

Turinys

IVADAS	3
1. MAŠININIO MOKYMOSI APŽVALGA	7
1.1. Mokymasis su mokytoju	7
1.1.1. Klasifikavimas	7
1.1.2. Regresinė analizė	9
1.2. Mokymasis be mokytojo	9
1.2.1. Klasterizavimas	9
1.3. Mokymosi su mokytoju ir mokymosi be mokytojo palyginimas	11
1.4. Kombinuotasis mokymasis	11
1.5. Atraminių vektorių klasifikatoriai	11
1.6. <i>Random Forest</i> klasifikatorius	13
2. PAGRINDINIAI MATŲ ATRINKIMO METODAI	14
2.1. <i>Fisher</i> įvertis	15
2.2. <i>Relief</i> metodas	15
2.3. Asimetrinis priklausomybės koeficientas	16
2.4. Absoliučių svorių SVM	17
2.5. Rekursyvus matų eliminavimas pagal SVM	17
3. STABIILIŲ MATŲ ATRINKIMO METODAI	18
3.1. Matų atrinkimo stabilumas	19
3.1.1. Stabilumo matavimas	19
3.1.2. <i>Kuncheva</i> indeksas	20
3.1.3. <i>Jaccard</i> indeksas	21
3.1.4. <i>Hamming</i> atstumas	21
3.2. Svoriais grįstas multikriterinis suliejimas	21
3.3. Reitingais grįstas multikriterinis suliejimas	23
3.4. Svoriais ir reitingais grįstas multikriterinis suliejimas	23
3.5. Multikriterinis rekursyvus matų eliminavimas	24
3.6. Konsensuso grupėmis grįstas stabilių matų atrinkimo metodas	25
4. EKSPERIMENTAI	27
4.1. Eksperimentuose naudoti duomenys	27

4.2. Metodologija	28
4.3. Matų atrinkimo metodų sparta	29
4.4. Klasifikavimo pagal atrinktus matus tikslumas	29
4.5. Matų atrinkimo stabilumas	32
REZULTATAI IR IŠVADOS	35
ŠALTINIAI	35
SĄVOKŲ APIBRĖŽIMAI	39

ĮVADAS

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

Šiame darbe yra nagrinėjama biomedicinoje kaupiamų genetinių daugiamačių duomenų analizės specifika. Šie duomenys yra specifiski tuo, kad jie turi šimtus kartų daugiau matų nei mėginių. Kadangi mėginio gavimo kaina yra aukšta, turimas mažas mėginių skaičius. Biomedicininis duomenų analizę apsunkina ir tai, kad matavimai, kuriais tie duomenys gaunami, yra triukšmingi. Triukšmas matavimo metu atsiranda dėl cheminių reakcijų netikslumo, tiriamo organizmo sudėtingumo. Kai duomenys yra triukšmingi, tai didėjant juos apibūdinančių matų skaičiui, didėja tikimybė duomenyse, kad bus rasta atsitiktinių priklausomybių. Tai yra pagrindinė priežastis, kodėl biomedicininis duomenų analizės procesas yra sudėtingas.

Biomedicininis duomenų klasifikavimo užduotis yra atskirti sveikųjų pacientų mėginius nuo sergančiųjų. Klasifikavimu siekiama nustatyti, kurie matai, veikdami drauge, geriausiai paaikšina skirtumą tarp ligos paveiktų ir nepaveiktų mėginių. Labiausiai ligą paaikšinančių 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 naudojant turimus mėginius – treniravimosi duomenis – ir informaciją apie jų būklę (sveikas ar sergantis). Klasifikatoriaus ruošimo procesas yra vadinamas apmokymu. Klasifikatoriai yra validuojami su testiniais duomenimis, o naudojami nustatant naujų, dar nematytų, mėginių būklę.

Dėl „daugiamatiškumo prakeikimo“ (angl. *the curse of dimensionality*) didėjant matų kiekiui mėginiai pasidaro panašūs, todėl bandymas juos klasifikuoti tolygus spėliojimui [Bel66]. Biomedicininis duomenų kontekste galima daryti prielaidą, kad ne visi matai yra susiję su tiriamą problema, pvz. gaubtinės žarnos vėžiu, dėl to, kad duomenys yra daugiamačiai. Paprastai nagrinėjamai problemai svarbus yra mažas, palyginus su visu, matų kiekis.

Todėl biomedicininį 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.

Matų atrinkimas yra svarbi biomedicininį duomenų apdorojimo (angl. *preprocessing*) etapo dalis. Naudojant matų atrinkimo metodus, galima kovoti su daugiamatiškumo prakeiksmu matų skaičių priartinant prie mėginių skaičiaus. Todėl svarbu yra pasirinkti geriausiai tinkančią matų atrinkimo strategiją. Kadangi 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.

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

Naudojant kryžminio patikrinimo metodą, daug kartų sudaromos skirtingos treniravimosi ir testinės mėginių imtys. Taikant atskirą šio metodo variantą, kryžminį patikrinimą paliekant vieną mėginį (angl. *leave-one-out cross-validation*), iš mokymosi duomenų išimamas vienas (testinis) mėginys, o su likusiais apmokomas klasifikatorius, kuris klasifikuoja išbrauktąjį mėginį. Procesas tęsiamas tol, kol suklasifikuojami visi objektai. Kitais kryžminio patikrinimo metodo variantais iš treniravimosi mėginių yra išimama po keletą mėginių. Pagal tai, kiek testinių mėginių klasifikatorius priskyrė klaidingai kategorijai, yra nustatoma vidutinė klaidingo klasifikavimo tikimybė. Šiuo metodu gauti įverčiai pasižymi dideliu

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

klasifikavimo rezultatų variabilumu [BND04].

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. *0.623² bootstrap*) metodas. Šiuo metodu gautas vidutinio klasifikavimo nuostolio įvertis pasižymi mažu variabilumu [MST94].

Kadangi biomedicininuose duomenyse reikšmingų matų kiekis tiriamai problemai yra nedidelis, todėl tyrėjams norint geriau suprasti nagrinėjamus biomedicininis duomenis yra svarbu orientuotis į mažesnę 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 matais [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 biomedicininų duomenų tyrimus – tyrėjams reikia atlikinėti bandymus su mažesniu mėginių skaičiumi, taip pat kuriant medicininius diagnostikos įrankius, naudojamų matų kiekis įtakoja įrankio kainą. Todėl stabilių matų atrinkimas dirbant su biomedicininiais duomenimis yra svarbus.

Matų atrinkimo stabilumo problemą Yang ir Mao [YM11] siūlė spręsti reitinguojant matus remiantis keletos matų atrinkimo metodų rezultatais. Galutinis matų reitingo sąrašas gaunamas, kai po kiekvieno matų atrinkimo yra išmetama viena 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 daugiamatų duomenų analizėje.

Matų atrinkimo stabilumo problemą siūlyta spręsti surandant matų grupių tankio centrus ir naudoti matus, kurie artimiausi tiems centrams [YDL08]. Pasiūlytas grupių tankių

²0.623 yra tikimybė mėginiui būti įtrauktam į treniravimosi imtį.

algoritmas užtrunka $O(\lambda n^2 m)$ laiko, kur n yra matų kiekis, o m - mėginių skaičius. Vėliau Loscalzo ir kt. pasiūlė mokymo duomenis skaidyti poaibiais ir kiekviename poaibyje ieškoti tankių grupių, o tada imti sprendimą balsavimo principu [LYD09]. Nors šie metodai siūlo stabilų matų atrinkimą, tačiau šių metodų panaudojamumą daugiamatčiuose duomenyse riboja skaičiavimo sudėtingumas.

Šiame bakalauriniame darbe remiantis Yang, Mao bei Loscalzo darbuose pateiktomis išvalgomis, bus stengiamasi pasiūlyti tyrimų kryptis, kurios galėtų padėti sukurti metodus, skirtus spręsti stabilų matų atrinkimo problemą. Idėja yra sugrupuoti matus pagal greitą klasterizacijos algoritmą, išrinkti reprezentatyviausius matus, transformuoti matų erdvę ir joje vykdyti matų atrinkimą remiantis keletu matų atrinkimo metodų.

