu variabilumu [MST94].

[2]0.623 yra tikimybė mėginiui būti įtrauktam į treniravimosi imtį.

1.1.2. Regresinė analizė

Mašininio mokymosi kontekste regresine analize yra vadinama problema, kai pagal patirtį apibūdinančius duomenis reikia nustatyti kiekybines duomenų charakteristikas. Regresinė analizė naudoja standartinius statistinius metodus, tokius kaip mažiausių kvadratų metodas (angl. *least squares*). Regresinė analizė dažniausiai naudojama įvertinti (ang. *forecast*) ateities duomenų vertes bei interpoliacijai – tikėtinos reikšmės tarp keleto taškų įvertinimui.

Dirbant su biomedicininiais duomenimis regresinė analizė taikoma nustatant, pvz., mėginio vėžio stadiją. Šiame darbe pagrindinis dėmesys skiriamas klasifikavimui. Daugiau informacijos apie regresiją galima rasti [GH07].

1.1.3. Atraminių vektorių klasifikatoriai

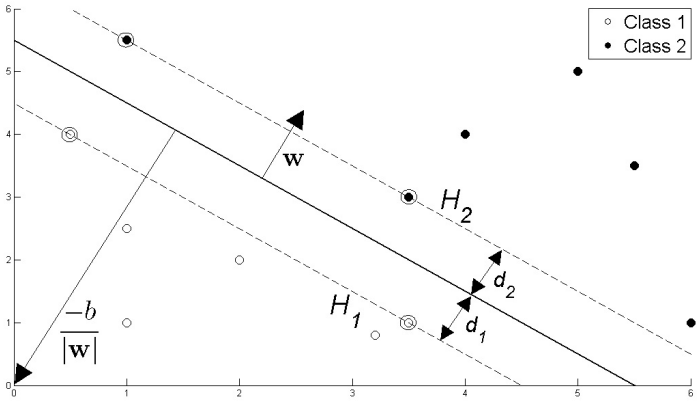
Atraminių vektorių klasifikatoriai (angl. *support vector machines*, SVM) - tai mokymosi su mokytoju algoritmas, kuris gali būti taikomas tiek klasifikavimui, tiek regresinei analizei [Vap00]. Atraminių vektorių klasifikatorių algoritmo tikslas yra mėginių erdvėje orientuoti atskiriančiąją hiperplokštumą, galimai pašalinant triukšmą bei išimtis (angl. *outlier*), tokiu būdu, kad atstumas tarp jos ir atraminių vektorių būtų didžiausias [CV95]. Atraminiais vektoriais (angl. *support vectors*) yra vadinami abiejų klasių mėginiai esantys arčiausiai atskiriančiosios hiperplokštumos (angl. *decision boundary*).

1.1.3.1. Tiesiškai atskiriami duomenys Tarkime, kad turime L mokymosi mėginių, kurių kiekvienas mėginys x_i turi D matų ir priklauso vienai iš dviejų klasių $y_i = -1$ arba $y_i = +1$. Taigi turime mokymosi duomenis, kurių pavidalas yra:

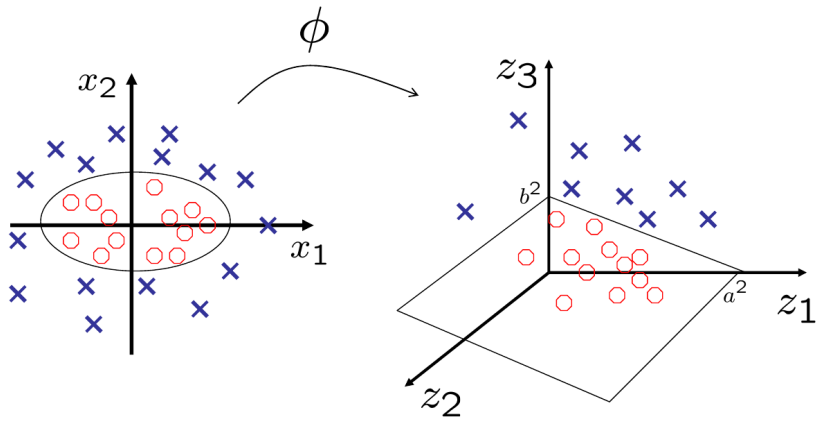
$$\{x_i, y_i\}, kur\ i = 1..L, y_i \in \{-1, 1\}, x \in \Re^D \tag{1}$$

Kai duomenys yra tiesiškai atskiriami, galima nupiešti tiesę plokštumoje x_1 ir x_2, kuri atskiria dvi klases, kai $D = 2$ ir hiperplokštumą erdvėje $x_1, x_2, ...x_D$, kai $D > 2$. Atskiriančioji tiesė pavaizduota 2 pav. Hiperplokštuma apibrėžta $w \cdot x_i + b = 0$, kur w – hiperplokštumos normalės vektorius, $\frac{b}{||w||}$ – statmens einančio nuo hiperplokštumos iki koordinačių pradžios taško ilgis (sisteminis nuokrypis (angl. *bias*)).

Taigi, atraminių vektorių klasifikatorių sukūrimas yra parametrų w ir b, tenkinančių



2 pav.: Atskiriančiosioji tiesė nubrėžta plokštumoje, kurioje pavaizduoti tiesiškai atskiriami duomenys [Fle09].



3 pav.: Tiesiškai neatskiriamos matų erdvės transformavimas į tiesiškai atskiriamą [Alb12].

minėtas sąlygas, radimas. Klasifikavimo taisyklę f testavimo duomenims X apibrėžiama formule:

$$f(X) = sign(w^T X + b), \tag{2}$$

kur w^T yra transponuotas matų svorių vektorius.

1.1.3.2. Tiesiškai neatskiriami duomenys Kai duomenys yra tiesiškai neatskiriami, matų erdvės transformavimas gali padėti duomenis padaryti tiesiškai atskiriamais. Branduolio metodai (angl. *kernel methods*) transformuoja matų erdvę taip, kad jie tampa tiesiškai atskiriami [ABR64]. Tokios transformacijos pavyzdys pateiktas 3 pav.

Branduoliu (angl. *kernel*) vadinama funkcija:

$$K(X, Y) = \langle \phi(X)^T \cdot \phi(Y) \rangle \tag{3}$$

Branduolio funkcija naudojama tiesiškai neatskiriamų duomenų erdvės projektavimui į kitą erdvę, kurioje duomenys būtų tiesiškai atskiriami. Branduolio savybė, kai jį galima išreikšti funkcija nuo pradinių kintamųjų skaliarinės sandaugos, vadinama „branduolio gudrybe“ (angl. *kernel trick*). „Branduolio gudrybė“ dažniausiai naudojama norint tiesinį metodą paversti netiesiniu.

Pavyzdžiui, turime mėginius $X = (x_1, x_2)$, $Y = (y_1, y_2)$, branduolys yra $K(X, Y) = \langle X \cdot Y \rangle^2$, o $\phi(X) = \phi(x_1, x_2) = (x_1^2, \sqrt{2}x_1x_2, x_2^2)^2$ transformuoja matų erdvę iš R^2 į R^3. Branduolį taikome mėginiams $K(X, Y) = \langle X \cdot Y \rangle^2 = \langle x_1y_1 + x_2y_2 \rangle^2 = \langle x_1^2y_1^2 + 2x_1y_1x_2y_2 + x_2^2y_2^2 \rangle = \langle (x_1^2, \sqrt{2}x_1x_2, x_2^2) \cdot (y_1^2, \sqrt{2}y_1y_2, y_2^2) \rangle = \langle \phi(X) \cdot \phi(Y) \rangle$.

Atraminių vektorių klasifikatorių algoritmo naudojimas dirbant su biomedicininiais duomenimis populiarus, nes jis demonstruoja gerus rezultatus, kai turima maža daugiamačių mokymosi duomenų aibė.

1.1.4. *Random Forest* klasifikatorius

Random Forrest klasifikatorius yra kombinuotojo mokymosi algoritmas, kuris sukuria keletą klasifikavimo medžių (angl. *decision tree*) [HK00], kurie nepriklausomai klasifikuoja mėginius, ir daugumos balsavimo (angl. *majority voting*) būdu yra skelbiamas galutinis klasifikavimo rezultatas [BFOS84]. Kiekvienas klasifikavimo medis yra konstruojamas pagal algoritmą nr. 1.

Algoritmas nr. 1 *Random Forest* klasifikavimo medžių konstravimas

1. Turima N mėginių, kurie turi M matų;
2. Pasirenkamas m matų, kurie bus naudojami klasifikavimo medžių kūrimui; $m << M$;
3. Sudaroma treniravimosi mėginių aibė, n kartų pasirenkant mėginius su pasikartojimais iš visų N mėginių. Visi nepasirinkti mėginiai paliekami klasifikatoriaus testavimui;
4. Kiekvienam medžio mazgui atsitiktinai pasirenkama m matų, kuries sudarys sąlygą tam mazgui. Randamas geriausia atskyrimo sąlyga treniravimos duomenims pagal tuos m matų;
5. Pilnai užauginti medžiai nėra genėjami (angl. *pruning*) taip stengiantis išlaikyti žemą sisteminį nuokrypį.

Random forest algoritmo tikslumas priklauso nuo koreliacijos tarp sukurtų klasifikavimo medžių ir atskirų klasifikavimo medžių skiriamosios galios. Didesnė koreliacija lemia mažesnį klasifikavimo tikslumą, o kuo didesnė atskiro klasifikavimo medžio skiriamoji galia, tuo geresnis klasifikavimo tikslumas.

Randon forest klasifikatoriai yra tikslūs, greiti, bei sugeba išvengti persimokymo (angl. *overfitting*). Šios trys klasifikavimo algoritmo savybės yra labai svarbios dirbant su biomedicininiais duomenimis.

1.2. Mokymasis be mokytojo

Mašininio mokymosi kontekste dažnai sutinkamas uždavinys yra į prasmingas grupes sugrupuoti turimus duomenis, kurių grupavimas iš anksto nėra žinomas. Mokymosi be mokytojo metodų pagrindinis principas – sugrupuoti duomenis taip, kad vienoje grupėje esantys mėginiai būtų kuo panašesni, o grupės tarpusavyje nepanašios.

