VILNIAUS UNIVERSITETAS
MATEMATIKOS IR INFORMATIKOS FAKULTETAS
INFORMATIKOS KATEDRA

TIESIOSIOS ŽARNOS VĖŽIO EPIGENETINIŲ MODIFIKACIJŲ TYRIMAS NAUDOJANT MAŠININIO MOKYMOSI METODUS

Study of Epigenetic Modifications in Colorectal Cancer Using Machine Learning Methods

Magistro baigiamasis darbas

Atliko:	Ieva Meržvinskaitė	(parašas)
Darbo vadovas:	Dr. Juozas Gordevičius	(parašas)
Recenzentas:	Lekt. dr. Valdas Dičiūnas	(parašas)

Vilnius – 2016

Santrauka

Magistrinio darbo tikslas taikant mašininio mokymosi metodus ištirti, ar mikrogardelių metodu gauti cirkuliuojančios DNR ir leukocitų DNR metilinimo duomenys tinkami vėžio diagnostikai.

Tyrimo metu analizuotos dvi skirtingos duomenų aibės. Pirmoji aibė – tiesiosios žarnos vėžio cirkuliuojančios DNR metilinimo duomenys gauti tiriant sveikų ir sergančių pacientų mėginius su *GeneChip®Human Tiling 2.0R Array Set* mikrogardelėmis. Antroji duomenų aibė – apjungti trijų skirtingų eksperimentų duomenys imti iš *Gene Expression Omnibus* internetinės duomenų bazės, kuriuose buvo tirtas kraujyje esančių leukocitų DNR metilinimas šlapimo pūslės vėžio atvejais, galvos ir kaklo suragėjusių ląstelių karcinomos atvejais ir kiaušidžių vėžio atvejais. Visuose trijuose eksperimentuose periferinio kraujo mėginiai tirti mikrogardelių metodu naudojant *Illumina Infinium 27k Human DNA methylation Beadchip v1.2* mikrogardeles. Magistriniame darbe siūlomi duomenų apdorojimo ir analizės žingsniai šiems duomenims tirti. Analizėje taikomi statistiniai ir mašininio mokymosi metodai (atsitiktiniai miškai, atraminių vektorių mašinos). Taikant suformuluotą duomenų apdorojimo ir analizės strategiją tiesiosios žarnos vėžio duomenims sveiki ir sergantys pacientai atskiriami su 0.79 testavimo tikslumu. Taikant suformuluotą duomenų apdorojimo ir analizės strategiją kitų apjungtų eksperimentų duomenims tirti, skirtingas vėžio formas galima atskirti su 0.7 testavimo tikslumu.

Apibendrinant, cirkuliuojančios DNR metilinimo duomenys yra labiau tinkami vėžio diagnostikai nei leukocitų DNR metilinimo duomenys. Keltą hipotezę, kad cirkuliuojančios DNR metilinimo duomenyse esantis signalas yra ne tik iš vėžinių ląstelių DNR, bet ir iš žuvusių kraujo ląstelių DNR, patikrinti būtų galima atliekant tolesnius tyrimus.

Raktiniai žodžiai: Atsitiktiniai miškai, Atraminių vektorių mašinos, Mašininis mokymasis, Mikrogardelių duomenų analizė

Summary

The aim of my Master thesis is to find out, by applying machine learning algorithms, whether the obtained circular DNA methylation data and blood cells' DNA methylation data can be used for the diagnosis.

In the analysis, two different datasets were investigated to find out whether they can be used in order to separate samples by cancer type or by diagnosis. The first dataset is from colorectal cancer study where circular DNA methylation levels from healthy-unhealthy patients' blood samples were analyzed using *GeneChip®Human Tiling 2.0R Array Set* microarrays. The combined data from the other three datasets (bladder cancer, head and neck squamous cell carcinoma, and ovarian cancer) are from *Gene Expression Omnibus* database. Peripheral blood samples were analyzed using the *Illumina Infinium 27k Human DNA methylation Beadchip v1.2* microarrays. The routine steps for pre-processing and analysis of the data are suggested in the present research. Statistical and machine learning methods like *Random Forests* and *Support vector machines* were used in the analysis. The results show that by applying the first analysis strategy it is possible to separate case-control samples with 0.79 testing accuracy in colorectal cancer data, and that by using the second strategy it is possible to separate different types of cancer with 0.7 accuracy, and separate case-control samples in combined datasets from the other three studies.

In conclusion, the circular DNA methylation data from colorectal cancer study is more applicable for the diagnosis than the blood-based DNA methylation data from the other three studies. The hypothesis was that the signal comes mostly from circular DNA of dead blood cells rather than from circular tumor DNA in colorectal cancer data. This statement can be proved by carying out further analysis on the matter.

Keywords: Random Forests, Support vector machines, Machine learning, Microarray data analysis

Turinys

Įvadas

Remiantis Lietuvos sveikatos statistikos duomenimis 2012 m. 100 000 gyventojų teko 271,9 piktybinių navikų sukeltos mirtys. Tai sudarė 21,6% pagrindinių mirties priežasčių tarp vyrų ir 17,4% tarp moterų [GMT13]. Pasaulyje 2012 m. nuo vėžio mirė 8,2 milijonai žmonių, tarp kurių 694 000 nuo tiesiosios žarnos vėžio. Tiesiosios žarnos vėžys yra viena iš pagrindinių mirtį sukeliančių vėžio formų pasaulyje [WHO14]. Laiku nustačius paciento sergamumą, jo pasveikimo šansai padidėtų iki 95%. Tačiau kol kas nėra greito ir patikimo neinvazinio metodo, kuris padėtų paprastai ir tiksliai nustatyti sergamumą vėžiu.

Vėžio susidarymas iš pradinės nepiktybinės neoplazmos ir virtimas piktybiniu augliu yra tiriamas nagrinėjant genetinius ar epigenetinius pakitimus DNR sekoje [KLS10]. Genetiniai pakitimai apima mutacijas onkogenuose (genai lemiantys vėžio susidarymą) ir vėžio susidarymą slopinančiuose genuose (pav. DNR taisymo aparato genuose). Vėžio susidarymas ir vystymasis taip pat gali būti lemiamas epigenetinių mechanizmų, paveldimų geno funkcijos pokyčių, kurie nėra susiję su pokyčiais DNR sekoje [KLS10].