Šio darbo tikslas yra išanalizuoti daugiamatčių duomenų klasifikavimo ypatybes. Šiam darbui yra keliamos tokios užduotys:

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

Tolimesnė darbo struktūra yra tokia: skyriuje nr.1 apžvelgiami su atliekamu tyrimu susiję darbai bei reikalinga 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 daugiamatčių duomenų analizės metu gauti duomenys.

1. MAŠININIO MOKYMOSI APŽVALGA

Mašininis mokymasis (angl. *machine learning*) yra dirbtinio intelekto šaka, kurios tyrėjai siekia įgalinti kompiuterius tobulinti savo elgseną (mokyti) empirinių duomenų atžvilgiu [DHS]. 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*). Toliau šiame skyriuje apžvelgiami mašininio mokymosi pagrindai: mokymasis su mokytoju, mokymasis be mokytojo, atraminių vektorių klasifikatoriai, *Random Forest* klasifikatorius.

1.1. Mokymasis su mokytoju

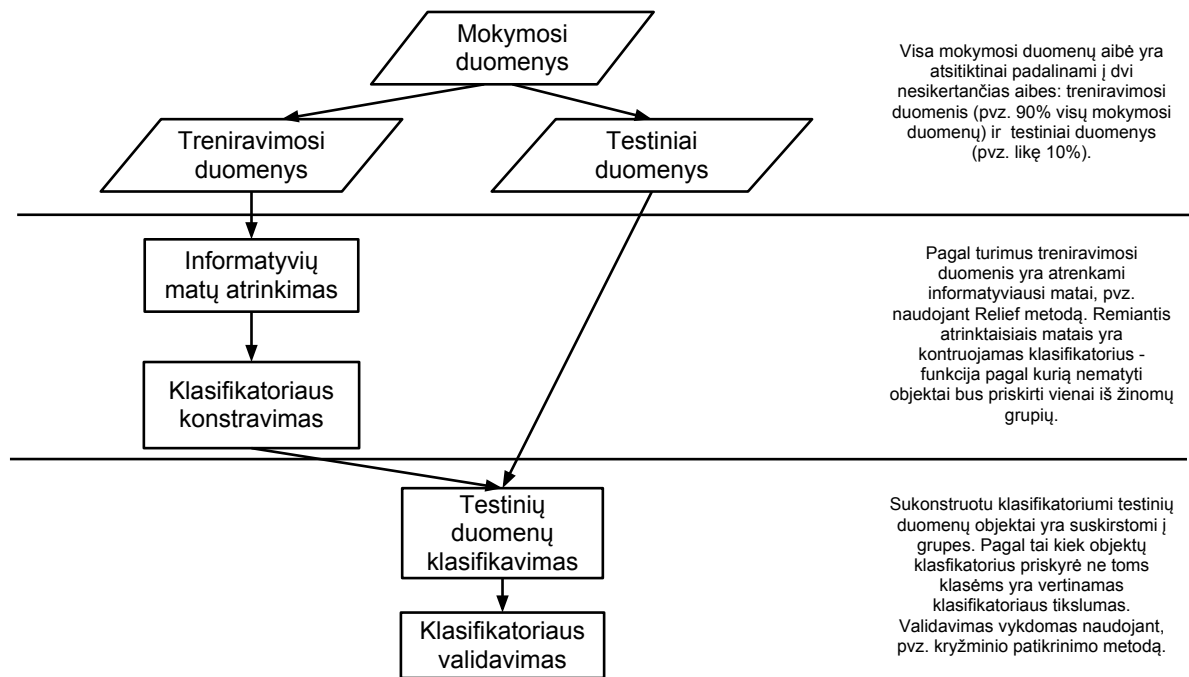
Žmonės mokosi iš patirties, tačiau, skirtingai nei žmonės, kompiuteriai patirties neturi, todėl kompiuteriai turi mokytis iš patyrimą apibūdinančių duomenų – mokymosi duomenų (angl. *training data*). Mokymosi su mokytoju tikslas yra sukonstruoti funkciją, kuri galėtų būti naudojama nuspėti testavimo duomenų (angl. *testing data*) charakteristikų reikšmes pagal mokymosi duomenis. Šiame kontekste mokytoją reikia suprasti kaip išankstinį mokymosi duomenų spėjamų charakteristikų žinojimą. Kitaip tariant, mokymosi su mokytoju metodais yra sprendžiami uždaviniai, kuriems atsakymus galima patikrinti. 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

Mašininio mokymosi kontekste klasifikavimu yra vadinama problema, kai pagal mokymosi duomenis reikia nustatyti, kuriai klasei priklauso objektas. Klasifikavimo procesas trimis etapais pavaizduoti 1 pav. srautų diagramoje. Klasifikavimo etapai:

1. Visa mokymosi duomenų aibė yra atsitiktinai padalinama į dvi nesikertančias aibes:



1 pav.: Klasifikavimo srautų diagrama su paaiškinimais.

treniravimosi duomenys (pvz. 90% visų mokymosi duomenų) ir testiniai 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 testiniai 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 grupei – 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 – *****TODO***** absci-

sių ašyje tikimybės (angl. *false positive*) įvykio, ordinačių ašyje (angl. *true positive*) įvykio.

1.1.2. Regresinė analizė

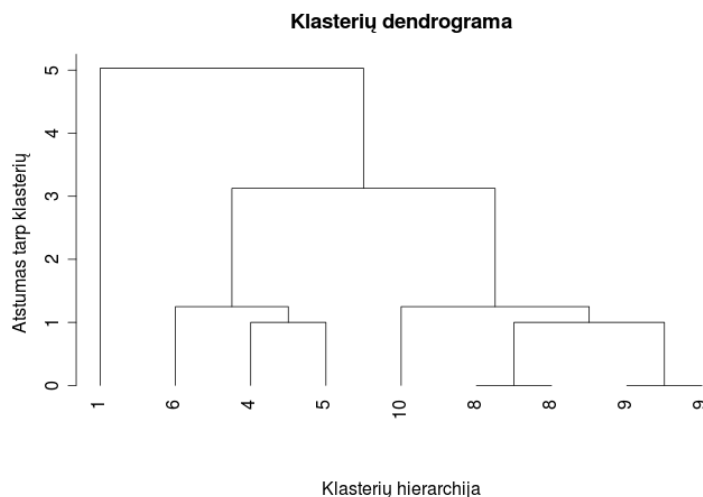
Mašininio mokymosi kontekste regresine analize yra vadinama problema, kai pagal patirtį apibūdinančius duomenis reikia nustatyti kiekybines duomenų charakteristikas. Regresinė analizė naudoja standartinius statistinius metodus, tokius kaip mažiausių kvadratų metodas (angl. *least squares*). Regresinė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

Dirbant su biomedicininiais duomenimis regresinė analizė gali būti taikoma bandant nuspėti vėžio stadiją mėginiui. Tačiau regresinė analizė dirbant su biomedicininiais duomenimis yra naudojama rečiau negu klasifikavimas, todėl toliau šiame darbe bus nagrinėjama klasifikavimo problema.

1.2. Mokymasis be mokytojo

Mašininio mokymosi kontekste dažnai sutinkamas uždavinys yra į prasmingas grupes sugrupuoti turimus duomenis, kurių grupavimas iš anksto nėra žinomas. Tokie uždaviniai yra sprendžiami mokymosi be mokytojo metodais. Mokymosi be mokytojo metodų pagrindinis principas – maksimizuoti mėginių, esančių toje pačioje grupėje, tarpusavio panašumą ir minimizuoti mėginių panašumą esančių skirtingose grupėse.

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



2 pav.: Hierarchinio klasterizavimo rezultatų grafinis pavyzdys.