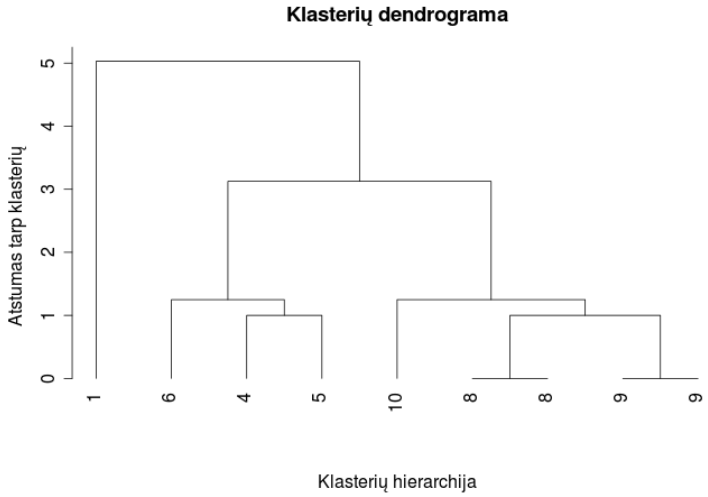
Mokymosi su mokytoju metu galima išmatuoti gautos funkcijos tikslumą įvairiais metodais, pvz., kryžminiu patikrinimu. Mokymosi be mokytojo proceso rezultato tiesioginio patikrinimo procedūrų nėra, yra tik įvairių sudarytų grupių – klasterių – kokybės įvertinimo metodų (angl. *cluster validity methods*) [HBV02], pvz., *Davies-Bouldin* indeksas. Dėl to yra sunkiau išsiaiškinti rezultatų, gautų pagal mokymosi be mokytojo algoritmų darbo rezultatus, patikimumą.

1.2.1. Klasterizavimas

Klasterizavimas yra viena iš mokymosi be mokytojo algoritmų rūšių. Klasterizavimas – tai turimų mėginių suskirstymas į klasterius taip, kad klasterio viduje esantys mėginiai būtų kuo panašesni tarpusavyje, o mėginiai iš skirtingų klasterių būtų kiek įmanoma skirtingesni. Klasterizavimu siekiama atrasti nežinomas struktūras turimuose duomenyse.

Klasterizavimo algoritmuose yra matuojamas mėginių panašumas. Panašumui matuoti yra naudojamos atstumo tarp mėginių metrikos, tokios kaip *Manhattan*, Euklido, *Mahalanobis* atstumai. Pasirinktosios atstumo metrikos rezultatai priklauso nuo to, kokioje skalėje yra atlikti paskirų matų matavimai. Todėl rekomenduojama prieš klasterizavimą visus matus normalizuoti. Dažniausiai naudojami normalizavimo parametrai: vidurkis lygus 0, standartinis nuokrypis – 1 matavimo vienetas (angl. *unit*). Normalizavimu siekiame apsisaugoti nuo situacijos, kai matas su didelėmis skaitinėmis reikšmėmis gali iškreipti atstumo matavimus.

Dirbant su biomedicininiais duomenimis klasterizavimo algoritmus galime panaudoti panašių matų sugrupavimui. Iš panašių matų grupės pasirinkus tik vieną reprezentatyviausią matą, būtų galima sumažinti bendrą matų skaičių. Toks matų skaičiaus sumažinimas



4 pav.: Hierarchinio klasterizavimo rezultatų grafinis pavyzdys.

pagerintų matų atrinkimo procesą.

Hierarchinis klasterizavimas (angl. *hierarchical clustering*) yra klasterizavimo algoritmas, kuris arba visą duomenų aibę panariui skaido į vis mažesnius klasterius (angl. *divisive clustering*), arba pradeda nuo klasterių sudarytų tik iš vieno mėginio ir kiekvienoje iteracijoje sujungia panašiausius klasterius (angl. *agglomerative clustering*) [HK00]. Hierarchinio klasterizavimo rezultatas – klasterių medis, dendrograma, rodanti, kaip klasteriai yra hierarchiškai susiję. Pasirinktame lygyje nupjovus dendrogramą gaunama klasterizavimo struktūra [Mar08]. Klasterių dendrogramos pavyzdys yra pateiktas 4 pav. Hierarchinis klasterizavimas yra informatyvesnis nei paprastas – plokščias – klasterizavimas. Tačiau šių algoritmų sudėtingumas didesnis nei, pvz., tankiu grįstų algoritmų [HK00].

1.3. Mokymosi su mokytoju ir mokymosi be mokytojo palyginimas

Mokymosi su ir be mokytojo procesai panašūs savo esme – siekia išgauti žinias apie turimus duomenis, tačiau jų panaudojimas skiriasi iš esmės:

- Mokymosi duomenys – mokymosi su mokytoju proceso įeities duomenyse yra išreikštinai pasakyta, kokio rezultato mes laukiame, o mokymosi be mokytojo įeities duomenyse tokios papildomos informacijos nėra.
- Naudojimo tikslai – mokymasis su mokytoju siekia iš pavyzdžių išmokti vertinti naujus duomenis, o mokymasis be mokytojo siekia atrasti vidinę duomenų struktūrą.

1.4. Kombinuotasis mokymasis

Kombinuotasis mokymasis (angl. *ensemble learning*) - tai toks mašininis mokymasis, kai problemos sprendimui yra kombinuojami keli mašininio mokymos metodai. Pristatant kombinuotojo mokymosi principus naudojama klasifikatoriaus sąvoka, tačiau principai galioja ir kitiems mašininio mokymosi metodams, pvz., matų atrinkimui ir klasifikavimui. Kombinuotojo mokymosi metodai turi sugebėti konstruoti pavienius klasifikatorius ir kombinuoti tų klasifikatorių gautus rezultatus. Yra keletas kombinuotojo mokymosi metodų, kurių populiariausi yra *bagging* [Bre96], *boosting* [Sch03], *stacking* [Wol92], etc.

Kombinuotasis mokymasis pirmiausia yra naudojamas tam, kad pagerintų kuriamo klasifikatoriaus tikslumą arba sumažintų prasto klasifikatoriaus sukūrimo tikimybę, nes dažniausiai nėra žinoma, kuris klasifikavimo algoritmas ir kokia jo konfigūracija geriausiai tinka nagrinėjamai problematikai. Tipinės situacijos, kai kombinuotojo mokymosi metodų taikymas pasiteisina [Pol06]:

1. Statistinis reikšmingumas – imant keleto klasifikatorių vidutinį rezultatą, yra sumažinama rizika pasirinkti netinkamą klasifikatorių.

2. Dideli duomenų kiekiai – kartais duomenų būna tiek, kad juos apdoroti vienam klasifikatoriui yra labai neefektyvu, o kombinuojant keletą klasifikatorių pasiekiamas geresnis efektyvumas.

3. Mažas mėginių skaičius – daugelio klasifikatorių apmokymas su vis kitu mėginių poaibiu padeda tiksliau nuspėti tikrąją duomenų funkciją.

4. Skaldyk ir valdyk – kai kurie duomenys yra per sudėtingi, kad jie būtų išmokti vienu klasifikatoriumi. Tačiau kombinuojant keletą klasifikatorių galima išmokti net ir labai komplikuotus duomenis.

5. Duomenų suliejimas – kartais duomenys gaunami iš keleto šaltinių, vienam klasifikatoriui apdoroti tokius duomenimis gali būti per sunku, todėl pasiteisina skirtingų klasifikatorių naudojimas atskiriems duomenų rinkiniams, kombinuojant jų rezultatus.

Kombinuotasis mokymasis gali būti palygintas su keleto ekspertų apklausa, todėl kombinuotojo mokymosi metodų naudojimas sudėtingose problemose, pvz., darbas su biomedicininiais duomenimis, pasiteisina.

2. PAGRINDINIAI MATŲ ATRINKIMO METODAI

Dėl „daugiamatiškumo prakeiksmo“ (angl. *the curse of dimentionality*) didėjant matų 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 dėl to, kad duomenys yra daugiamačiai. Todėl biomediciniinių duomenų daugiamatiškumui sumažinti yra naudojami informatyviausių matų atrinkimo metodai [GE03] (angl. *feature selection*). 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. Kadangi matų atrinkimo rezultatai priklauso nuo konkrečių duomenų, todėl svarbu yra pasirinkti tinkamiausią matų atrinkimo strategiją

Pagal tai, kaip matų atrinkimo metodai yra susiję su klasifikatoriumi, matų atrinkimo metodus galima skirstyti į tris kategorijas [SAVdP08]:

- Filtruojantys metodai (angl. *filter methods*), pvz. *Fisher* įvertis. Jie dirba tiesiogiai su duomenimis, o jų darbo rezultatas gali būti matų įvertinimas svoriais, matų reitingavimas ar tiesiog geriausių matų poaibis, kuriuo remiantis vėliau apmokomas klasifikatorius. Tokių metodų pagrindinis privalumas yra tai, kad jie yra greiti, tinka paskirstytų skaičiavimų aplinkoms ir nepriklausomi nuo klasifikavimo metodo, tačiau remiantis atrinktaisiais matais nebūtinai bus sukurtas geriausias klasifikatorius.

- Prisitaikantieji metodai (angl. *wrapper methods*). Pirma, apmokomas klasifikatorius su visais matais, antra, parenkamas matų poaibis ir apmokomas klasifikatorius. Po daugkartinio matų poaibių įvertinimo pagal klasifikavimo rezultatus yra nusprendžiama, kuris matų poaibis yra labiausiai tinkamas klasifikavimui. Įterptinių metodų atveju matų atrinkimo procesas yra neatsiejamas nuo klasifikavimo proceso – matai yra atrenkami pagal klasifikatoriaus darbo rezultatus. Prisitaikantieji metodai dažnai duoda geresnius rezultatus negu filtravimo metodai, bet yra reiklūs resursams.

- Įterptiniai metodai (angl. *embedded methods*), pvz. AW-SVM[Vap00]. Jie matų atrinkimui naudoja vidinius klasifikatoriaus duomenis (pvz. matų svoriai gauti pagal atraminių vektorių klasifikatorius). Šie metodai dažnai siūlo gerą santykį tarp klasifikavimo tikslumo ir skaičiavimų sudėtingumo.

Šiame skyriuje yra nagrinėjai pagrindiniai matų atrinkimo metodai: *Fisher* įvertis,

Relief metodas, asimetrinis priklausomybės koeficientas, absoliučių svorių SVM, rekursyvus matų eliminavimas pagal SVM.