Vienas iš epigenetinių pakitimų yra DNR metilinimas, cheminė DNR molekulės modifikacija, kurios metu metilo grupė (CH3) prijungiama nukleotido citozino aromatiniame žiede (bazėje) prie 5 anglies atomo, ir histonų (baltymai, reikalingi DNR supakavimui ląstelės branduolyje) modifikacijos. DNR metilinimo ir histonų modifikacijos pakitimai dažnai pasitaiko vėžinėse ląstelėse [Est07]. Šios DNR modifikacijos, dažniausiai lemiančios genų raiškos slopinimą, susijusios viena su kita ir gali lemti somatinės ląstelės virtimą vėžine [CB09].

Išskiriami du DNR metilinimo atvejai: hipometilinimas ir hipermetilinimas. Hipometilinimas dažnas daugelyje vėžio formų, ypač vėlyvesnėse stadijose, ir būna paplitęs visame genome. Genų promotorių hipometilinimas yra susijęs su padidėjusia transkripcija. DNR hipermetilinimas pasireiškia specifinėse promotorių srityse ar pasikartojančiose sekose ir yra specifiškas įvairioms vėžio formoms, turinčioms tam tikrus patologinius, klinikinius ir molekulinius požymius. Daugelyje vėžio atvejų pastebėtas padidėjęs citozino metilinimas CpG lokusuose greta vėžį slopinančių genų. Taip blokuojama šių genų transkripcija ir jie inaktyvuojami [KLS10].

Magistriniame darbe analizuojami DNR metilinimo duomenys gauti iš kraujyje cirkuliuojančios DNR (cirDNR), kurią sudaro mirusių kraujo ląstelių, kitų audinių ląstelių, tarp jų ir vėžinių ląstelių DNR patekusi į kraują. DNR metilinimo duomenys gauti mikrogardelių metodu tiriant pacientų kraujo mėginius ir juose nustatant cirkuliuojančios DNR metilinimo lygį įvairiose genomo vietose. CirDNR leidžia neinvaziniu būdu atpažinti vėžinės ląstelės DNR [CKL$^+$12]. Nustačius cirDNR metilinimo pakitimus bus galima diagnozuoti tiesiosios žarnos vėžį iš kraujo mėginio. Toks metodas galėtų tapti rutininiu naudojamu medicinos įstaigose ir padedančiu greitai nustatyti, ar pacientas serga.

Siekiant patikrinti, ar vėžiui specifiniai pakitimai randami tik cirDNR, ar gali būti randami ir kraujo ląstelių DNR, naudojant panašius metodus, kuriais tirti tiesiosios žarnos vėžio atvejai, kraujyje esančios cirkuliuojančios DNR metilinimo duomenys, analizuojami ir 3 skirtingų eksperimentų duomenys [LHA$^+$14] (duomenys imti iš *Gene Expression Omnibus*

internetinės duomenų bazės). Juose buvo tirtas kraujyje esančių leukocitų DNR metilinimas šlapimo pūslės vėžio atvejais (angl. *Bladder cancer*), galvos ir kaklo suragėjusių ląstelių karcinomos atvejais (angl. *Head and neck squamous cell carcinoma (HNSCC)*) ir kiaušidžių vėžio atvejais (angl. *Ovarian cancer*). Visuose trijuose eksperimentuose periferinio kraujo mėginiai tirti mikrogardelių metodu naudojant *Illumina Infinium 27k Human DNA methylation Beadchip v1.2* gardeles.

Mikrogardelių metodu gautų duomenų tyrimas – sudėtingas uždavinys, todėl pirmiausia suformuluojama duomenų analizės strategija. Duomenys apdorojami ir analizuojami panaudojant literatūros apžvalgoje nagrinėjamus požymių atrinkimo, klasifikavimo metodus. Duomenų analizės tikslas išsiaiškinti, ar turint mikrogardelių duomenis įmanoma atskirti sveikus ir sergančius pacientus, ar įmanoma atskirti skirtingas vėžio formas. Analizės rezultatas, klasifikatorius, atskiriantis sveikus pacientus nuo vėžiu sergančių naudojant metilinimo duomenis ir jo tikslumo įvertinimas.

1 Sprendžiama problema

Magistriniame darbe tiriami cirkuliuojančios DNR duomenys – 393 pacientų mėginiai, kurių 193 serga tiesiosios žarnos vėžiu (CRC), o likę 200 neserga – kontrolinė grupė (CTR). Požymių (p) aibė – cirDNR metilinimo lygis įvairiose genomo vietose. Duomenys gauti naudojant *GeneChip®Human Tiling 2.0R Array Set* mikrogardelių technologiją. Iš viso 393 mėginiai (N) ir daugiau nei 6.5 milijono požymių.

Kiti tiriami duomenys – kraujo leukocitų DNR metilinimo duomenys iš 3 skirtingų eksperimentų [LHA+14]. Pirmajame eksperimente tirtas šlapimo pūslės vėžys (angl. *Bladder cancer*), 223 mėginiai sudaro sergančiųjų grupę, 205 mėginiai – kontrolinę grupę. Antrajame eksperimente tirta galvos ir kaklo suragėjusių ląstelių karcinoma (angl. *Head and neck squamous cell carcinoma (HNSCC)*), 92 mėginiai sudaro sergančiųjų ir 92 – kontrolinę grupes. Paskutiniajame eksperimente tirtas kiaušidžių vėžys (angl. *Ovarian cancer*), 266 pacienčių mėginiai sudaro sergančiųjų, 274 – kontrolinę grupes. Visuose eksperimentuose kraujo mėginiai tirti mikrogardelių metodu naudojant *Illumina Infinium 27k Human DNA methylation Beadchip v1.2* gardeles, kurios padengia apie 27,000 CpG lokusų žmogaus genome. Iš viso 1152 mėginiai, 27578 požymių.

Tiriant mikrogardelių duomenis susiduriama su $p \gg N$ problema [HTF09], kai požymių daug daugiau nei mėginių. Kadangi požymių tiek daug, nėra paprasta atskirti, kurie jų tinkami klasifikatoriaus konstravimui. Tokių duomenų tyrimas – sudėtingas uždavinys.