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 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 objekto ir kiekvienoje iteracijoje sujungia panašiausius klasterius (angl. *agglomerative clustering*) [HK00]. Hierarchinio klasterizavimo rezultatas – klasterių medis, dendrograma, rodanti, kaip klasteriai yra hierar-

chiškai susiję. Pasirinktame lygyje nupjovus dendrogramą gaunama klasterizavimo struktūra [Mar08]. Klasterių dendrogramos pavyzdys yra pateiktas 2 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 mokymosi metodai. Pristatant kombinuotąjį mokymąsi bus kalbama apie klasifikavimą, tačiau principai yra pritaikomi ir kitiems mašininio mokymosi metodams, pvz. matų atrinkio uždaviniams.

Kombinuotasis mokymasis pirmiausia yra naudojamas tam, kad pagerintų kuriamo klasifikatoriaus tikslumą arba sumažintų prasto klasifikatoriaus sukūrimo tikimybę. Prielaidos šiam teiginiui yra:

- 1.

1.5. Atraminių vektorių klasifikatoriai

Atraminių vektorių klasifikatoriai (angl. *support vector machines*, SVM) - tai mašininio mokymosi algoritmas, kuris gali būti taikomas tiek klasifikavimui, tiek regresinei analizei. Šis algoritmas priskiriamas prie mokymosi su mokytoju algoritmų [Vap00].

Atraminiai vektoriai (angl. *support vectors*) yra mėginiai esantys arčiausiai atskiriančios hiperplokštumos (angl. *decision boundary*). Atraminių vektorių klasifikatorių algoritmo tikslas yra mėginių erdvėje orientuoti atskiriančiąją hiperplokštumą tokiu būdu, galimai pašalinant triukšmą bei išimtis (angl. *outlier*), kad atstumas tarp jos ir artimiausių objektų iš abiejų klasių būtų didžiausias [CV95]. Atskiriančiosios tiesės pavyzdys pavaizduotas 3 pav.

Tarkime, kad turime L mokymosi objektų, kurių kiekvienas objektas 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:

$$\{x_i, y_i\}, \text{ kur } i = 1..L, y_i \in \{-1, 1\}, x \in \mathbb{R}^D \quad (1)$$

Tarkime, kad duomenys yra tiesiškai atskiriami. Tai reiškia, kad galima nupiešti tiesę grafe x_1 ir x_2 , kuri atskiria dvi klases, kai $D = 2$ ir hiperplokštumą grafuose x_1, x_2, \dots, x_D , kai $D > 2$. 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.

Taigi, atraminių vektorių klasifikatorių sukūrimas yra parametrų w ir b tenkinančių minėtas sąlygas radimas. Tai galima užrašyti tokia nelygybe:

$$y_i(x_i \cdot w + b) - 1 > 0 \quad (2)$$

Jei abiejų klasių objektai nėra tiesiškai atskiriami, reikia „atpalaiduoti“ (2) sąlygą įvedant parametą ξ_i :

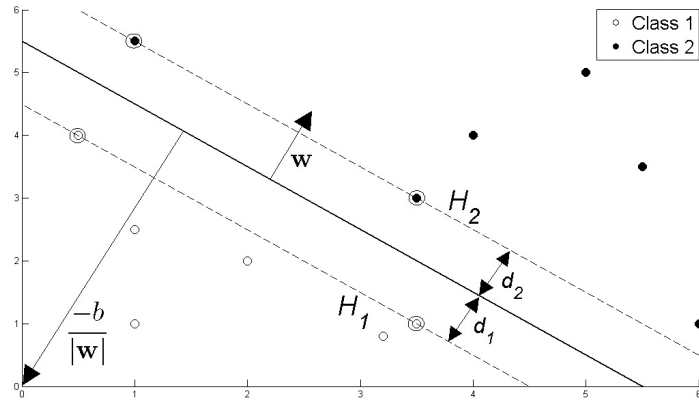
$$y_i(x_i \cdot w + b) - 1 + \xi_i > 0, \text{ kur } \xi_i \geq 0, \forall i, \quad (3)$$

kur ξ_i yra baudos dydis už neteisingai klasei priskirtą mėginį.

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

1.6. *Random Forest* klasifikatorius

Random Forrest klasifikatorius yra įrankis, kuris sukuria keletą klasifikavimo medžių (angl. *decision tree*), kurie visi nepriklausomai klasifikuoja mėginius, ir daugumos balsavimo (angl. *majority voting*) būdų yra skelbiamas galutinis klasifikavimo rezultatas [BFOS84].



3 pav.: Hiperplokštuma nubrėžta per dvi tiesiškai atskiriamas klases.

Toks daugelio klasifikatorių panaudojimas yra vadinamas kombinuotoju mokymusi (angl. *ensemble learning*).

Kiekvienas klasifikavimo medis yra konstruojamas pagal procedūrą aprašytą algoritme nr. 1.

Algoritmas nr. 1 *Random Forest* klasifikavimo medžių konstravimas

1. Turimas N mėginių, kurie turi M matų;
 2. Pasirenkamas m matų, kurie bus naudojami klasifikavimo medžių kūrimui; $m \ll M$;
 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*).
-

Random forest algoritmo tikslumas priklauso nuo: koreliacijos tarp sukurtų klasifikavimo medžių (didesnė koreliacija lemia mažesnį klasifikavimo tikslumą.); atskiro klasifikavimo medžio skiriamoji galia (kuo didesnė atskiro klasifikavimo medžio skiriamoji galia, tuo geresnis klasifikavimo tikslumas).

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

2. PAGRINDINIAI MATŲ ATRINKIMO METODAI

Dėl „daugiamatiškumo prakeiksmo“ (angl. *the curse of dimensionality*) – didėjant matų kiekiui mėginiai pasidaro panašūs, todėl bandymas juos klasifikuoti tolygus spėliojimui [Bel66]. Biomedicininį duomenų kontekste galima daryti prielaidą, kad ne visi matai yra susiję su tiriamąja problema, pvz. gaubtinės žarnos vėžiu, dėl to, kad duomenys yra daugiamatiai. Paprastai nagrinėjamai problemai svarbus yra mažas, palyginus su visu, matų kiekis. Todėl biomedicininį duomenų daugiamatiškumui sumažinti yra naudojami informatyviausių matų atrinkimo metodai [GE03] (angl. *feature selection*). Matų atrinkimas yra svarbi biomedicininį duomenų apdorojimo (angl. *preprocessing*) etapo dalis. Naudojant matų atrinkimo metodus, galima kovoti su daugiamatiškumo prakeiksmu matų skaičių priartinant prie mėginių skaičiaus. Todėl svarbu yra pasirinkti geriausiai tinkančią matų atrinkimo strategiją. Kadangi ir pačių matų atrinkimo metodų veikimas priklauso nuo konkrečių duomenų, tai metodo pasirinkimas yra sudėtinga užduotis.

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

1. 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.
2. Prisitaikantieji metodai (angl. *wrapper methods*). Pirma, apmokomas klasifikatorius su visais matais, antra, parenkamas matų poaibis ir apmokomas klasifikatorius. 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 – matai yra atrenkami pagal klasifikatoriaus darbo rezultatus. Prisitaikantieji metodai dažnai duoda geresnius rezultatus negu filtravimo metodai, bet yra reiklūs resursams.
3. 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 klasifi-

kavimo tikslumo ir skaičiavimų sudėtingumo.

Šiame skyriuje nagrinėsiu pagrindinius matų atsirinkimo metodus:

1. *Fisher* įvertis (angl. *Fisher ratio*)[PWCG01];
2. *Relief* metodas[RSK03];
3. Asimetrinis priklausomybės koeficientas (angl. *Asymmetric Dependency Coefficient, ADC*) [Sha01];
4. Absoliučių svorių SVM (angl. *Absolute Weight SVM, AW-SVM*) [Vap00];
5. Rekursyvus matų eliminavimas pagal SVM (SVM-RFE) (angl. *Recursive Feature Elimination by SVM*) [GWBV02b].