2.1. *Fisher* įvertis

Fisher įvertis (angl. *Fisher ratio*) vertina individualius matus pagal matų klasių atskiriamąją galią [PWCG01]. Mato įvertis yra sudarytas iš atskirų klasių vidurkių skirtumo santykio su klasių standartinių nuokrypių suma:

$$FR(j) = \frac{(\mu_{j1} - \mu_{j2})^2}{\sigma_{j1}^2 + \sigma_{j2}^2}, \tag{4}$$

kur, j – mato indeksas, μ_{jc} – mato j reikšmių vidurkis klasėje c, σ_{jc}^2 – mato j reikšmių standartinis nuokrypis klasėje c, kur $c = 1, 2$. Kuo didesnis yra *Fisher* įvertis, tuo geriau tas matas atskiria klases. Šis metodas neįvertina matų tarpusavio sąveikų.

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 klasių, ir atnaujinamos visų matų „susietumo“ reikšmės [RSK03]. 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:

$$W(j) = W(j) - \frac{diff(j, x, x_H)}{n} + \frac{diff(i, x, x_M)}{n}, \tag{5}$$

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 atsitiktinai pasirinkto mėginio x ir atitinkamo jo kaimyno. Skirtumą į intervalą $[0, 1]$ normalizuojanti funkcija yra:

$$diff(j, x, x') = \frac{|x_j - x'_j|}{x_{j_{max}} - x_{i_{min}}}, \tag{6}$$

kur $x_{j_{max}}$ ir $x_{j_{min}}$ yra maksimali ir minimali j-ojo mato 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].

2.3. Asimetrinis priklausomybės koeficientas

Asimetrinis priklausomybės koeficientas (angl. *Asymetric Dependency Coefficient*, ADC) yra matų reitingavimo motodas, kuris matuoja mėginio klasės tikimybinę priklausomybę nuo j-ojo mato, naudodamas informacijos prieaugio metriką (angl. *information gain*) [Sha01]:

$$ADC(Y, j) = \frac{MI(Y, X_j)}{H(Y)}, \tag{7}$$

kur $H(Y)$ – klasės Y entropija (angl. *entropy*), o $MI(Y, X_j)$ – yra tarpusavio informacija (angl. *mutual information*) tarp mėginio klasės Y ir j-ojo mato [Ken83].

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

$$H(X_j) = -\sum_{x} p(X_j = x) logp(X_j = x), \tag{9}$$

$$MI(Y, X_j) = H(Y) + H(X_j) - H(Y, X_j), \tag{10}$$

$$H(Y, X_j) = -\sum_{y, x_j} p(y, x_j) logp(y, x_j), \tag{11}$$

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

2.4. Absoliučių svorių SVM

Viena priežasčių, kodėl atraminių vektorių klasifikatoriai (SVM) yra vienas populiariausių klasifikavimo algortimų, yra tai, kad jis gerai susidoroja su daugiamačiais duomenimis [GWBV02]. Yra keletas bazinių SVM variantų, bet šiame darbe naudojamas tiesinis SVM, nes jis demonstruoja gerus rezultatus analizuojant genų ekspresijos duomenimis [Vap00].

Tiesinis SVM yra hiperplokštuma apibrėžta kaip:

$$\sum_{j=1}^{p} w_j x_j + b_0 = 0, \tag{12}$$

kur p – matų kiekis, w_j – j-ojo mato svoris, x_j – j-ojo mato kintamasis, b_0 – sisteminis nuokrypis. 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. Todėl metodas ir vadinamas absoliučių svorių SVM (angl. *Absolute Weight SVM*, AW-SVM). Pastebėtina, kad svorių nustatymas yra atliekamas tik vieną kartą (SVM-RFE matų atrinkimo metodas svorius matams nustato daug kartų).

2.5. Rekursyvus matų eliminavimas pagal SVM

Rekursyvus matų eliminavimas pagal SVM (angl. *Support Vector Machines – Recursive Feature Elimination*, SVM-RFE) yra vienas populiariausių matų atrinkimo algoritmų [GWBV02]. Todėl, jis yra naudojamas kaip atskaitos taškas (angl. *benchmark*) vertinant kitus matų atrankimo metodus. Iš esmės šis metodas yra daugkartinis absoliučių svorių SVM metodo taikymas nuolat išmetinėjant matus su mažiausiais svoriais. Rekursyvus matų eliminavimas mums padeda surasti klasifikavimui tinkamą matų poaibį, kas ne visada pavyksta su matų reitingavimo metodais. Rekursyvus matų eliminavimas yra aprašytas algoritme nr. 2. Jei trečiajame algoritmo žingsnyje iš matų aibės yra pašalinamas tik viena matas, tai

Algoritmas nr. 2 Rekursyvus matų eliminavimas

1. Turime pilną matų rinkinį F_0, nustatome $i = 0$;
2. Įvertiname kiekvieno mato kokybę matų aibėje F_i;
3. Išmetame mažiausiai kokybišką matą iš F_I tam, kad gautume matų rinkinį F_{i+1};
4. Nustatome $i = i+1$ ir grįžtame į antrąjį žingsnį kol nėra patenkinta algoritmo pabaigos sąlyga.

gaunamas matų reitingavimą, o jei pašalinamas fiksuotas skaičius ar dalis (pvz. 50%) matų, tai matų reitingavimas negaunamas. Pastebėtina, kad rekursyvus matų eliminavimas labai padidina algoritmo sudėtingumą. Algoritmo pabaigos sąlyga gali būti koks nors konkretus matų skaičius arba tiesiog matų aibę mažinama tol, kol visi matai yra išmetami.

3. STABILIŲ MATŲ ATRINKIMO METODAI

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

1. Duomenys yra triukšmingi ir kai kurie matai gali būti palaikyti informatyviais dėl atsitiktinių priežasčių;
2. Daugiamačiuose duomenyse dalis matų koreliuoja, 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.

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. Kuo triukšmingesni duomenys, kuo mažiau turima mėginių ir kuo daugiau yra matų, tuo ryškesnė yra ši problema [LYD09].

Stabilių matų atrinkimo problematika yra populiarėjanti tyrimų kryptis. Stabilumas aktualus, nes biomedicininiuose duomenyse konkrečiai problemai aktualūs yra tik tam tikri matai. Todėl dalykinės srities ekspertams yra svarbu naudoti tuos matų atrinkimo metodus, kurie atrenka stabilius ir susijusius su analizuojama problema matus.

Vienas iš būdų didinti matų atrinkimo stabilumą galėtų būti multikriterinių matų atrinkimo metodų naudojimas. Multikriterinių matų atrinkimo metodų esmė yra panaudoti kelis matų atrinkimo metodus suliejant jų rezultatus į vieną bendrą rezultatą. Yra skiriamos trys priežastys, kodėl keletas agreguotų silpnų ir nestabilių matų atrinkimo metodų gali duoti stabilesnius matų atrinkimo rezultatus [Die00]:

1. Keletas skirtingų, bet vienodai gerų hipotezių gali būti teisingos, todėl kriterijų kombinavimas sumažiną tikimybę, kad bus pasirinkta neteisinga hipotezė;
2. Atskiri matų atrinkimo metodai gali dirbti skirtinguose lokaliuose optimumuose, o kombinavimas gali geriau reprezentuoti tikrąją duomenis generuojančią funkciją;

3. Kombinuojant keletą kriterijų praplečiama galimų hipotezių erdvė. Todėl, jei pavieniai pagal pavienius kriterijus nerandama teisinga hipotezė, tai kombinuojant keletą kriterijų didėja tikimybė pasirinkti teisingą hipotezę.

Suliejant keletą skirtingų matų atrinkimo metodų rezultatų suliejamos gerosios pavienių matų atrinkimo metodų savybės, taip kompensuojant metodų silpnybes.

Šiame skyriuje aptariama stabilumo matavimų problematika bei matų atrinkimo stabilumą didinantys metodai: svoriais grįstas multikriterinis suliejimas, reitingais grįstas multikriterinis suliejimas, svoriais ir reitingais grįstas multikriterinis suliejimas, multikriterinis rekursyvus matų eliminavimas, konsensuso grupėmis grįstas stabilių matų atrinkimo metodas.

3.1. Matų atrinkimo stabilumas

Matų atrinkimo metodų stabilumas gali būti apibrėžtas kaip matų atrinkimo rezultatų variacijų lygis dėl mažų pakeitimų duomenų rinkinyje. Pakeitimai duomenų rinkinyje gali būti mėginių lygio (pvz. mėginiai pridedami arba atimami), matų lygio (pvz. pridedant matams triukšmo) ar abiejų lygių kombinacija.

Svarbu tai, kad matų stabilumas nėra matuojamas nepriklausomai – jis yra matuojamas atsižvelgiant į klasifikavimo rezultatus. Matuoti stabilumą verta tada, kai pagal atrinktus matus sukuriamas tikslus klasifikatorius.

3.1.1. Stabilumo vertinimas

Vertinant matų atrinkimo metodų stabilumą yra svarbu kaip panašiai yra atrenkami matai, kai yra atliekamas matų atrinkimas su vis kitu mėginių ar matų poaibiu. Kuo mažiau skiriasi atrinktoji matų aibė darant pakeitimus duomenyse, tuo matų atrinkimo stabilumas yra didesnis. Vidutinis matų atrinkimo stabilumas gali būti apibrėžtas kaip vidurkis visų reitingavimo metu gautų atrinktų matų poaibių porų tarpusavio panašumo įverčių [KPH07]:

$$S_{tot} = \frac{2\sum_{i=1}^{k-1}\sum_{j=i+1}^{k} S(f_i, f_j)}{k * (k-1)}, \tag{13}$$

kur k žymi kiek kartų buvo imtas skirtingas mėginių poaibis matų atrinkimui, f_i, f_j – matų atrinkimo rezultatas – reitingai, $S(f_i, f_j)$ – aibių panašumo įvertinimo funkcija.

Matų atrinkimo stabilumo įvertis priklauso nuo to, kokią aibių panašumo funkciją naudosime. Tradicinės panašumo įvertinimo funkcijos (persidengimo procentas, *Pearson* koreliacija, *Spearman* koreliacija) yra linkusios priskirti didesnes panašumo reikšmes, kai pasirenkamas didesnis matų poaibis, nes imant didesnį matų poaibį padidėja tikimybė tiesiog atsitiktinai pasirinkti matą.