Analizės metu susiduriama su daugiau problemų. Viena jų – kokybės problema, duomenyse gali būti techninis triukšmas, t. y. variabilumas, atsiradęs atliekant eksperimentą ir nesusijęs su biologiniu variabilumu, kurį norima tirti. Duomenyse pasitaiko sisteminės paklaidos (angl. *Batch effects*), duomenų poslinkis, atsiradęs dėl atlikto eksperimento niuansų, pavyzdžiui, paros meto, kada buvo atliekamas eksperimentas, analizėje naudotų priemonių ar įrangos. Duomenyse gali būti mėginių atsiskyrėlių (angl. *Outliers*) – mėginių, kurie labai

skiriasi nuo kitų mėginių, pavyzdžiui, nepavykus analizei. Neatsižvelgus į mėginius atsiskyrėlius, sukonstruotas klasifikatorius prastai veiktų. Juos galima aptikti naudojant, pavyzdžiui, principinių komponenčių analizę.

Tarp mėginių gali egzistuoti netiesinės priklausomybės, į kurias būtina atsižvelgti konstruojant klasifikavimo modelį. Tai reiškia, kad mėginių gali nepavykti atskirti naudojant tiesinį modelį, tuomet reikėtų naudoti netiesinį.

Tiriamų duomenų savybės nėra žinomos, todėl nėra aišku, kuris mašininio mokymosi metodas tiktų labiausiai, dėl to būtina išsiaiškinti ir išbandyti keletą jų. Egzistuoja daug ir įvairių mašininio mokymosi algoritmų. Skiriasi algoritmų veikimo greitis, sudėtingumas ir atvejai, kada jie naudojami, todėl būtina įsigilinti į jų veikimą ir taikymą.

Tiriant mikrogardelių metodu gautus duomenis, reikia atsižvelgti į technologijų, kuriomis jie gauti skirtumus, skirtingais metodais (skirtingo modelio mikrogardelėmis) gauti duomenys nėra palyginami.

Būtina atsižvelgti į tiriamą biologinį objektą, pavyzdžiui, kai tiriami cirkuliuojančios DNR metilinimo duomenys kyla klausimas dėl juose esančio signalo – ar įmanoma iš cirDNR metilinimo duomenų nustatyti pakitimus tarp sveikų ir sergančių pacientų mėginių, kadangi auglys, ypač pradinėse stadijose, būna nedidelis, ir jo žuvusių ląstelių DNR patekusios į kraują kiekis taip pat būna nedidelis ir susimaišęs kartu su žuvusių kraujo ir kitų ląstelių DNR [WS15]. Kai analizuojamas kraujyje esančių leukocitų DNR metilinimas taip pat nėra aišku, ar DNR metilinimas leukocituose pasikeičia susirgus vėžiu, ir jei atsiranda pokyčiai, ar jie būna pakankamai dideli, kad būtų pastebėti ir iš jų būtų galima atskirti, sergančių/sveikų pacientų mėginius. Ar pokyčiai priklauso nuo vėžio formos, kuria sergama, t. y. ar būtų galima atskirti sergantį pacientą ir pasakyti, kokiu vėžiu jis serga turint jo leukocitų DNR metilinimo duomenis.

Analizuojant mikrogardelių metodu gautus duomenis būtina atsižvelgti į visus prieš tai aptartus analizės niuansus. Tinkamai apdorojus pradinius duomenis, išbandžius pasirinktus mašininio mokymosi metodus, reikia išsiaiškinti, ar duomenys gali būti naudojami diagnozei nustatyti. Jei duomenys tinkami, gautas klasifikatorius turi patikimai priskirti naują pacientą prie sergančiųjų vėžiu arba sveikų pacientų grupės.

2 Darbo tikslas ir uždaviniai

Darbo tikslas – nustatyti ar leukocitų DNR ir cirDNR metilinimo duomenys tinkami vėžio diagnostikai. Jei tinkami, gauti klasifikatorių turimiems metilinimo duomenims, įvertinti jo tikslumą.

Siekiant įgyvendinti darbo tikslą, bus sprendžiami tokie uždaviniai:

- aptartus mašininio mokymosi metodus klasifikavimui: SVM, Random Forests, pritaikyti tiriant DNR metilinimo duomenis;
- panaudoti apžvelgtą principinių komponenčių analizės metodą duomenų požymių aibei sumažinti ir mėginiams atsiskyrėliams pašalinti;

- panaudoti kitus aptartus požymių atrinkimo metodus;
- palyginti mašininio mokymosi modeliais gautus rezultatus, nustatyti jų pranašumus ir trūkumus;
- suformuluoti pasiūlymus, kaip būtų galima pagerinti gautus rezultatus.

3 Laukiami rezultatai

Darbo rezultatai – suformuluota strategija leukocitų DNR ir cirDNR metilinimo duomenims tirti, nustatytas duomenų tinkamumas vėžio diagnostikai. Duomenyse esant signalui, rezultatas – klasifikatorius pacientų mėginiams klasifikuoti.

4 Literatūros apžvalga ir metodai

Skyriuje apžvelgiama mašininio mokymosi teorija, šiuo metu populiariausi klasifikavimo algoritmai. Aptariami naudoti duomenų apdorojimo ir analizės metodai.

4.1 Mašininio mokymosi taikymai

Mašininio mokymosi algoritmai taikomi daugelyje sričių, ne vien tiriant biologinės kilmės duomenis, ir kiekvienoje srityje jie tobulinami. Vienoje srityje patobulintas mašininio mokymosi algoritmas gali būti sėkmingai panaudotas kitoje srityje, todėl būtina apžvelgti pagrindines mašininio mokymosi taikymo sritis.

VLDB (angl. *Very Large Data Bases*), SIGIR, ICDM (angl. *International Conference on Data Mining*), KDD (angl. *Knowledge Discovery and Data Mining*), ECDA (angl. *European Conference on Data Analysis*) – tai su mašininiu mokymusi susijusios kasmet vykstančios pagrindinės konferencijos, kuriose pristatomos aktualios mašininio mokymosi taikymo sritys ir problemos, su kuriomis susiduriama taikant mašininio mokymosi modelius tose srityse.