2.1. *Fisher* įvertis

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

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

kur, j – yra 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. Nors ir paprastas, 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 grupių, ir atnaujinamos visų matų „susietumo“ reikšmės. Dėl atsitiktinumo faktoriaus klasifikavimo ir matų atrinkimo stabilumo rezultatai naudojant šį metodą varijuoja. Mato įvertis yra vidurkis visų objektų atstumų skirtumų iki artimiausių kaimynų iš kitos ir tos pačios klasių:

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

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

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

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

2.3. Asimetrinis priklausomybės koeficientas

Asimetrinis priklausomybės koeficientas (angl. *asymmetric dependency coefficient*, ADC) yra matų reitingavimo motodas, kuris matuoja mėginio grupės tikimybinę priklausomybę j -ajam matui, naudodamas informacijos prieaugį (angl. *information gain*) [Ken83]:

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

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

$$H(Y) = - \sum_y p(Y = y) \log p(Y = y), \quad (8)$$

$$H(X_j) = - \sum_x p(X_j = x) \log p(X_j = x), \quad (9)$$

$$MI(Y, X_j) = H(Y) + H(X_j) - H(Y, X_j), \quad (10)$$

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

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

2.4. Absoliučių svorių SVM

Atraminių vektorių klasifikatorius (SVM) yra vienas populiariausių klasifikavimo algoritmų, nes jis gerai susidoroja su daugiamačiais duomenimis [GWBV02a]. 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:

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

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

2.5. 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 [GWBV02a]. Todėl, jis yra naudojamas kaip atskaitos taškas (angl. *benchmark*) vertinant kitus matų atrankos metodus. Iš esmės šis metodas yra daugkartinis absoliučių svorių SVM metodo taikymas nuolat išmetinėjant matus su mažiausiais svoriais. Rekursyvus matų eliminavimas mums padeda surasti tinkamą matų poaibį, kas ne visada pavyksta su matų reitingavimo metodais. Bendroji rekursyvaus matų eliminavimo procedūra: Jei trečiajame

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 antrąją žingsnį kol nėra patenkinta algoritmo pabaigos sąlyga.
-

algoritmo žingsnyje iš matų aibės yra pašalinamas tik viena matas, tai gauname matų reitingavimą, o jei pašalinamas fiksuotas skaičius ar dalis (pvz. 50%) matų, tai matų reitingavimo negauname. Pastebėtina, kad rekursyvus matų eliminavimas labai padidina algoritmo sudėtingumą. Algoritmo pabaigos sąlyga gali būti koks nors konkretus matų skaičius arba tiesiog matų aibę mažiname tol, kol matų visai nebeliks.

3. STABILIŲ MATŲ ATRINKIMO METODAI

Naudodajant matų atrinkimo metodus, biomedicininis 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 tikėtina, kad dalis matų koreliuoja, todėl, kuris iš koreliuojančių matų bus pasirinktas, priklauso nuo to, kuriuos mėginius pasirinksimė 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].

Vienas iš būdų didinti matų atrinkimo stabilumą yra naudoti multikriterinius matų atrinkimo metodus. Jų 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 optimalių hipotezių gali būti teisingos, ir kriterijų agregavimas sumažina tikimybę, kad bus pasirinkta neteisinga hipotezė;
2. Atskiri matų atrinkimo metodai gali dirbti skirtinguose lokaliuose optimumuose, tuo tarpu agregavimas gali geriau reprezentuoti tikrąją duomenis generuojančią funkciją;
3. Tikroji duomenų funkcija negali būti reprezentuojama jokia hipoteze paskiro algoritmo hipotezių erdvėje ir agreguojant pavienių metodų rezultatus galima praplėsti hipotezių erdvę.

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 metodus:

1. Svoriais grįstas multikriterinis suliejimas;
2. Reitingais grįstas multikriterinis suliejimas;
3. Svoriais ir reitingais grįstas multikriterinis suliejimas;
4. Multikriterinis rekursyvus matų eliminavimas;
5. Konsensuso grupėmis grįstas stabilių matų atrinkimo metodas.

3.1. Matų atrinkimo stabilumas

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

Stabilių matų atrinkimo problematika yra populiarėjanti tyrimų kryptis. Stabilumo aktualumas yra sąlygotas fakto, kad biomediciniuose duomenyse konkrečiai problemai yra aktualūs tik tam tikri matai. Todėl dalykinės srities ekspertams yra svarbu naudoti tuos matų atrinkimo metodus, kurie yra stabilūs ir susiję su modeliuojama problema, nes tai atpigina duomenų analizę.

Svarbu paminėti, kad matų stabilumas nėra matuojamas visiškai nepriklausomai – jis yra matuojamas atsižvelgiant į klasifikavimo rezultatus. Matuoti stabilumą verta tada, kai atrinkami matai duoda gerus klasifikavimo rezultatus. Kitaip tariant, nėra naudingi tie matų atrinkimo metodai, kurie duoda labai stabilius rezultatus, bet jais remiantis atrinktais matais pavyksta sukurti tik atsitiktinius rezultatus duodančius klasifikatorius.

3.1.1. Stabilumo matavimas

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 sąrašų porų tarpusavio panašumo įverčių [KPH07]:

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

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

Matų atrinkimo stabilumo įvertis priklauso nuo to, kokią aibių panašumo funkciją naudosime. Tradicinės panašumo funkcijos (persidengimo procentas, *Pearson* koreliacija, *Spearman* koreliacija) gali būti taikomos, bet jos yra linkusios priskirti didesnes panašumo reikšmes, kai pasirenkamas didesnis matų poaibis. Taip yra dėl padidėjusio sisteminio nuokrypio (ang. *bias*), nes imant didesnę 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:

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

kur $s = |f_i| = |f_j|$ yra atrinktų matų aibės dydis, $r = |f_i \cap f_j|$ - abiem 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 didesnę panašumą, o artimos nuliui reikšmės reiškia, kad matai atrenkami daugiausia atsitiktinai. *Kuncheva* indekso ypatybė yra ta, kad jis atsižvelgia tik į persidengiančias, tačiau visiškai nekreipia dėmesio į koreliuojančius matus.

3.1.3. *Jaccard* indeksas

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

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

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. Kitaip tariant, *Hamming* atstumas yra minimalus skaičius pakeitimų, kad vieną vektorių padarytume lygų kitam.

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

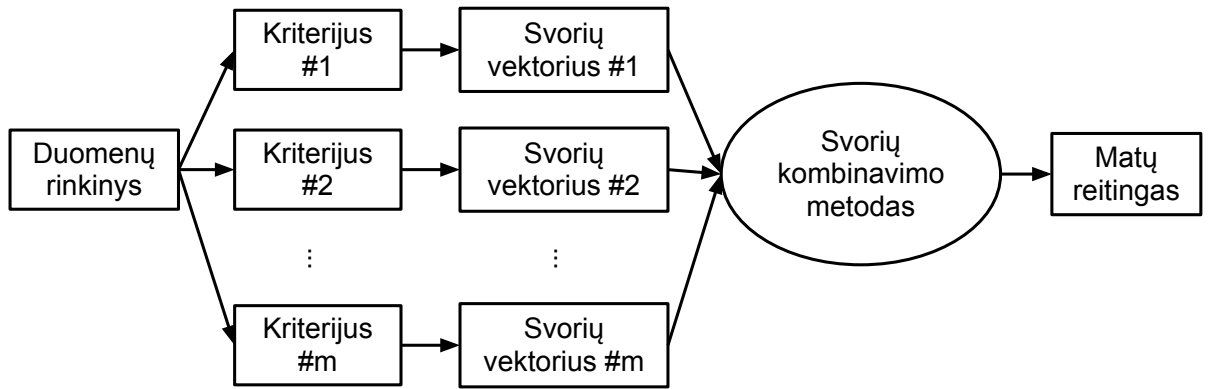
kur \oplus - sumos moduliu 2 arba XOR operacija.