3.1.2. *Kuncheva* indexas

Kuncheva indexas [Kun07] yra funkcija skirta matuoti aibių panašumui. Ši funkcija gerai tinka matuoti matų atrinkimo atabilumą, nes atsižvelgia į paimto matų poaibio dydį. *Kuncheva* indeksas:

$$KI(f_i, f_j) = \frac{r * N - s^2}{s * (N - s)} = \frac{r - (s^2/N)}{s - (s^2/N)}, \tag{14}$$

kur $s = |f_i| = |f_j|$ yra atrinktų matų aibės dydis, $r = |f_i \bigcap f_j|$ - abiems atrinktiems matų poaibiams bendrų matų skaičius, N - bendras duomenų aibės matų skaičius. Pastebėtina, kad formulėje esantis atėminys s^2/N ištaiso sisteminį nuokrypį atsirandantį dėl galimybės atsitiktinai pasirinkti matus.

Kuncheva indeksas gali įgyti reikšmes iš intervalo $[-1, 1]$, kur didesnė reikšmė reiškia didesnį panašumą, o artimos nuliui reikšmės reiškia, kad matai atrenkami beveik atsitiktinai.

3.1.3. *Jaccard* indeksas

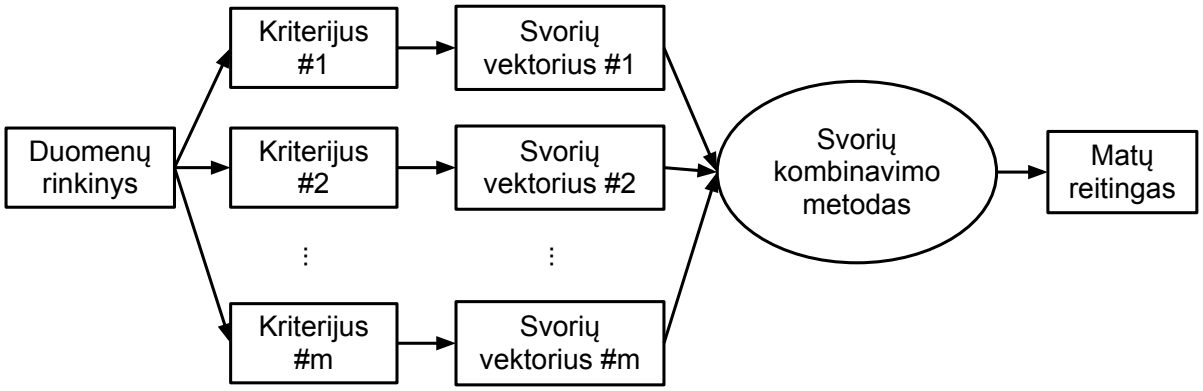
Vienas paprasčiausi aibių panašumo įverčių yra *Jaccard* indeksas [Jac01]. *Jaccard* indexas yra santykis tarp aibių sankirtos ir aibių sąjungos:

$$JI(f_i, f_j) = \frac{|f_i \bigcap f_j|}{|f_i \cup f_j|} = \frac{\sum_l I(f_i^l = f_j^l = 1)}{\sum_l I f_i^l + f_j^l > 0)}, \tag{15}$$

kur f_i ir f_j yra matų reitingai, $I(x)$ - funkcija grąžinanti 1, jei $x = TRUE$, ir 0 kitu atveju.

3.1.4. *Hamming* atstumas

Informacijos teorijoje *Hamming* atstumas [Ham50] tarp dviejų vienodo ilgio vektorių yra apibrėžtas kaip pozicijų skaičius, kuriose esantys simboliai nesutampa. *Hamming* atstu-



5 pav.: Svoriais grįstas multikriterinis suliejimas.

mas yra minimalus skaičius pakeitimų, kad vieną vektorių padarytume lygų kitam.

$$HD(X,Y) = \sum_{i=1}^{n}(x_i \oplus y_i), \tag{16}$$

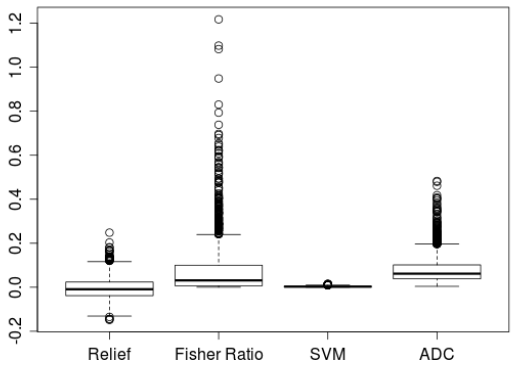
kur $\oplus$ - sumos moduliu 2 arba XOR operacija.

Šiuo metodu matuojant matų atrinkimo stabilumą, prieš atstumo matavimą reikia atlikti rezultatų pertvarkymą: pirma, iš atrinktų matų vektorių padaryti bendro matų skaičiaus ilgio binarinius vektorius; antra, vektoriaus elementus, kurių indeksai gaunami matų atrinkimo metodu, nustatyti lygius vienetui. Atlikus tokius pertvarkymus galima matuoti atstumą tarp dviejų matų atrinkimo rezultatų.

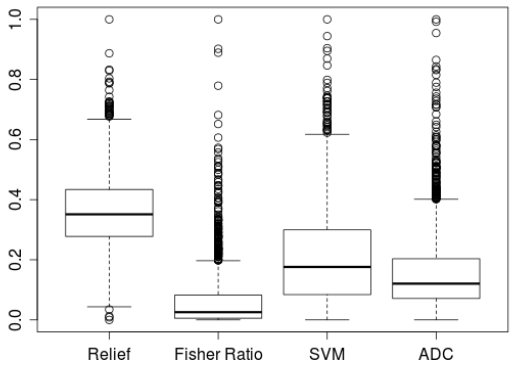
3.2. Svoriais grįstas multikriterinis suliejimas

Svoriais grįsto multikriterinio matų atrinkimo suliejimo pagal svorius algoritmo pirmajame žingsnyje kiekvienas bazinis metodas priskiria duomenų rinkinio matams svorius, tada tie svoriai yra kombinuojami į vieną sutarties (angl. *consensus*) svorių vektorių, kurio pagrindu yra gaunami matų reitingai. Algoritmas yra pavaizduotas 5 pav.

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 svoriai yra nepalyginami. Paveikslėlyje 6 pav. nenormalizuotų pavienių matų vertinimo metodų skiriasi suteiktų svorių intervalai. Paveikslėlyje 7 pav. pavaizduota, kad normalizavus matų svorius pastebimai skiriasi svorių kvartiliai – į tai reikia atkreipti dėmesį interpretuojant galutinius matų vertinimo rezultatus. Svorių



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



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

normalizavimas į intervalą $[0, 1]$ atliekamas pagal formulę:

$$u'_i = \frac{u_i - u_{i_{min}}}{u_{i_{max}} - u_{i_{min}}}, \tag{17}$$

kur u_i – matų svorių vektorius pagal i kriterijų, $u_{i_{min}}$ – minimali u_i svorių vektoriaus reikšmė, $u_{i_{max}}$ – maksimali u_i svorių vektoriaus reikšmė, u'_i – normalizuotų svorių vektorius.

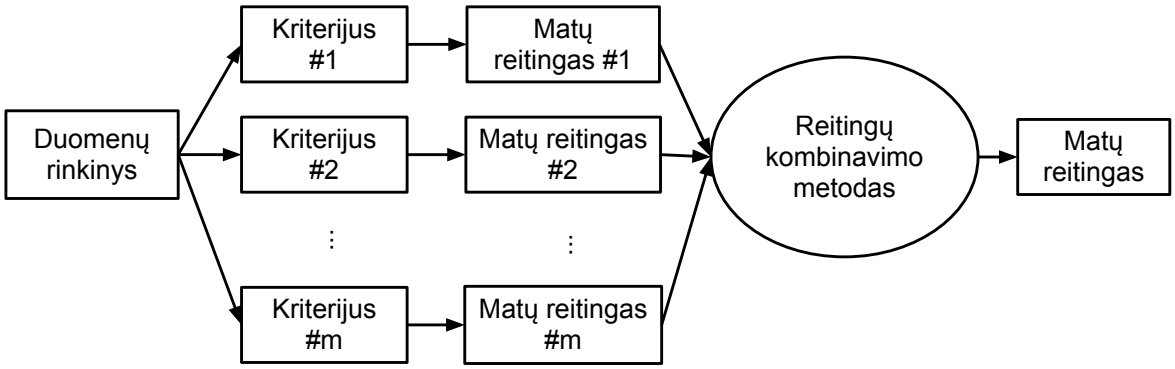
Sutarties svorių vektorius u yra vidurkis normalizuotų svorių vektorių:

$$u = \frac{1}{m} \sum_{i=1}^{m} u'_i, \tag{18}$$

kur m yra bazinių kriterijų skaičius. Didesnė svorio reikšmė reiškia, kad matas yra reikšmingesnis klasifikavimui.

3.3. 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 8 pav. Suliejimo pagal reitingus metodas nereikalauja matų atrinkimo metodų rezultatų normalizavimo, todėl galima matams priskirtus reitingus kombinuoti iškart. Skirtingai nei suliejimo pagal svorius algoritme, baziniai matų atrinkimo kriterijai turi gražinti matų reitingus, o ne svorius.



8 pav.: Reitingais grįstas multikriterinis suliejimas.