2014 – 2015 m. populiariausios mašininio mokymosi taikymo sritys buvo pardavimai (angl. *Marketing*), finansai ir ekonomika, medicina ir gyvybės mokslai, socialiniai tinklai, rekomenduojančios sistemos (angl. *Recommendation Systems*), vaizdų, garso atpažinimo sistemos, robotika, literatūros analizė, paieškos sistemos ir kt. Naudojant mašininį mokymąsi duomenyse iš minėtų sričių ieškoma dėsningumų, pagal kuriuos būtų galima daryti tikslius spėjimus naujuose duomenyse.

Populiariausi mašininio mokymosi algoritmai, kurių patobulinimai ir modifikacijos, taikomi prieš tai minėtose srityse yra atsitiktiniai miškai (angl. *Random Forests*), atraminių vektorių mašinos (angl. *Support Vector Machines*), K artimiausių kaimynų metodas (angl. *K Nearest Neighbours*), principinių komponenčių analizė (angl. *Principal Components Analysis*), gilus mokymasis (angl. *Deep Learning*), neuroniniai tinklai (angl. *Neural Networks*)

ir kt. Atsitiktinių miškų, atraminių vektorių mašinų, principinių komponenčių analizės algoritmai plačiau apžvelgiami sekančiuose skyriuose.

Taikant mašininio mokymosi algoritmus susiduriama su įvairiomis problemomis, pavyzdžiui, kai duomenys nėra patikimi, juose gali pasitaikyti mėginių atsiskyrėlių (angl. *Outliers*). Pastarieji labai nutolę nuo kitų mėginių imtyje/populiacijoje, yra reti/nebūdingi atvejai. Jei tiriami biologinio eksperimento duomenys, tai mėginiai atsiskyrėliai gali atsirasti, pavyzdžiui, nepasisekus eksperimentui. Yra mašininio mokymosi algoritmų, pavyzdžiui, atraminių vektorių mašinos SVM, kurie jautrūs mėginiams atsiskyrėliams, todėl prieš taikant šį algoritmą reikia pašalinti tokius mėginius arba taikyti modifikuotą SVM [SOST14], kuris atsižvelgia į šią problemą.

Kita problema, iškylanti naudojant mašininį mokymąsi, kai žymėtuose dviejų klasių duomenyse viena klasė pasitaiko kur kas rečiau nei kita. Tai dažnas atvejis mediciniuose duomenyse, kai sergančiųjų tam tikra liga, kur kas mažiau nei sveikųjų [ANKA14].

Sudėtinga taikyti mašininio mokymosi algoritmus, kai turimų duomenų požymių p skaičius didelis (milijonai), o tiriamų mėginių N imtis maža ($p \gg N$ problema). Didelės dimensijos erdvėje, kai turima daug požymių, atstumai, pavyzdžiui, euklidiniai, tarp mėginių tampa dideli ir darosi nebeįmanoma apskaičiuoti panašumo tarp mėginių, jų atskirti. Tokia problema sprendžiama tiek mašinininiame mokymesi be mokytojo (angl. *Unsupervised Learning*), kai bandoma suklasterizuoti duomenis, turinčius daug požymių, ir kurių mėginių klasės nėra iš anksto žinomos [YZW+14], tiek mašinininiame mokymesi su mokytoju (angl. *Supervised Learning*). Be to dauguma mašininio mokymosi algoritmų ilgai veikia apdorodami duomenis su daug požymių.

Sekanti problema – skirtingas duomenų pasiskirstymas mokymo ir testavimo imtyse. Mašinininiame mokymesi visuomet daroma prielaida, kad testavimo ir mokymo duomenų pasiskirstymai vienodi, tačiau tarp jų gali egzistuoti skirtumas/poslinkis [WYG14], į kurį mašininio mokymosi algoritmas turėtų atsižvelgti.

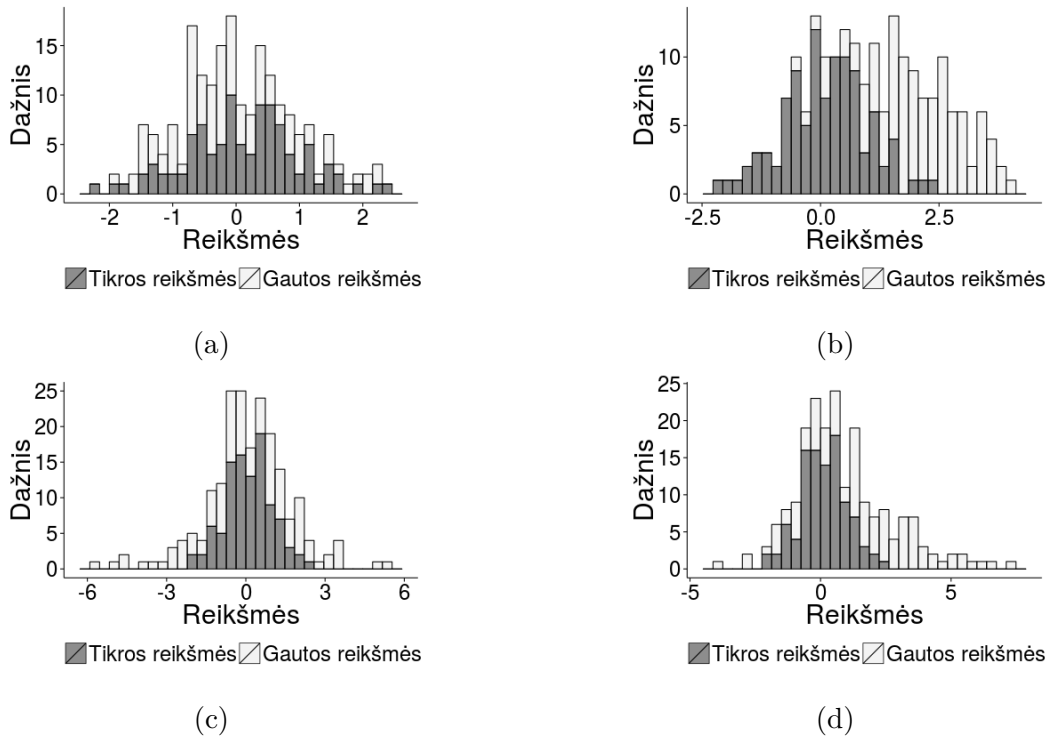
Mašininis mokymasis dažnai susiduria su problema, kai mėginio klasės nustatymas yra sudėtingas procesas, kuriame pasitaiko klaidų. Pasitaiko, kad dalis mokymo duomenų būna iš karto gerai priskirti tam tikrai klasei, o kita dalis turi netiksliai arba visai nenustatytą klasę. Tuomet taikomi dalinio mokymosi su mokytoju algoritmai (angl. *Partially supervised learning*) [ST14].