Šiuo metodu matuojant matų atrinkimo stabilumą, prieš atstumo matavimą reikia atlikti pertvarkymus. Pirma, iš atrinktų matų vektorių padaryti bendro matų skaičiaus ilgio binarinius vektorius. Antra, vienetukus sudėti tose vektoriaus elementuose, kurių indeksai gaunami matų atrinkimo metodu. Tada jau galima matuoti atstumą tarp dviejų matų atrinkimo rezultatų.

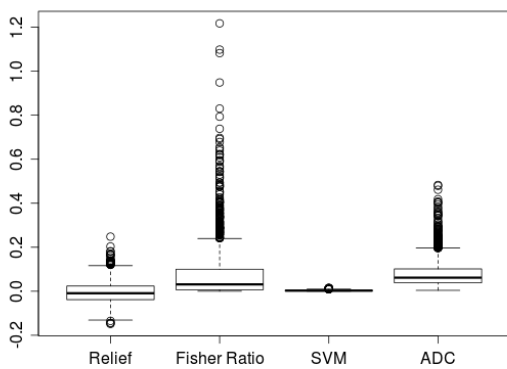
3.2. Svoriais grįstas multikriterinis suliejimas

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 4 pav.

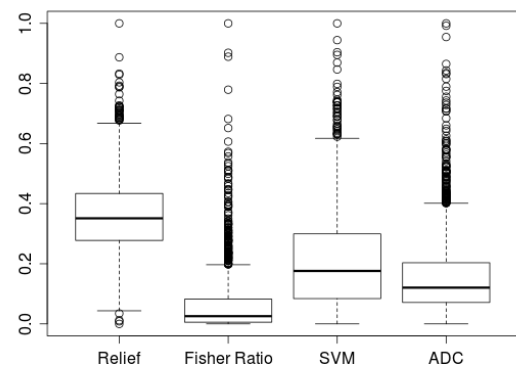
Suliejant svorius svarbu yra užtikrinti, kad svoriai, gauti naudojant skirtingus bazinius kriterijus, būtų palyginami. Todėl svorių normalizavimas turi būti atliekamas prieš svorių kombinavimą. Kitu atveju matų įvertinimai bus nepalyginami. Paveikslėlyje 5 pav. nenormalizuotų pavienių matų vertinimo metodų skiriasi netgi suteiktų svorių intervalai.



4 pav.: Svoriais grįstas multikriterinis suliejimas.



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



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

Paveikslėlyje 6 pav. matome, kad net ir normalizavus svorius skiriasi svorių kvartiliai – į tai reikia atkreipti dėmesį interpretuojant galutinius matų vertinimo rezultatus. Šiame darbe svoriai yra normalizuoti intervale $[0, 1]$ pagal formulę:

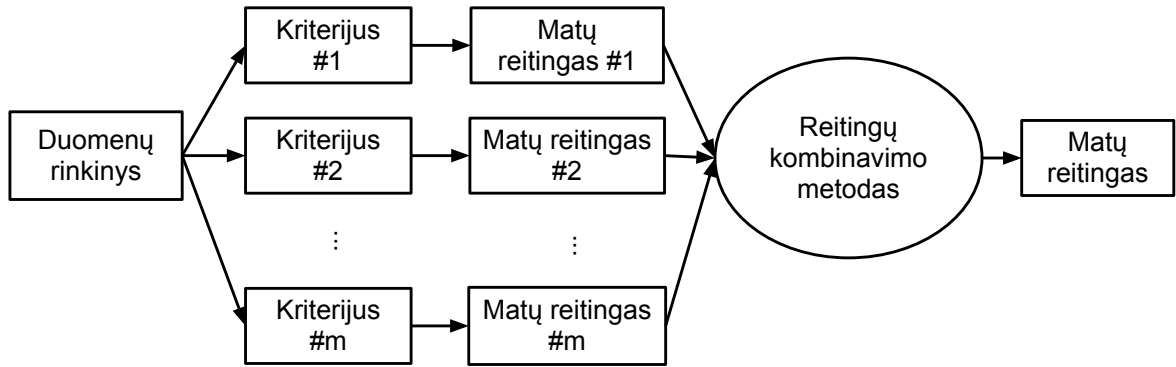
$$u'_i = \frac{u_i - u_{i_{\min}}}{u_{i_{\max}} - u_{i_{\min}}}, \quad (17)$$

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:

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

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



7 pav.: Reitingais grįstas multikriterinis suliejimas.

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 bendrą matų reitingą. Algoritmas yra pavaizduotas 7 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 grąžinti matų reitingus, o ne svorius.

Matų reitingų kombinavimui yra keletas metodų [DKNS01], tačiau paprastumo dėlei šiame darbe naudosiu Borda balsavimą³ (angl. *Borda count*). Tarkime, kad turime m balsuotojų ir p kandidatų aibę. Tada Borda balsavimo metodas kiekvienam i -ajam balsuotojui sukuria balsų vektorius 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

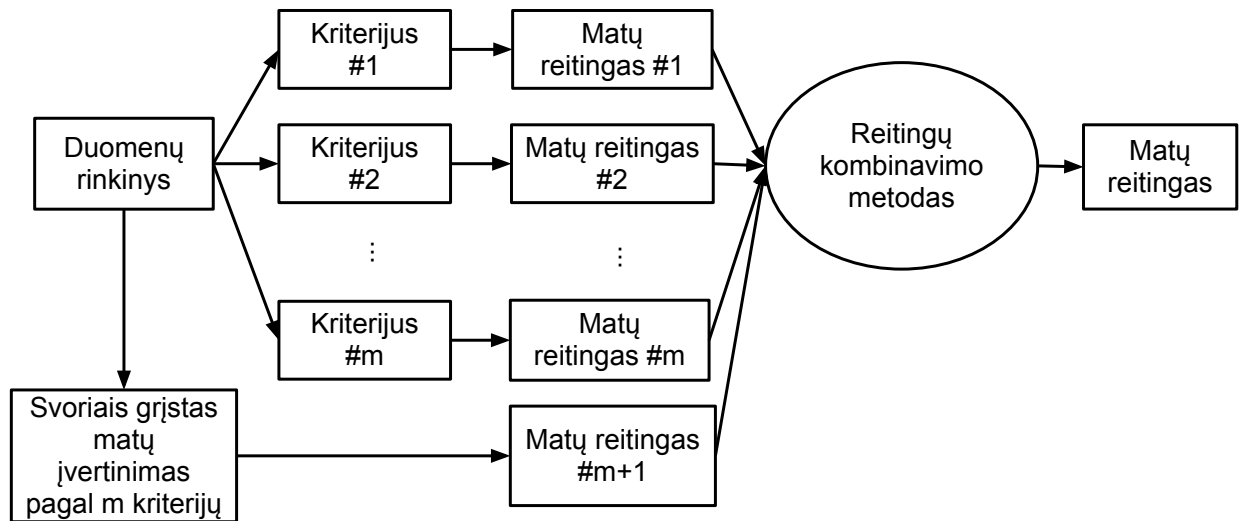
$$v = \sum_{i=1}^m v_i, \quad (19)$$

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 panaudo-

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



8 pav.: Svoriais ir reitingais grįstas multikriterinis suliejimas.

jamas svoriais grįsto multikriterinio matų atrinkimo metu gautas reitingas. Multikriterinio matų įverčių ir pagal svorius, ir pagal reitingus metodas vyksta trimis žingsniais:

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

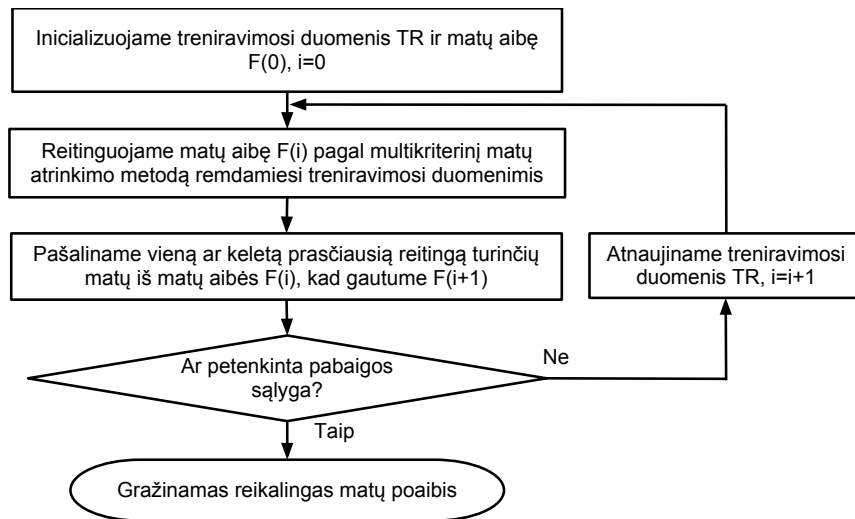
Algoritmas yra pavaizduotas 8 pav.