Matų reitingų kombinavimui yra keletas metodų[DKNS01], tačiau dėl paprastumo naudojamas *Borda* balsavimas[3] (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 taškus

$$v = \sum_{i=1}^{m} v_i, \tag{19}$$

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

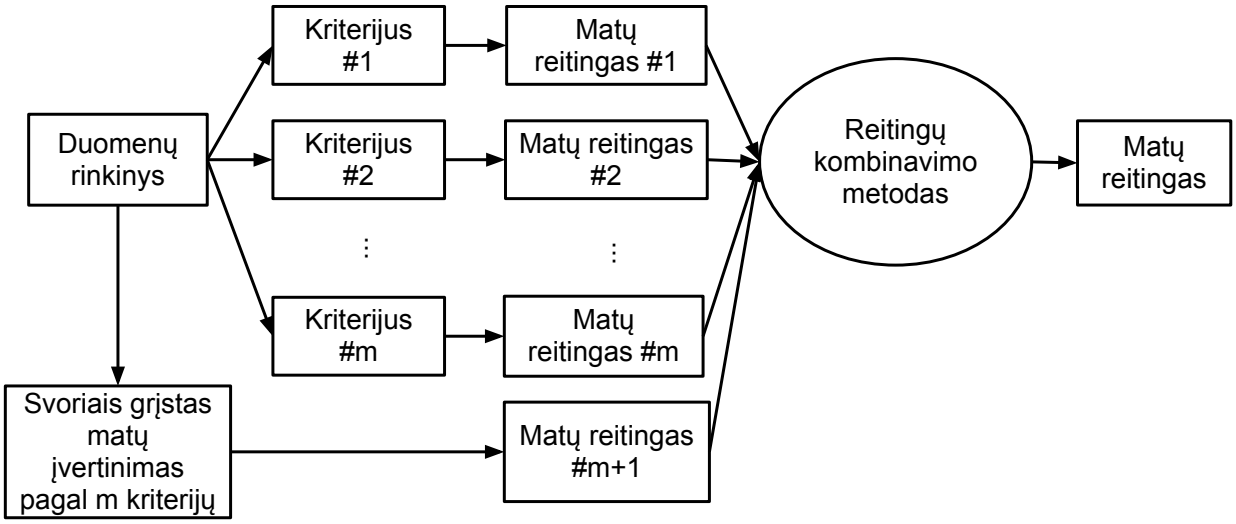
3.4. Svoriais ir reitingais grįstas multikriterinis suliejimas

Svoriais ir reitingais grįsto multikriterinio suliejimo metodas nuo reitingais grįsto multikriterinio suliejimo metodo skiriasi tuo, kad kaip dar vienas matų reitingas yra panaudojamas svoriais grįsto multikriterinio matų atrinkimo metu gautas reitingas. Algoritmas pavaizduotas 9 pav. Multikriterinis matų atrinkimas pagal svorius ir pagal reitingus atliekamas trimis etapais:

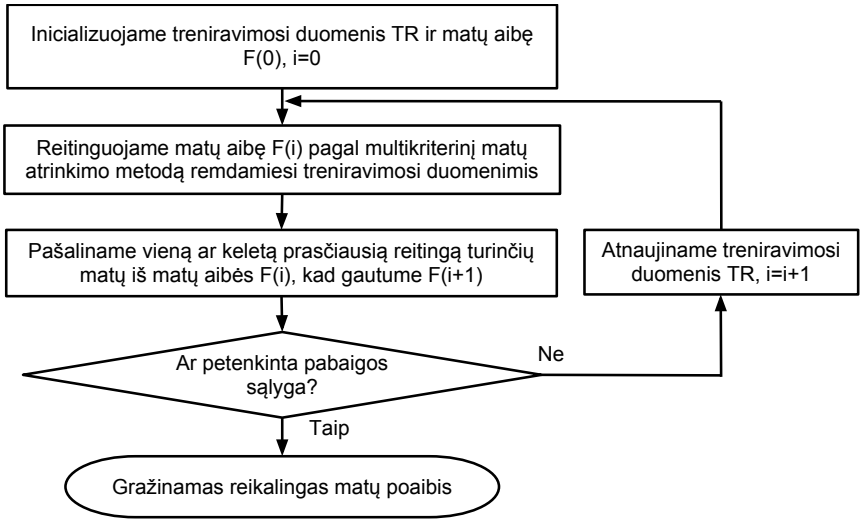
1. Gaunamas matų reitingus pagal m pavienių matų atrinkimo motodų;
2. Suliejamas matų įverčius pagal svorius, taip gaunamas vieną matų reitingą;
3. Reitinguojami matai pagal visus turimus $m+1$ pavienius reitingus.

Suliejant keletą silpnai koreliuojančių matų reitingavimo metodų rezultatų, yra siekiama didesnio matų atrinkimo stabilumo, kai varijuoja treniravimosi duomenų poaibis (angl. *subsampling*) [YM11].

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



9 pav.: Svoriais ir reitingais grįstas multikriterinis suliejimas.

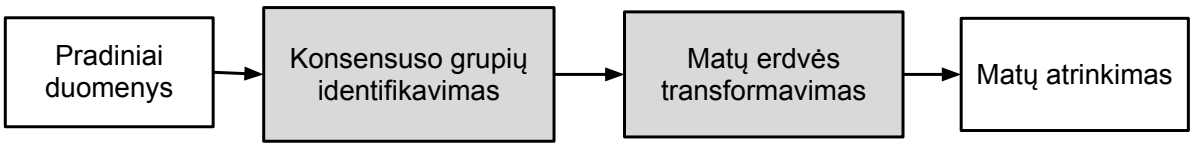


10 pav.: Multikriterinio rekursyvaus matų eliminavimo algoritmas.

3.5. Multikriterinis rekursyvus matų eliminavimas

Jei matų atrinkimo tikslas yra pagerinti klasifikavimo rezultatus, tai taikant multikriterinius matų atrinkimo metodus nebūtinai surandamas geriausias matų poaibis. Tam, kad būtų surastas geriausias matų poaibis reikia kombinuoti multikriterinį matų reitingavimą su matų paieškos strategija. Rekursyvus matų eliminavimas yra dažnai naudojama paieškos strategija matų atrinkimui.

Multikriterinis rekursyvus matų eliminavimas (angl. *Multicriterion Fusion Based Recursive Feature Elimination*, MCF-RFE) susideda iš dviejų dalių [YM11]: keleto matų atrinkimo kriterijų suliejimo pagal svorius ir pagal reitingus, ir rekursyvaus matų eliminavimo aprašyto algoritme nr. 2. Algoritmas pavaizduotas 10 pav.



11 pav.: Konsensuso grupėmis grįstas stabilių matų atrinkimas.

Standartinis rekursyvus matų eliminavimas, kai vienos iteracijos metu yra eliminuojamas vienas matas, gali labai padidinti algoritmo sudėtingumą. Todėl genų ekspresijos duomenims prasmingiau yra eliminuoti keletą matų vienu metu.

Nors SVM-RFE 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.

3.6. Konsensuso grupėmis grįstas stabilių matų atrinkimo metodas

Konsensuso grupėmis grįstas stabilių matų atrinkimo metodas (angl. *Consensus Group Stable feature selection*, CGS), pirma, identifikuoja tankias matų grupes, antra, pagal surastas grupes transformuoja matų erdvę, trečia, transformuotoje matų erdvėje atlieka matų atrinkimą [LYD09]. Schematiškai šis algoritmas pavaizduotas 11 pav.

CGS metodo pagrindinė dalis yra tankių matų identifikavimas. Šio uždavinio sprendimui naudojamas *Dense Group Finder* (DGF) algoritmas. DGF aprašytas algoritme nr. 3. CGS algoritme 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 erdvę. Transformuotoje matų erdvėje vykdomas matų atrinkimas kuriuo nors matų atrinkimo metodu Φ, pvz., *Relief* matų atrinkimo metodu.

Algoritmas nr. 3 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 (20)
  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 $||p_r - x_i|| < h$

$$y_{i,j+1} = \frac{\sum_{i=1}^{n} x_i K(\frac{y_j - x_i}{h})}{\sum_{i=1}^{n} K(\frac{y_j - x_i}{h})} j = 1, 2, ... \tag{20}$$

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

Algoritmas nr. 4 Konsensuso grupėmis grįstas stabilių matų atrinkimas

Įeitis: mėginių aibė D, iteracijų skaičius t, matų atrinkimo metodas Φ
Išeitis: atrinktos konsensuso matų grupės $CG_1, CG_2, ..., CG_k$
// 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 $CG_1, CG_2, ..., CG_L$ 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 $CG_1, CG_2, ..., CG_L$ pagal $\Phi(X_i)$
Pasirinkti k matų, turinčių geriausią reitingą

4. EKSPERIMENTAI

Šiame skyriuje yra aprašyti naudoti biomedicininių duomenų rinkiniai, atlikti eksperimentai, įvardinti eksperimentų nustatymai, pateikti matų atrinkimo metodų spartos matavimai, pristatyti klasifikavimo tikslumo matavimai ir apžvelgti stabilių matų atrinkimo rezultatai. Eksperimentais siekiama išsiaiškinti pavienių ir kombinuotų matų atrinkimo metodų įtaką klasifikavimui bei matų atrinkimo stabilumui.

4.1. Eksperimentuose naudoti duomenys

Šiame darbe eksperimentai buvo atliekami su biomedicininiais viešai prieinamais genų ekspresijos mėginių rinkiniais. Informacija apie mėginių rinkinius pateikta 1 lentelėje.

Mėginių rinkinius apibūdinantis dydis OMS (Objektų-Matų Santykis), kuris turimiems mėginių rinkiniams yra nuo 0,403% iki 3,01% procento, reiškia, kad turimi mėginiai turi šimtus kartų daugiau matų nei mėginių. Tai apsunkina duomenų tyrimo procesą ir gali sukelti persimokymo (angl. *overfitting*) problemą.

Šizofrenijos ir maniakinės depresijos mėginių rinkinys ypatingas tuo, kad jis turi tris klases. Šiame darbe nagrinėjamas tik dviejų klasių atvejis, todėl šizofrenija sergančių pacientų mėginiai nebuvo naudojami.

1 lentelė. Darbe naudoti mėginių rinkiniai

Pavadinimas	**Šaltinis**	**Mėginių skaičius (+/-)**	**Matų skaičius**	**OMS**
Gaubtinės žarnos auglys (angl. Colon)	[ABN+99]	62 (40/22)	2000	0.031
Centrinės nervų sistemos auglys (CNS)	[PTG+02]	60 (39 / 21)	7129	0.0084
Prostatos auglys	[SFR+02]	102 (52/50)	6033	0.0169
Šizofrenija ir maniakinė depresija	[Ins12]	90 (bp[4]: sz[5]: cc[6] =30:31:29)	22283	0.00404

[4]bp (angl. *Bipolar Disorder*) - maniakine depresija sergantys pacientai.
[5]sz (angl. *Schizophrenia*) - šizofrenija sergantys pacientai.
[6]cc (angl. *Control Crowd*) - kontrolinė grupė.