Mašininio mokymosi algoritmai dažnai būna sudėtingi, linkę persimokyti. Jei duomenyse tarp mėginių yra ne tiesinė priklausomybė, tuomet ne visi mašininio mokymosi algoritmai gerai klasifikuos tokius duomenis.

Renkantis klasifikatorių būtina atsižvelgti į prieš tai minėtas problemas, su kuriomis susiduria mašininio mokymosi algoritmai, ir tiriant duomenis išbandyti keletą jų.

4.2 Poslinkis (angl. *Bias*) ir variabilumas (angl. *Variance*)

Klasifikavimo modelio tikslumas priklauso nuo poslinkio (angl. *Bias*) ir variabilumo (angl. *Variance*) [GBD92], Pav. 1. Poslinkis yra skirtumas tarp gauto įverčio ir tikrosios



1 pav.: (a) Tikrosios ir klasifikatoriumi gautos reikšmės pasiskirsčiusios panašiai esant mažam klasifikatoriumi gautų reikšmių variabilumui ir poslinkiui. (b) Klasifikatoriumi gautų reikšmių variabilumas nedidelis, tačiau didelis poslinkis. (c) Klasifikatoriumi gautos reikšmės labai variabilios ir nepaslinktos. (d) Klasifikatoriumi gautos reikšmės labai variabilios ir paslinktos.

vertės. Klasifikatoriaus tikslumas yra didelis tuomet, kai mėginiai įvertinami taip, jog nėra nutolę nuo tikrosios reikšmės (angl. *Unbiased*) ir nėra labai variabilūs, Pav. 1 (a). Kai klasifikatoriumi gautos reikšmės koncentruojasi apie tikrąją reikšmę, tačiau yra paslinktos, Pav. 1 (b) – klasifikatoriaus tikslumas bus mažesnis. Tokiu atveju, jei žinomas poslinkis, galima atlikti įverčių korekcijas. Esant dideliam klasifikatoriumi gautų reikšmių variabilumui, Pav. 1 (c), nors poslinkio ir nėra, bet klasifikavimo tikslumas vis tiek nebus didelis. Jei klasifikatoriumi gautos reikšmės labai variabilios ir paslinktos, toks klasifikatorius naujus duomenis klasifikuos prastai, Pav. 1 (d).

Variabilumas – mėginių reikšmių išsibarstymas aplink tikrąją reikšmę. Dvi mėginių aibės gali turėti tokį patį vidurkį (juo įvertinama tikroji reikšmė, kai ji nežinoma), tačiau vienoje mėginių reikšmės gali būti labiau variabilios, nei kitoje. Aibėje, kurioje mėginių variabilumas mažesnis, galima tikėtis tikslesnio reikšmių įvertinimo.

Biologinių mėginių tyrimuose egzistuoja kelių tipų variabilumas: biologinis, kuris yra natūralus ir pasitaiko gamtoje, bei eksperimentinis/techninis, kuris egzistuoja dėl procedūros aspektų: mėginių paruošimo, dienos kada buvo vykdomas eksperimentas, naudotos įrangos ir kt. Visada norima pašalinti techninį variabilumą, kad būtų galima toliau tirti tik biologinį. Be to tyrimuose pasitaiko ir sisteminis poslinkis (angl. *Systematic experimental bias*), kuris taip pat atsiranda dėl eksperimento procedūros aspektų [WE05].

Klasifikuojant duomenis būtina atsižvelgti į abu šiuos parametrus: poslinkį ir variabilumą, nuo kurių priklauso kuriamo klasifikatoriaus klasifikavimo tikslumas.

Jei egzistuoja klasifikavimo modelis $Y = f(X) + \varepsilon$, kuris nėra žinomas, tai tokį modelį galima įvertinti $\hat{Y} = \hat{f}(X)$. Y žymi tikrąją reikšmę, t. y. vidutinę tikruoju klasifikatoriumi, kuris žymimas $f(X)$, gautą reikšmę. X žymi mokymo duomenų aibę. ε raide žymimas triukšmas (angl. *Irreducible Error*), kurio vidurkis – $E(\varepsilon) = 0$ ir variabilumas $Var(\varepsilon) = \sigma_\varepsilon^2$ ($Var(\varepsilon)$ gali būti lygus 0, bet nebūtinai). Triukšmo negalima pašalinti, nes jis nėra žinomas ir atsiranda dėl skaičiavimo netikslumų. Turint didelę aibę mokymo duomenų ir pakartotinai įvertinus tikrąjį klasifikatorių f klasifikatoriumi $\hat{f}$ bei testuojant taške $X = x_0$, galima būtų įvertinti kuriamo klasifikatoriaus $\hat{f}$ vidutinę kvadratinę testavimo klaidą (angl. *Mean squere error*) [JWHT13], 34 psl.:

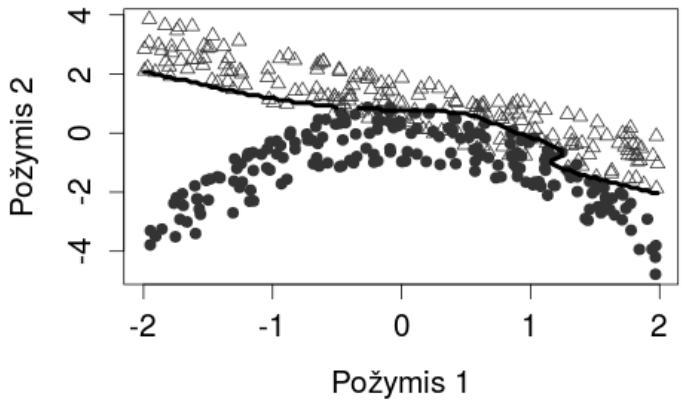
$$Err(x_0) = E[(f(x_0) - \hat{f}(x_0))]^2 = \tag{1}$$

$$\sigma_\varepsilon^2 + [f(x_0) - E\hat{f}(x_0)]^2 + E[\hat{f}(x_0) - E\hat{f}(x_0)]^2 \tag{2}$$

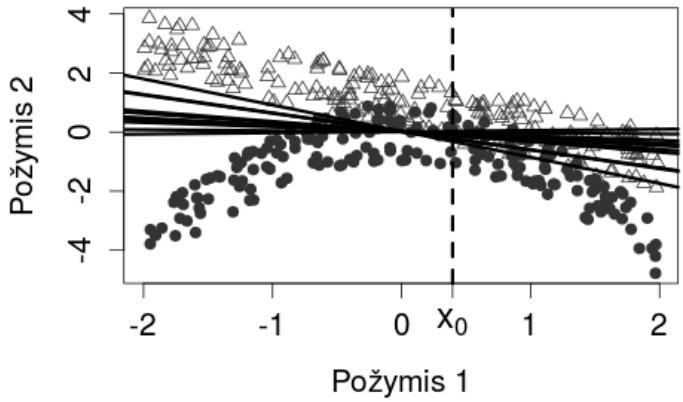
Vidutinę kvadratinę testavimo klaidą sudaro triukšmas, skirtumo taške x_0 tarp tikrosios reikšmės ir vidutinės $\hat{f}$ klasifikatoriumi gautos reikšmės kvadratas ir $\hat{f}$ klasifikatoriumi gautų reikšmių variabilumas taške x_0. Idealiu atveju poslinkis ir variabilumas turėtų būti lygūs 0. Tačiau dažniausiai taip nebūna ir kuriant $\hat{f}$ klasifikatorių siekiama sumažinti skirtumą tarp jį naudojant gautos vidutinės reikšmės atsitiktinai pasirinktame taške ir tikrosios reikšmės. Siekiama sumažinti naujai sukurto klasifikatoriaus variabilumą atsitiktinai pasirinktame taške. Dažnai sumažinę klasifikatoriaus variabilumą, padidiname poslinkį. Šis atvejis pavaizduotas Pav. 2 (b). Taškai žymi dviejų klasių mėginių pasiskirstymą. $f(x)$, tikrasis klasifikatorius, pažymėtas juoda kreive Pav. 2 (a). Tiesės žymi $\hat{f}(x)$ klasifikatorių Pav. 2 (b). Atsitiktinai pasirinktame taške x_0 matomas $\hat{f}$ klasifikatoriumi gautų reikšmių tame taške pasiskirstymas. Ryškus skirtumas tarp tikruoju klasifikatoriumi gautų reikšmių ir nauju klasifikatoriumi gautų reikšmių taške x_0 reiškia didelį poslinkį. $\hat{f}$ klasifikatoriumi gautų reikšmių taške x_0 pasiskirstymas nėra didelis – jos koncentruojasi apie vidurkį, tai reiškia, kad klasifikatoriaus variabilumas nėra didelis.

Jeigu sumažinamas poslinkis – variabilumas padidėja. Pav. 2 (c) pavaizduoti taškai žymi dviejų klasių mėginių pasiskirstymą. $f(x)$ pažymėtas juoda kreive Pav. 2 (a). $\hat{f}(x)$ pažymėtas kreivėmis artimomis tikrajai $f(x)$ kreivei Pav. 2 (c). Atsitiktinai pasirinktame taške x_0 matomas didelis klasifikatoriumi gautų reikšmių tame taške pasiskirstymas. Skirtumas tarp tikruoju klasifikatoriumi gautų reikšmių ir $\hat{f}$ klasifikatoriumi gautų reikšmių taške x_0 nedidelis, tai reiškia mažą poslinkį. $\hat{f}$ klasifikatoriumi gautų reikšmių taške x_0 pasiskirstymas didelis – tai reiškia didelį klasifikatoriaus variabilumą.

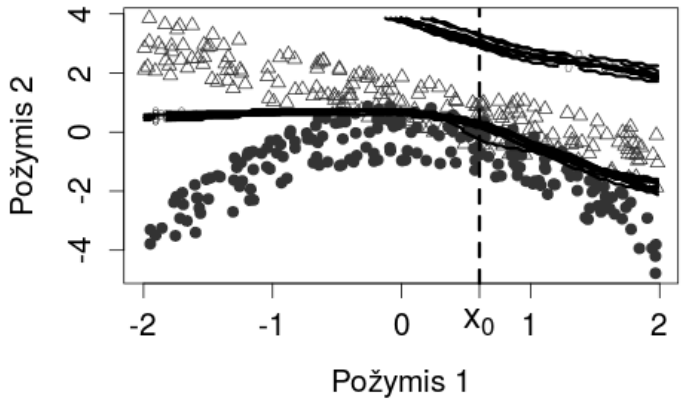
Modelio $\hat{f}$ variabilumas priklauso nuo jo sudėtingumo. Kuo kuriamas klasifikatorius sudėtingesnis, t. y. turi daugiau parametrų, tuo jo variabilumas didesnis. Taip yra, nes sudėtingesnis modelis labiau prisitaiko prie mokymo duomenų – persimoko (angl. *Overfitting*). Ir atvirkščiai, kuo klasifikatorius paprastesnis, tuo jo variabilumas mažesnis. Paprastesnis klasifikatorius priešingai, negu sudėtingas, nepersimoko mokymo duomenimis, greičiau nepakankamai išmoksta (angl. *Underfitting*). Kai klasifikatorius nesudėtingas, jo variabilumas



(a)



(b)



(c)

2 pav.: (a) „Tikrasis" klasifikatorius sukonstruotas naudojant visą populiaciją. (b) Esant nedideliam klasifikatoriaus variabilumui, atsiranda poslinkis. (c) Esant dideliam klasifikatoriaus $\hat{f}$ variabilumui, poslinkis tarp $\hat{f}$ ir f reikšmių atsitiktiniame taške tampa nedidelis.

yra nedidelis, tačiau atsiranda didelis klasifikatoriaus generuojamų reikšmių poslinkis. Dėl to klasifikatorius prastai apsimoko ir testuoja. Kai klasifikatorius labai sudėtingas, su daug parametrų, jis persimoko ir jo generuojamų reikšmių atsitiktiniame taške variabilumas didelis. Tada poslinkis nuo tikrosios reikšmės tampa nedideliu, tačiau esant dideliam reikšmių variabilumui sunkiau „atspėti", kuri yra tikroji. Todėl, norint sukurti gerą klasifikatorių, būtina atsižvelgti į kuriamo klasifikatoriaus sudėtingumą, variabilumą ir jo generuojamų reikšmių galimą poslinkį.