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

3.5. Multikriterinis rekursyvus matų eliminavimas

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

Multikriterinis rekursyvus matų eliminavimas susideda iš dviejų dalių [YM11]: keletos



9 pav.: Multikriterinio rekursyvaus matų eliminavimo algoritmas.

matų atrinkimo kriterijų suliejimo pagal svorius ir pagal reitingus, ir rekursyvaus matų eliminavimo aprašyto algoritme nr. 2. Algoritmas pavaizduotas 9 pav.

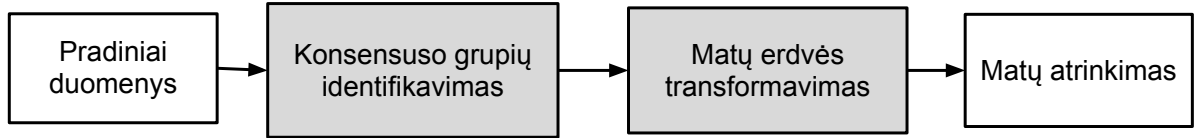
Standartinis rekursyvus matų eliminavimas, kai vienos iteracijos metu yra eliminuojamas vienas matas, gali labai padidinti algoritmo sudėtingumą. Todėl genų ekspresijos duomenims prasmingiau yra eliminuoti keletą matų vienu metu.

Nors SVM-RFE matų atrinkimo algoritmas ir yra labai populiarus, tačiau yra žinoma, kad jam trūksta stabilumo [GWBV02a]. Todėl kombinuodami didesnį stabilumą turintį multikriterinį matų atrinkimą su rekursyvaus matų eliminavimo paieškos strategija, gauname stabilesnį matų atrinkimo algoritmą.

3.6. Konsensuso grupėmis grįstas stabilų matų atrinkimo metodas

Konsensuso grupėmis grįstas stabilų matų atrinkimo metodas(angl. *Consensus Group Stable feature selection*, CGS), pirma, identifikuoja panašių matų grupes, antra, pagal surastas grupes transformuoja matų aibę, trečia, transformuotoje matų aibėje atlieka matų atrinkimą [LYD09]. Schematiškai šis algoritmas pavaizduotas 10 pav.

CGS metodo pagrindinė dalis yra panašių matų identifikavimas. Šio uždavinio sprendimui naudojamas *Dense Group Finder* (DGF) algoritmas. DGF aprašytas algoritme nr. 3. CGS algoritme pagal matus pagal DGF algoritmą yra sugrupuojami keletą kartų. Po pakartotinio grupavimo yra ieškoma stabilų grupių – jei matas buvo sugrupuotas į konkrečią



10 pav.: Konsensuso grupėmis grįstas stabilų matų atrinkimas.

grupę daugiau nei pusėje grupavimų, tai matas ir priklausys tai konsensuso grupei. Matų aibės transformavimas vyksta iš kiekvienos konsensuso grupės išrenkant reprezentatyviausią matą – konkretų matą esantį arčiausiai konsensuso grupės vidurkio. Išrinktieji reprezentatyviausieji matai ir sudaro transformuotą matų aibę. Transformuotoje matų aibėje vykdomas matų antrinkimas kuriuo nors matų atrinkimo metodu Φ , pavyzdžiui, *Relief* matų atrinkimo metodu.

Algoritmas nr. 3 DGF – *Dense Group Finder*

Ieitis: duomenys $D = \{x_i\}_{i=1}^n$, branduolio plotis h
Išeitis: tankios matų grupės G_1, G_2, \dots, G_L
for $i = 1$ **to** n **do**
 Inicializuojame $j = 1, y_{i,j} = x_i$
 repeat
 Suskaiciuoti 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 $\|p_r - x_i\| < h$

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

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

Algoritmas nr. 4 Konsensuso grupėmis grįstas stabilių matų atrinkimas

Įėjimas: mėginių aibė D , iteracijų skaičius t , matų atrinkimo metodas Φ

Išėjimas: atrinktos konsensuso matų grupės CG_1, CG_2, \dots, 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, \dots, 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, \dots, CG_L pagal $\Phi(X_i)$

Pasirinkti k matų, turinčių geriausią reitingą

4. EKSPERIMENTAI

Šiame skyriuje yra aprašyti daugiamačių duomenų klasifikavimo analizės metu atliktų eksperimentų rezultatai. Skyrius susideda iš eksperimentuose naudotų biomedicininio duomenų rinkinių aprašymo, eksperimentų nustatymų įvardinimo, matų atrinkimo metodų spartos matavimų įvertinimo, klasifikavimo tikslumo matavimų rezultatų pristatymo, bei stabilių matų atrinkimo rezultatų pristatymo.

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.

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

Lentelės tęsinys kitame puslapyje...

lentelė 1 – Tęsinys

Pavadinimas	Šaltinis	Mėginių skaičius (+/-)	Matų skaičius	OMS
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	[Ins]	90 (bp ⁴ : sz ⁵ : cc ⁶ =30:31:29)	22283	0.00404

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.

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 parametru 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.

⁴bp (angl. *Bipolar disorder*) - maniakinė depresija sergantys pacientai.

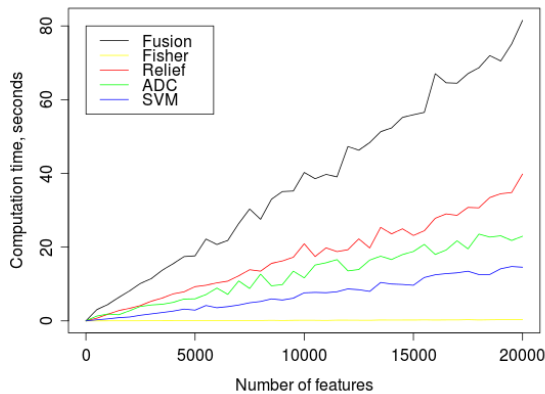
⁵sz (angl. *Schizophrenia*) - šizofrenija sergantys pacientai.

⁶cc (angl. *Control Crowd*) - kontrolinė grupė.

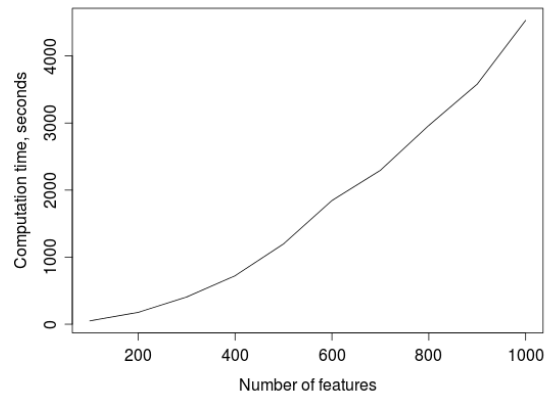
Kiekvienoje iteracijoje su vis kitu mėginių poaibiu buvo atliekamas matų atrinkimas, po to būdavo atliekamas klasifikavimas su 10, 20, ..., 500 aukščiausią reitingą turinčių matų.