4.2. Metodologija

Eksperimentuose buvo naudojami skyrelyje nr. 4.1. aprašyti mėginių rinkiniai. Mėginių rinkiniai nebuvo atskirai normalizuojami, nes daryta prielaida, jog duomenys jau yra apdoroti.

Klasifikavimui naudota atraminių vektorių klasifikatorių algoritmo R programavimo kalbos paketo „e1071“ implementacija. Naudotas tiesinis atraminių vektorių klasifikavimo algoritmas su parametro C reikšme 0,01, kuri buvo nustatyta empiriškai.

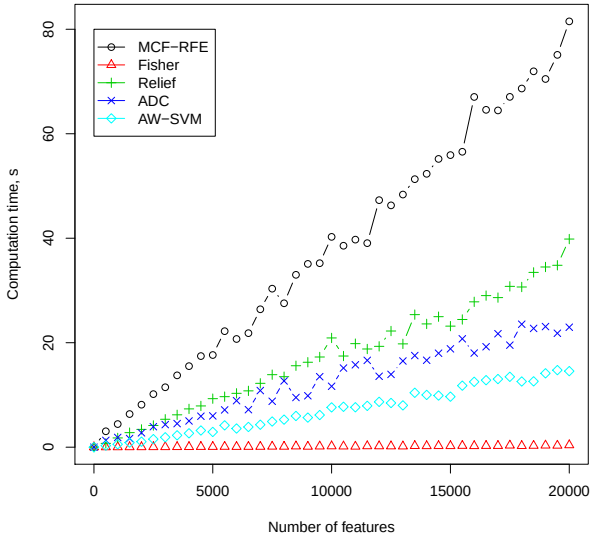
Matų atrinkimo metodai buvo suprogramuoti šio darbo autoriaus, nes nėra standartinių R kalbos paketų, kuriuose matų atrinkimo metodai jau būtų implementuoti.

Dėl to, kad turima mažai mėginių ir daug matų, klasifikavimas buvo kartojamas 300 kartų, kai treniravimosi duomenų aibę sudarė kaskart atsitiktinai parenkami 90% mėginių. Kiekvienoje iteracijoje su vis kitu mėginių poaibiu buvo atliekamas matų atrinkimas, po to būdavo atliekamas klasifikavimas su 10, 20, .., 500 aukščiausią reitingą turinčių matų.

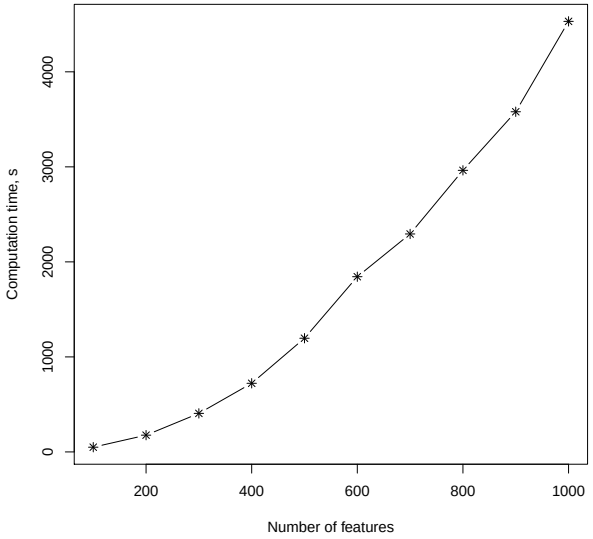
4.3. Matų atrinkimo metodų sparta

Matų atrinkimo metodų darbo laikas buvo palygintas naudojant vieną šizofrenijos ir maniakinės depresijos duomenų rinkinį [Ins12]. Skaičiavimai buvo atlikti kompiuteryje naudojant vieną 2.66GHz spartos procesoriaus branduolį, bei 2 GB RAM atminties. 12 pav. ir 13 pav. pavaizduota matų atrinkimo metodo darbo laiko priklausomybė nuo mėginius apibūdinančių matų skaičiaus.

Matų atrinkimo metodų darbo laikas yra atvaizduotas dviem grafikais, nes pagal atliktų eksperimentų rezultatus buvo pastebėta, kad CGS matų atrinkimo metodas yra apie 1000 kartų lėtesnis už kitus suprogramuotus matų atrinkimo metodus, todėl viename grafike neįmanoma atvaizduoti visų turimų matų atrinkimo metodų. Pagal 12 pav. galime daryti išvadą, kad sparčiausias matų atrinkimo metodas yra *Fisher* įvertis. Remiantis matų darbo laiko priklausomybės nuo matų kiekio grafikais galime daryti išvadą, kad CGS algoritmas daugiamačių duomenų matų atrinkimui nėra tinkamas, nes jo skaičiavimų laikas yra per ilgas.



12 pav.: Pagrindinių matų atrinkimo metodų darbo laikas.



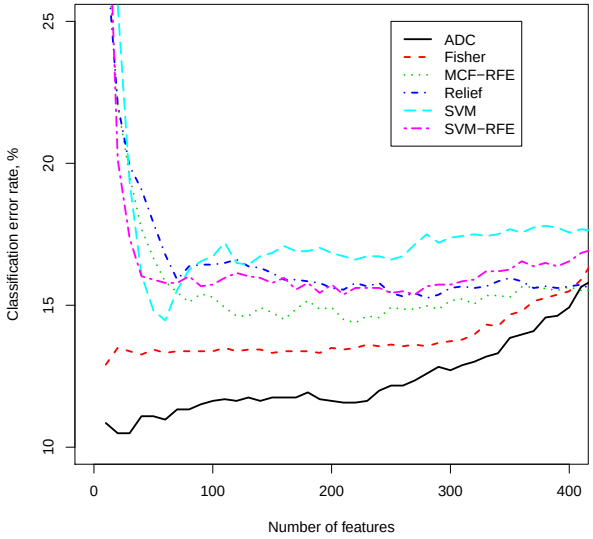
13 pav.: Konsensuso grupėmis grįsto matų atrinkimo metodo darbo laikas.

4.4. Klasifikavimo pagal atrinktus matus tikslumas

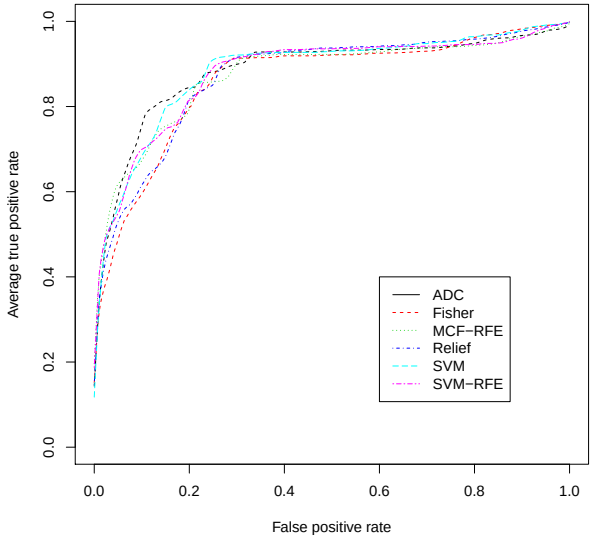
Matų atrinkimo metodų įtaką klasifikavimo klasifikavimo tikslumui buvo matuojama naudojant tris biomediciniinių duomenų rinkinius: gaubtinės žarnos auglio (angl. *colon*), centrinės nervų sistemos (CNS), prostatos. Klasifikavimui buvo naudojami tiesiniai atraminių vektorių klasifikatoriai (SVM), su empiriškai nustatytu parametru $C = 0.01$. 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: klasifikavimo nuostolio priklausomybės nuo atrinktų matų skaičiaus, bei ROC kreivėmis, kurios buvo apskaičiuotos pagal duomenis gautus klasifikuojant su tiek atrinktų matų, su kiek klasifikavimo tikslumas buvo pats geriausias [GS66].

14 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.

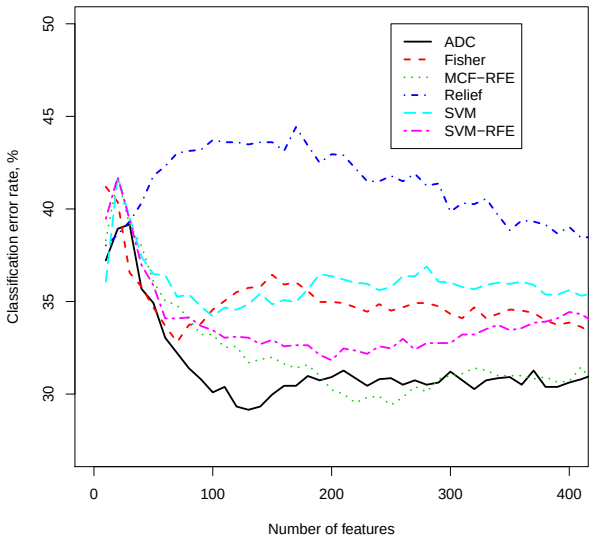
Centrinės nervų sistemos duomenų rinkinys yra sunkiai klasifikuojamas, nes vidutinis klaidų skaičius yra apie 35%, kai, pvz. gautinės žarnos auglio duomenų rinkinio vidutinis klaidų skaičius yra tik 15%. 16 pav. matome, kad šiam duomenų rinkiniui vidutiniškai



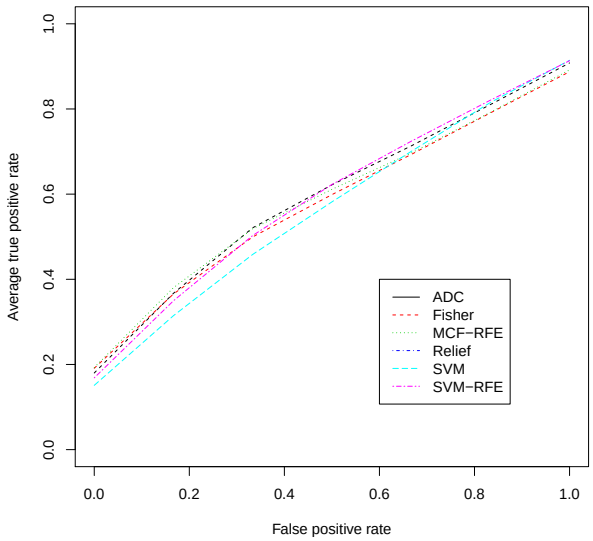
14 pav.: Gaubtinės žarnos auglio meginių klasifikatorių tikslumas.