4.3 Klasifikavimo metodai

4.3.1 Maksimalaus tarpo klasifikatorius

Atraminių vektorių mašina (SVM) [Vap95] paprasto ir intuityvaus maksimalaus tarpo klasifikatoriaus (angl. *Maximal margin classifier*) [BGV92] apibendrinta forma. Vladimiras Vapnikas 1970 m. pasiūlė pradinius SVM variantus. SVM pradėta labiau vystyti nuo 1990 m.

Maksimalaus tarpo klasifikatorius atskiria duomenis hiperplokštuma – padalina į dvi klases Pav. 3 (a).

Apibrėžiama pradinių duomenų aibė: n žymi mėginių aibę nuo 1 iki n, o p – požymių aibę. *mėginių* $\times$ *požymių* matrica:

$$X_{n,p} = \begin{pmatrix} x_{1,1} & x_{1,2} & \cdots & x_{1,p} \\ x_{2,1} & x_{2,2} & \cdots & x_{2,p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n,1} & x_{n,2} & \cdots & x_{n,p} \end{pmatrix} \tag{3}$$

Atsitiktinis mėginys žymimas x_i, kur i = 1, 2, ..., n:

$$x_\mathrm{i} = x_{\mathrm{i},1}, x_{\mathrm{i},2}, \ldots, x_{\mathrm{i,p}} \tag{4}$$

Kiekvienas mėginys priklauso tam tikrai klasei (sergančių/nesergančių). Mėginių klasių vektorius žymimas y, kur $\mathrm{y_i}$ atitinka *i*-tojo mėginio klasę, i = 1, 2, ..., n.

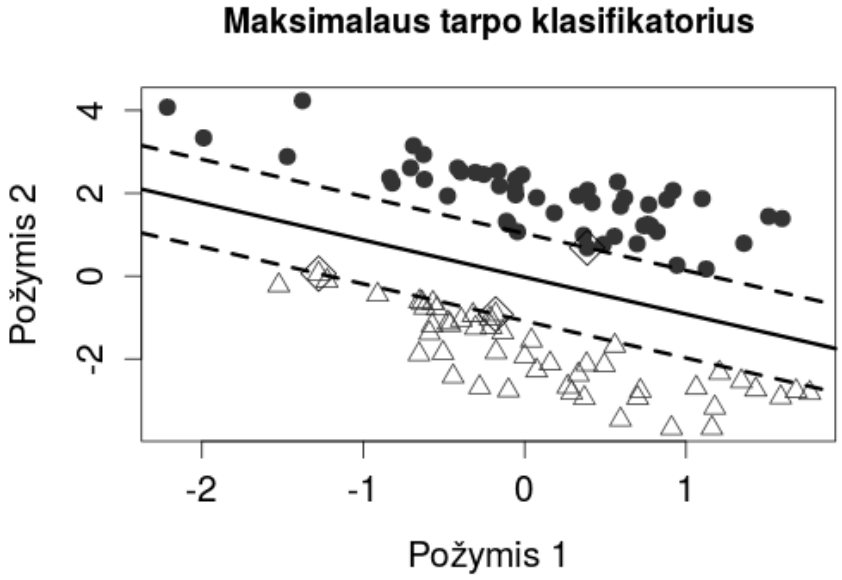
Tada apibrėžiamas maksimalaus tarpo klasifikatorius. Pirmiausia apibrėžiami jo hiperplokštumos koeficientai, kurie žymimi β. Kai požymių skaičius yra p, tada:

$$\beta = \beta_1, \beta_2, \ldots, \beta_\mathrm{p} \tag{5}$$

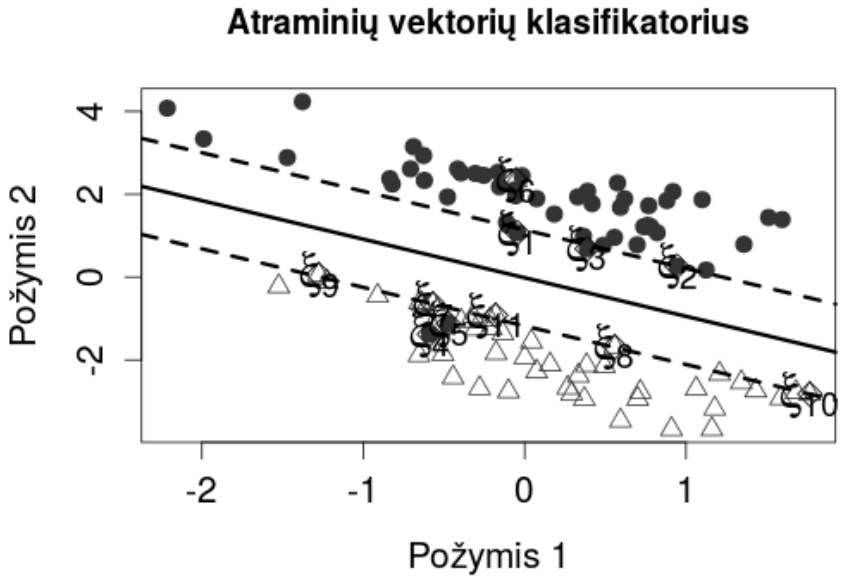
$$\beta_0{}^2 + \beta_1{}^2 + \beta_2{}^2 + \ldots + \beta_\mathrm{p}{}^2 = 1 \tag{6}$$