4.3. Matų atrinkimo metodų sparta

Matų atrinkimo metodų darbo laikas buvo palygintas naudojant vieną biomedicininį duomenų rinkinį - AltarA [Ins]. Skaičiavimai buvo atlikti kompiuteryje naudojant vieną procesoriaus branduolį veikiantį 2.66 GHz, bei 2 GB RAM atminties. 11 pav. ir 12 pav. pavaizduota matų atrinkimo metodo darbo laiko priklausomybė nuo mėginius apibūdinančių matų skaičiaus.



11 pav.: Pagrindini matų atrinkimo metodų darbo laikas.



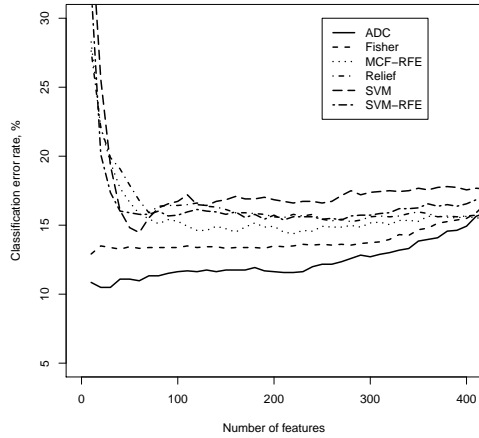
12 pav.: Konsensuso grupėmis grįsto matų atrinkimo metodo darbo laikas.

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 11 pav. galime daryti išvadą, kad sparčiausias matų atrinkimo metodas yra *Fisher* įvertis. Pagal gautus matų darbo laiko priklausomybės nuo matų kiekio grafikus galime daryti išvadą, kad CGS algoritmas daugiamačių duomenų matų atrinkimui nėra tinkamas, nes jo skaičiavimų laikas yra per ilgas.

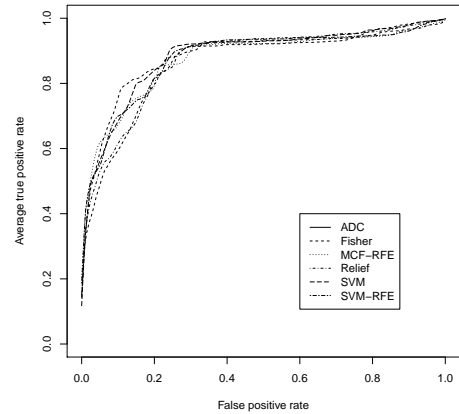
4.4. Klasifikavimo pagal atrinktus matus tikslumas

Matų atrinkimo metodų įtaką klasifikavimo tikslumui buvo matuojama naudojant tris biomedicininį duomenų rinkinius: Gaubtinės žarnos auglio (angl. Colon)

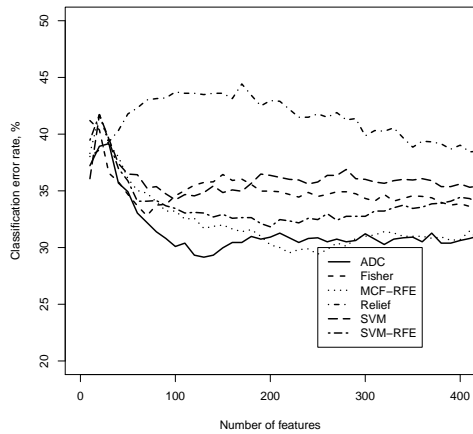
[ABN⁺99], Centrinės nervų sistemos (CNS) [PTG⁺02], prostatos [SFR⁺02]. Klasifikavimui buvo naudojami tiesiniai atraminių vektorių klasifikatoriai (SVM) [Vap00], su parametru $C = 0.01$, kurį nustačiau empiriškai. Keičiant parametrus keičiasi ir klasifikavimo tikslumas. Klasifikatoriui apmokyti buvo naudojama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



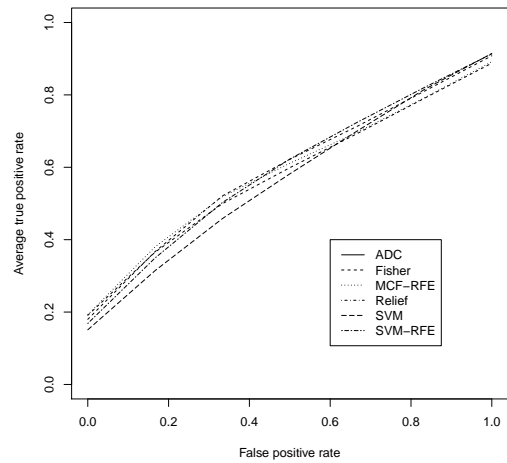
13 pav.: Gaubtinės žarnos auglio mėginių klasifikatorių tikslumas.



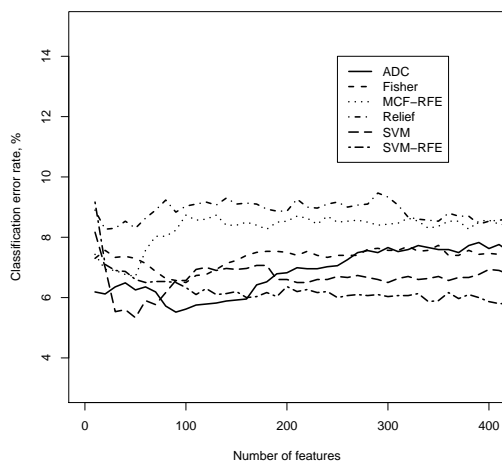
14 pav.: Gaubtinės žarnos auglio mėginių klasifikatorių ROC kreivės.



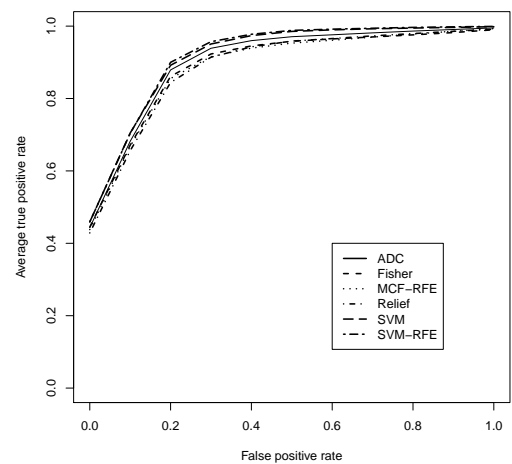
15 pav.: Centrinės nervų sistemos mėginių klasifikatorių tikslumas.



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



17 pav.: Prostatos mėginių klasifikatorių tikslumas.



18 pav.: Prostatos mėginių klasifikatorių ROC kreivės.

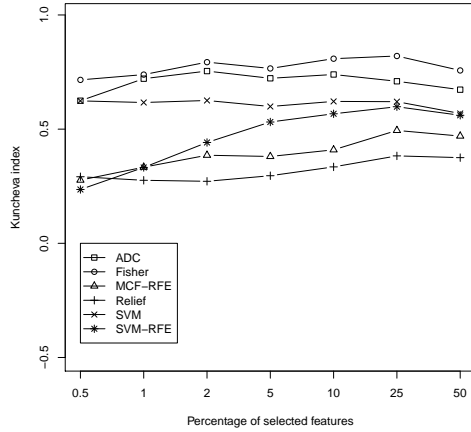
13 pav. matome, kad gaubtinės žarnos auglio duomenų rinkinio matus geriausiai atrenka ADC metodas. Tik šiek tiek prasčiau pasirodo *Fisher* įvertis. Blogiausiai su gaubtinės žarnos auglio mėginiais susidoroja absoliučių svorių SVM matų atrinkimo metodas.

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%. 15 pav. matome, kad šiam duomenų rinkiniui vidutiniškai geriausiai matus atrenka ADC ir multikriterinio rekursyvaus matų eliminavimo metodai. Prasčiausiai pasirodo *Relief* metodas. 17 pav. matome, kad prostatos duomenų rinkinio matus klasifikavimui geriausiai atrenka ADC absoliučių svorių SVM metodas. Prasčiausiai matus atrenka *Relief*.