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



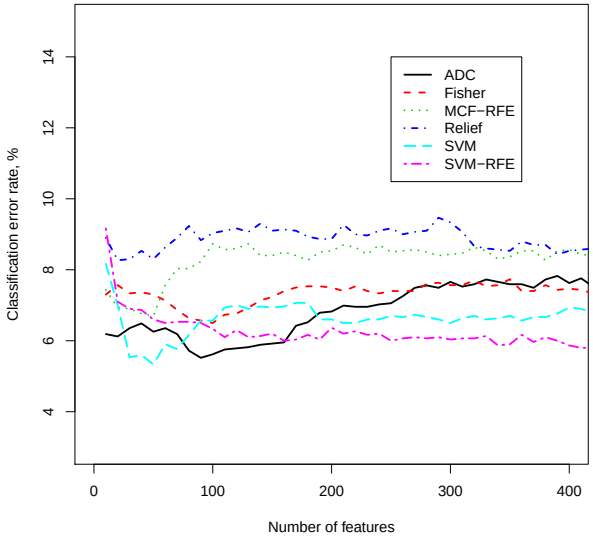
16 pav.: Centrinės nervų sistemos meginių klasifikatorių tikslumas.



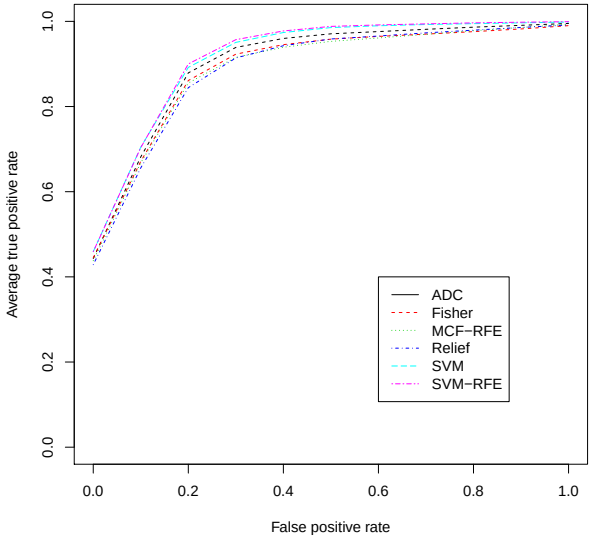
17 pav.: Centrinės nervų sistemos mėginių klasifikatorių ROC kreivės.

geriausiai matus atrenka ADC ir multikriterinio rekursyvaus matų eliminavimo metodai. Prasčiausiai pasirodo *Relief* metodas. 18 pav. parodyta, kad prostatos duomenų rinkinio matus klasifikavimui geriausiai atrenka ADC bei absoliučių svorių SVM metodas. Prasčiausiai matus atrenka *Relief*.

Klasifikavimo nuostolio grafikuose matoma, kad pagal kombinuotojo mokymosi algoritmą, multikriterinį rekursyvų matų eliminavimą (MCF-RFE), nėra sukuriamas geriausias klasifikatorius. Taip yra, nes MCF-RFE sudarytas iš atskirų matų atrinkimo metodų, kurių



18 pav.: Prostatos meginių klasifikatorių tikslumas.



19 pav.: Prostatos mėginių klasifikatorių ROC kreivės.

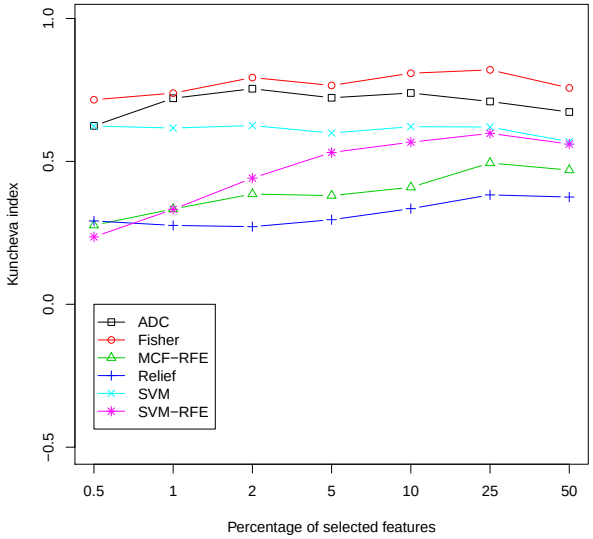
rezultatai varijuoja. Tačiau pagal MCF-RFE matų atrinkimo metodas pasirodo sąlyginai gerai, kai yra taikomas sunkiai klasifikuojamiems duomenims, pvz., CNS. Kadangi biomediciniai duomenys dažniausiai yra sunkiai klasifikuojami, todėl MCF-RFE metodo taikymas tokiems duomenims yra prasmingas.

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. Rezultatai parodė, kad matų atrinkimas svariai prisideda prie geresnio klasifikatoriaus sukūrimo.

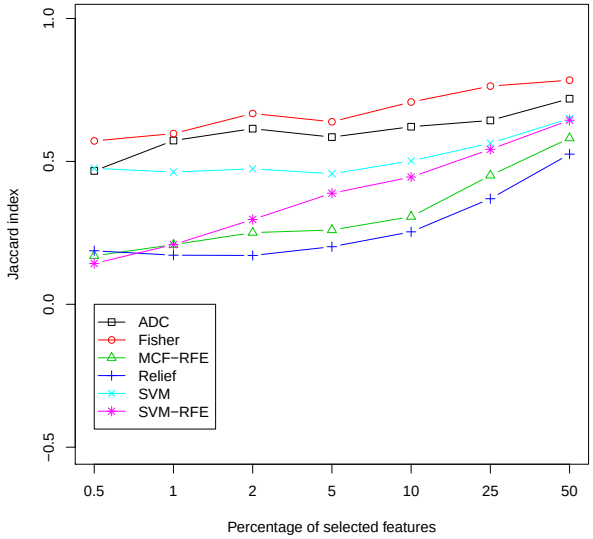
4.5. Matų atrinkimo stabilumas

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 poskyrio grafikus reikia nagrinėti atsižvelgiant į poskyrio, kuriame buvo nagrinėtas klasifikavimo pagal atrinktus matus tikslumas.

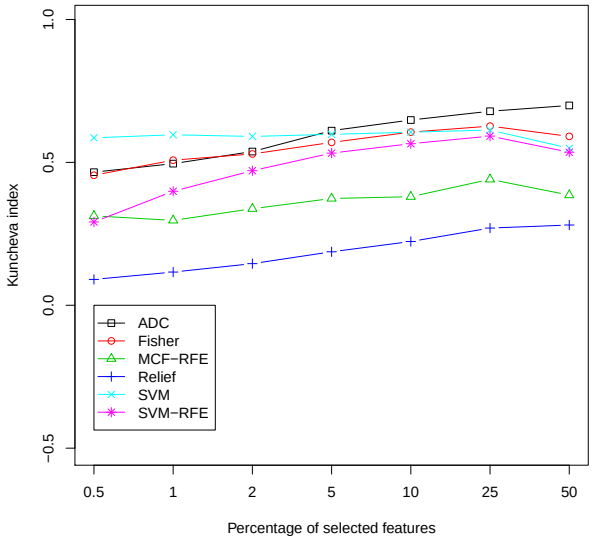
Pagal 20 pav. ir 21 pav. galima teigti, kad gaubtinės žarnos auglio duomenų rinkinio matus stabiliausiai atrenka *Fisher* įvertis. Mažiausiai stabiliai matus atrenka *Relief*



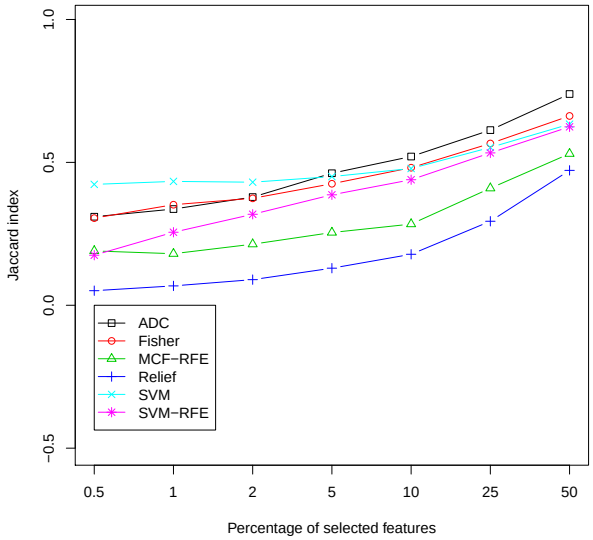
20 pav.: Matų atrinkimo gaubtinės žarnos auglio mėginiams stabilumo grafikas pagal *Kuncheva* indeksą.



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



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

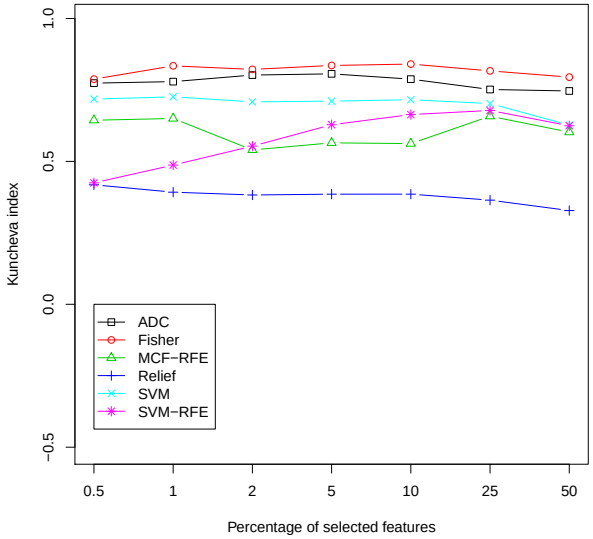


23 pav.: Matų atrinkimo CNS mėginiams stabilumo grafikas pagal *Jaccard* indeksą.