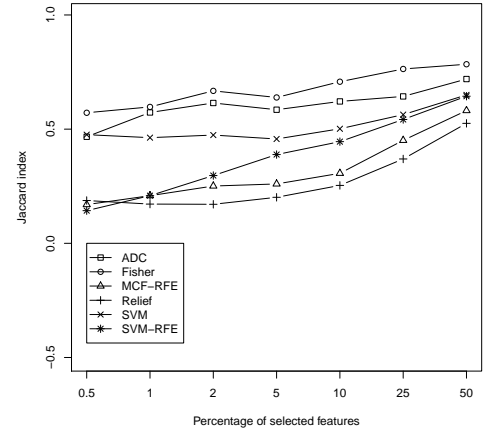
Apibendrinamas gautus klasifikavimo tikslumo matavimo rezultatus, galiu teigti, kad nėra vieno absoliučiai geriausio matų atrinkimo metodo. Reikia eksperimentuoti, kad būtų rastas konkrečiai problemai geriausiai tinkantis matų atrinkimo metodas. Tačiau rezultatai parodė, kad matų atrinkimas svariai prisideda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

4.5. Matų atrinkimo stabilumas

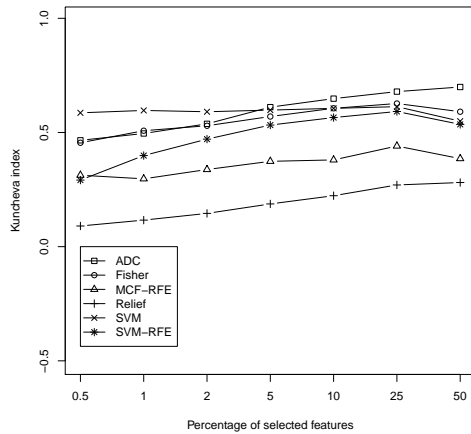
Matų atrinkimo stabilumas buvo tiriamas naudojant tuos pačius biomedicininį duomenų rinkinius kaip ir tiriant klasifikavimo pagal atrinktus matus tikslumą. Matų atrinkimo stabilumas buvo matuojamas pagal *Kuncheva* ir *Jaccard* indeksus. Stabilumas pats savaime nėra svarbus, jis turi būti matuojamas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



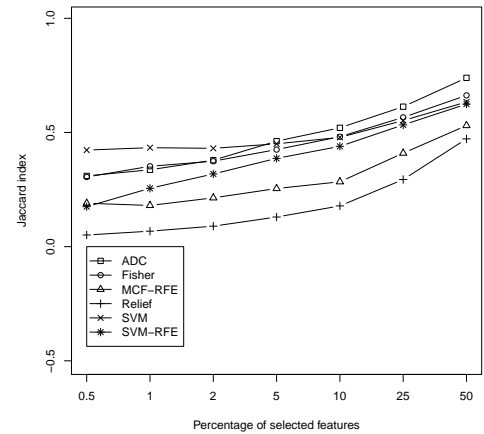
19 pav.: Matų atrinkimo gaubtinės žarnos auglio mėginiams stabilumo grafikas pagal Kuncheva indeksą.



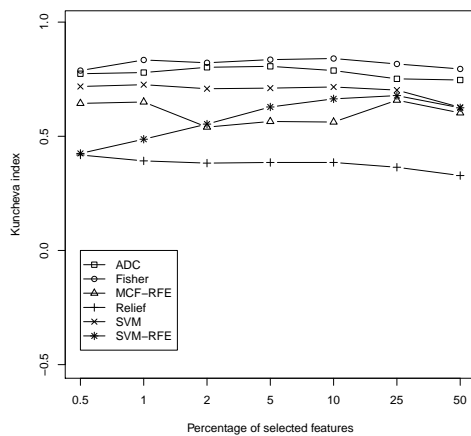
20 pav.: Matų atrinkimo gaubtinės žarnos auglio mėginiams stabilumo grafikas pagal Jaccard indeksą.



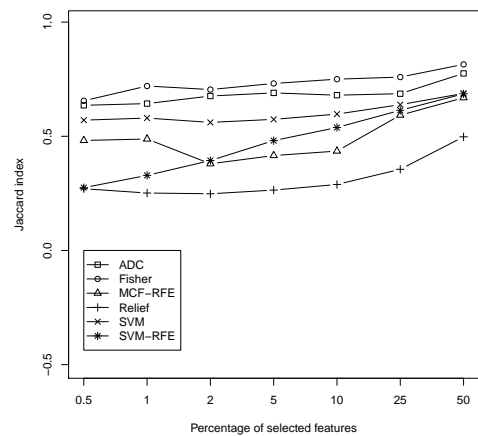
21 pav.: Matų atrinkimo CNS mėginiams stabilumo grafikas pagal Kuncheva indeksą.



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



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



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

Pagal 19 pav. ir 20 pav. matome, kad gaubtinės žarnos auglio duomenų rinkinio matus stabiliausiai atrenka *Fisher* įvertis. Mažiausiai stabiliai matus atrenka *Relief* metodas.

Pagal 23 pav. ir 24 pav. matome, kad prostatos duomenų rinkinio matus stabiliausiai atrenka *Fisher* įvertis. Mažiausiai stabiliai matus atrenka *Relief* metodas.

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

REZULTATAI IR TOLIMESNIŲ TYRIMŲ KRYPTYS

Čia bus rezultatai ir išvados. Ir nutarta, kad čia bus aprašytos tolimesnių tyrimų kryptys.

Literatūra

- [ABN⁺99] U. Alon, N. Barkai, D.A. Notterman, K. Gish, S. Ybarra, D. Mack, and A.J. Levine. Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *Proceedings of the National Academy of Sciences*, 96(12):6745, 1999.
- [Bel66] R.E. Bellman. *Adaptive control processes: A guided tour*. Princeton University Press, 1966.
- [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.
- [CV95] C. Cortes and V. Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.
- [DHS] R.O. Duda, P.E. Hart, and D.G. Stork. *Pattern classification*. Pattern Classification and Scene Analysis: Pattern Classification. <http://books.google.lt/books?id=YoxQAAAAAAAJ>.
- [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.

- [GE03] I. Guyon and A. Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
- [GS66] D.M. Green and J.A. Swets. *Signal detection theory and psychophysics*, volume 1974. Wiley New York, 1966.
- [GWBV02a] 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.
- [GWBV02b] Isabelle Guyon, Jason Weston, Stephen Barnhill, and Vladimir Vapnik. Gene selection for cancer classification using support vector machines. *Mach. Learn.*, 46(1-3):389–422, March 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.
- [Ins] Stanley Medical Research Institute. Online genomics database. [žiūrēta 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. Vaizdy 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.
- [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.
- [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.
- [tzz10] *Tarptautinių žodžių žodynas*. Vyriausioji enciklopedijų redakcija, 2010.
- [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.
- [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.

SAVOKŲ APIBRĖŽIMAI

Klasifikavimas –

Klasifikatorius –

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

Genėjimas (klasifikavimo medžių) (angl. *pruning*) – technika, kuri iš klasifikavimo medžių pašalina mazgus, kurie turi salyginai mažą atskiriamąją galią.

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

Normalės vektorius – vektorius, kuris yra statmenas tiesei arba plokštumai.

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

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

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

Atraminių vektorių klasifikatoriai (angl. support vector machines, SVM) – mokymosi su mokytoju metodas, taikomas ir klasifikavime, ir regresinei analizei.

Regresija [lot. regressio — grįžimas, traukimasis] – tikimybių teorijoje ir mat. statistikoje – atsitiktinio dydžio vidurkio priklausomybės nuo kt. dydžio (kelių dydžių) išraiška [tzz10].

Mokymosi duomenys (angl. *training data*) – duomenys, su kuriais dirbs mašininio mokymosi algoritmai.

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