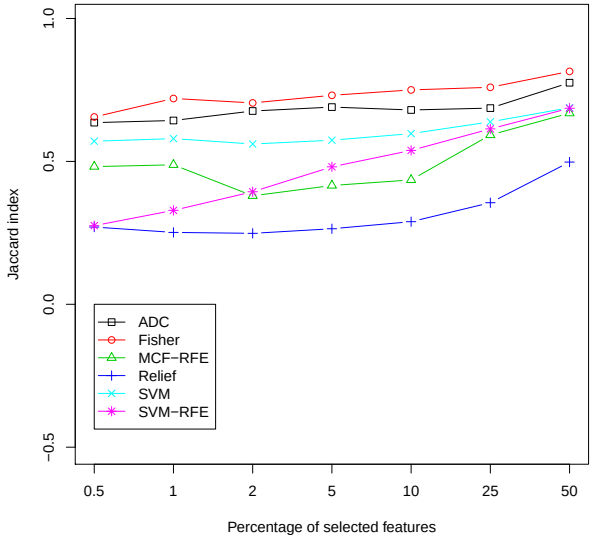
metodas.

Pagal 24 pav. ir 25 pav. galima teigti, kad prostatos duomenų rinkinio matus stabiliausiai atrenka *Fisher* įvertis. Mažiausiai stabiliai matus atrenka *Relief* metodas.

Multikriterinio rekursyvus matų eliminavimo metodas neatrenka stabilių matų, nors teoriškai šis metodas yra skirtas matų atrinkimo stabilumui padidinti. Taip yra, nes kombinuojami ne tik stabilesni, pvz. *Fisher* ar ADC, metodai bet ir nestabilumu pasižymintys



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



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

metodai, pvz., *relief*. Todėl galutiniam rezultatui įtakos turi labai varijuojantys veiksniai, dėl kurių nukenčia stabilumas. MCF-RFE algoritmo stabilumą padidintų dar keletos matų atrinkimo metodų pridėjimas ar svorių suteikimas pavieniams matų atrinkimo metodams.

Apibendrinant matų atrinkimo stabilumo matavimus galima daryti išvadą, kad matų atrinkimo stabilumas priklauso ne tik nuo matų atrinkimo metodo, bet ir nuo duomenų rinkinio, kurio matai yra atrenkami. 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 nestabiliausias iš tirtųjų. Gana geru stabilumu pasižymi ADC metodas bei *Fisher* įvertis.

REZULTATAI IR TOLIMESNIŲ TYRIMŲ KRYPTYS

Šiame darbe analizuota vis didesnį susidomėjimą kelianti daugiamačių duomenų klasifikavimo problematika ypatingą dėmesį kreipiant į matų atrinkimo problematiką. Atliekant analizę buvo susipažinta su daugiamačių duomenų klasifikavimo ir matų atrinkimo geriausiomis praktikomis ir atlikti matų atrinkimo metodų palyginimo eksperimentus.

Matų atrinkimo metodų palyginamųjų eksperimentų metu gauti rezultatai parodė, kad nėra vieno universaliai geriausio matų atrinkimo metodo. Pasirenkant matų atrinkimo metodą visada reikia atsižvelgti į turimus duomenis bei į darbui keliamus tikslus.

Matų atrinkimas daugiamačiams duomenims yra labai svarbus. Ieškant naujų kokybiškų matų atrinkimo metodų dvi pagrindinės kryptys yra naudoti multikriterinius matų atrinkimo ir panašių matų grupavimo metodus. Panašių matų grupavimui kol kas yra pasiūlytas tik CGS metodas, kuris netinka daugiamačiams duomenims, nes yra per lėtas. Multikriteriniai matų reitingavimo metodai kenčia situacijose, kai kurie nors iš kriterijų tam tikram duomenų rinkiniui demonstruoja prastus rezultatus, pavyzdžiui, duomenyse yra daug koreliuojančių matų, o matų reitingavimo metodai į tai neatsižvelgia. Kuo daugiau matų turi duomenys, tuo labiau pasireiškia minėtos problemos.

Šiame darbe sukaupta patirtis gali būti panaudota kaip tolimesnių daugiamačių biomedicininių duomenų tyrimų pagrindas. Toliau ieškant stabilių matų atrinkimo metodų, kurie maksimaliai padidina klasifikavimo tikslumą, reikėtų daugiau dėmesio kreipti į tinkmo matų grupavimo algoritmo paiešką. Tinkamas matų grupavimo algoritmas turėtų atsižvelgti į matų tarpusavio koreliacijas bei būti pakankamai spartus.

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.

[ABR64] A. Aizerman, E.M. Braverman, and LI Rozoner. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and remote control*, 25:821–837, 1964.

[Alb12] University Albany. Notes on machine learning, 2012.

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

[BFOS84] L. Breiman, J.H. Friedman, R.A. Olshen, and C.J. Stone. Classification and regression trees. 1984.

[BND04] U.M. Braga-Neto and E.R. Dougherty. Is cross-validation valid for small-sample microarray classification? *Bioinformatics*, 20(3):374–380, 2004.

[Bre96] L. Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[CV95] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

[DHS01] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Pattern Classification and Scene Analysis: Pattern Classification. Wiley, 2001. `http://books.google.lt/books?id=YoxQAAAAMAAJ`.

[Die00] T. Dietterich. Ensemble methods in machine learning. *Multiple classifier systems*, pages 1–15, 2000.

[DK82] P.A. Devijver and J. Kittler. *Pattern recognition: A statistical approach*. Prentice/Hall International, 1982.

[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.

[Fle09] T. Fletcher. Support vector machines explained. *Tutorial paper., Mar*, 2009.

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

[GH07] A. Gelman and J. Hill. *Data Analysis Using Regression And Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge University Press, 2007.

[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.

[Ham50] R.W. Hamming. Error detecting and error correcting codes. *Bell System technical journal*, 29(2):147–160, 1950.

[HBV02] Maria Halkidi, Yannis Batistakis, and Michalis Vazirgiannis. Cluster validity methods: Part i. *SIGMOD Record*, 31(2):40–45, 2002.

[HK00] Jiawei Han and Micheline Kamber. *Data Mining: Concepts and Techniques*. Morgan Kaufmann, 2000.

[Ins12] Stanley Medical Research Institute. Online genomics database, 2012. [žiūrėta 2012-04-03]. Prieiga per internetą: <www.stanleygenomics.org>.

[Jac01] P. Jaccard. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura*. 1901.

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

[KPH07] A. Kalousis, J. Prados, and M. Hilario. Stability of feature selection algorithms: a study on high-dimensional spaces. *Knowledge and information systems*, 12(1):95–116, 2007.

[Kun07] Ludmila I. Kuncheva. A stability index for feature selection. In *Artificial Intelligence and Applications*, pages 421–427, 2007.

[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.

[Mar08] Dalia Martišiūtė. Vaizdų klasterizavimas. Master's thesis, Vilniaus universitetas, 2008. Prieiga internetu: `http://vddb.laba.lt/fedora/get/LT-eLABa-0001:E.02~2008~D_20090908_201754-37094/DS.005.1.01.ETD`.

[MST94] D. Michie, D.J. Spiegelhalter, and C.C. Taylor. Machine learning, neural and statistical classification. 1994.

[PLA09] E. Pettersson, J. Lundeberg, and A. Ahmadian. Generations of sequencing technologies. *Genomics*, 93(2):105–111, 2009.

[Pol06] R. Polikar. Ensemble based systems in decision making. *Circuits and Systems Magazine, IEEE*, 6(3):21–45, 2006.

[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.

[PWCG01] Paul Pavlidis, Jason Weston, Jinsong Cai, and William Noble Grundy. Gene functional classification from heterogeneous data. In *Proceedings of the fifth annual international conference on Computational biology*, RECOMB '01, pages 249–255, New York, NY, USA, 2001. ACM.

[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.

[Sch03] R.E. Schapire. The boosting approach to machine learning: An overview. *LECTURE NOTES IN STATISTICS-NEW YORK-SPRINGER VERLAG-*, pages 149–172, 2003.

[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.

[Wol92] D.H. Wolpert. Stacked generalization*. *Neural networks*, 5(2):241–259, 1992.

[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.

[YM11] 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.

SĄVOKŲ APIBRĖŽIMAI

Klasifikavimas – procesas, kurio metu sukonstruojama funkcija, kuri pagal mėginių nepriklausomus kintamuosius apskaičiuoja priklausomus kintamuosius.

Klasifikatorius – funkcija, kuri pagal mėginių nepriklausomus kintamuosius apskaičiuoja priklausomus kintamuosius.

Jautrumas (angl. *sensitivity*) – įvertis, kuris parodo testo/modelio gebėjimą dignozuoti susirgimą, jeigu asmuo iš tikrųjų serga (ligonis identifikuojamas ligoniu).

Specifiškumas (angl. *specificity*) – įvertis, parodantis testo/modelio gebėjimą nustatyti, jog susirgimo nėra, kai jo iš tikrųjų nėra (sveikas identifikuojamas sveiku).

Persimokymas (angl. *overfitting*) – reiškinys, kai klasifikavimo algoritmas per daug prisitaiko prie treniravimosi duomenų. Sukurtas klasifikatorius pasižymi aukštu klasifikavimo tikslumu dirbant su treniravimosi duomenimis, tačiau klasifikavimo tikslumas yra žemas dirbant su testiniais duomenimis.

Genėjimas (klasifikavimo medžių) (angl. *pruning*) - mazgų, kurie turi sąlyginai mažą atskiriamąją galią, pašalinimas iš klasifikavimo medžio.

Hiperplokštuma (angl. *hyperplane*) – plokštumos generalizacija daugiamatėje erdvėje.

Triukšmas (angl. *noise*) – pašaliniai atsitiktiniai signalai, patekę į informaciją nešančių signalų srautą.

Išimtis (angl. *outlier*) – objektas, kuris savo skaitine reikšme daug didesnis arba daug mažesnis už imties vidurkį.

Mašininis mokymasis (angl. *machine learning*) – dirbtinio intelekto šaka, kurios tyrėjai siekia įgalinti kompiuterius tobulinti savo elgseną (mokytis) empirinių duomenų atžvilgiu.

Regresija (lot. *regressio* – grįžimas, traukimasis) – tikimybių teorijoje ir mat. statistikoje – atsitiktinio dydžio vidurkio priklausomybės nuo kt. dydžio (kelių dydžių) išraiška.

Mokymosi duomenys (angl. *training data*) – duomenys, pagal kuriuos yra kuriamas klasifikatorius.

Testavimo duomenys (angl. *testing data*) – duomenys, kuriais bus validuojamas sukurtas klasifikatorius.