Kai mėginio $\mathrm{x_i}$ klasė $\mathrm{y_i} = 1$, tuomet:

$$\beta_0 + \beta_1 x_{\mathrm{i},1} + \beta_2 x_{\mathrm{i},2} + \ldots + \beta_\mathrm{p} x_{\mathrm{i,p}} = \beta_0 + \beta^\mathrm{T} x_\mathrm{i} > 0 \tag{7}$$



(a)



(b)

3 pav.: (a) Maksimalaus tarpo klasifikatorius. Ryškiai juoda linija žymi atskiriančią hiperplokštumą. Tarp hiperplokštumos ir punktyrinių linijų esanti erdvė yra tarpas (angl. *Margin*). Mėginiai esantys ant punktyrinių linijų – atraminiai vektoriai, tai arčiausiai hiperplokštumos esantys taškai pagal kuriuos ji nustatoma. (b) Atraminių vektorių klasifikatorius. ξ_i žymi atraminius vektorius. Kintamieji $\xi_2, \xi_7, \xi_8, \xi_9, \xi_{10}$ yra lygūs 0. $\xi_1, \xi_3, \xi_{11} > 0$ tarpe atsidūrusiems mėginiams. $\xi_4, \xi_5, \xi_6 > 1$, kadangi mėginiai atsidūrė ne toje hiperplokštumos pusėje.

Kai mėginio x_i klasė y_i = -1, tuomet:

$$\beta_0 + \beta_1 x_{i,1} + \beta_2 x_{i,2} + \ldots + \beta_p x_{i,p} = \beta_0 + \beta^T x_i < 0 \tag{8}$$

Bendru atveju mėginio priskyrimas apibrėžiamas:

$$f(x_i) = \beta_0 + y_i \beta^T x_i > 0 \tag{9}$$

Turint mėginį x_i, pagal $f(x_i)$ ženklą mėginys priskiriamas tam tikrai klasei, o pagal $f(x_i)$ dydį galima pasakyti, ar mėginys toli nuo hiperplokštumos. Jei mėginys yra toliau nuo hiperplokštumos, galima būti tikresniems, kad mėginys teisingai priskirtas tam tikrai klasei.

Tiesiškai atskiriamą dviejų klasių mėginių aibę bandant atskirti hiperplokštuma, egzistuoja begalinis skaičius tokių hiperplokštumų, nes paslinkus ar pasukus senąją hiperplokštumą galima gauti naują atskiriančią hiperplokštumą. Maksimalaus tarpo klasifikatoriaus atveju pasirenkama hiperplokštuma esanti toliausiai nuo mokymo mėginių [JWHT13], 341 psl. Mėginius atskiriant maksimalaus tarpo hiperplokštuma laikoma, kad tokia hiperplokštuma turėtų atskirti ir testinių mėginių aibę. Maksimalaus tarpo klasifikatorius dažniausiai sėkmingai atskiria testinius mėginius, tačiau esant dideliam požymių p skaičiui gali persimokyti.

Norint rasti maksimalaus tarpo klasifikatoriaus hiperplokštumą pirmiausia apibrėžiama [Fle09]:

$$x_i \cdot \beta + \beta_0 \geq +1, \text{ kai } y_i = +1 \tag{10}$$

$$x_i \cdot \beta + \beta_0 \leq -1, \text{ kai } y_i = -1 \tag{11}$$

$$\tag{12}$$

Tuomet bendru atveju gaunama:

$$y_i(x_i \cdot \beta + \beta_0) - 1 \geq 0, \forall i, i = 1, ..., n \tag{13}$$

Tarpas tarp hiperplokštumos ir atraminių vektorių žymimas:

$$M = \frac{1}{||\beta||} \tag{14}$$

Norint maksimizuoti M, sprendžiama $||\beta||$ minimizavimo problema naudojant Lagranžo daugiklių metodą:

$$min\ \frac{1}{2}||\beta||^2 \text{ su sąlyga, kad } y_i(x_i \cdot \beta + \beta_0) - 1 \geq 0, \forall i, i = 1, ..., n \tag{15}$$

Naudojant Lagranžo daugiklių metodą gaunama:

$$L_P = \frac{1}{2}||\beta||^2 - \sum_{i=1}^{n} \alpha_i y_i (x_i \cdot \beta + \beta_0) + \sum_{i=1}^{n} \alpha_i \tag{16}$$

Norima minimizuoti β ir β_0 bei maksimizuoti α. Tada ieškoma sprendinių diferencijuojant L_P pagal β ir β_0. Gaunami sprendiniai:

$$\beta = \sum_{i=1}^{n} \alpha_i y_i x_i \tag{17}$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0 \tag{18}$$

Juos įstačius į (16) gaunama:

$$L_D = \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i,j}^{n} \alpha_i \alpha_j y_i y_j x_i x_j, \text{ kur } \alpha_i \geq 0, \forall i, i = 1, ..., n \tag{19}$$

Gavus L_P dualią formą (angl. *Dual form*) L_D ji maksimizuojama. Šiai problemai spręsti naudojamas kvadratinis programavimas (angl. *Quadratic Programming* (QP)). Gavus α, randama β ir apskaičiuojamas β_0.

Atraminių vektorių klasifikatoriui ir atraminių vektorių mašinoms (SVM) hiperplokštumos ieškoma sprendžiant tokią pat problemą naudojant Lagranžo daugiklių metodą, tik įvedamos papildomos sąlygos (kadangi atraminių vektorių klasifikatorius ir SVM leidžia daliai mėginių atsidurti tarpe M arba kitoje hiperplokštumos pusėje).

4.3.2 Atraminių vektorių klasifikatorius

Kai nėra egzistuojančios hiperplokštumos su tarpu tarp hiperplokštumos ir mokymo mėginių M > 0, galinčios atskirti mėginius į dvi skirtingas klases, naudojamas atraminių vektorių klasifikatorius. Jis yra maksimalaus tarpo klasifikatoriaus variantas, kuriuo klasifikuojant dalis atraminių vektorių gali būti mėginiai priskirti neteisingai klasei.

Naudojant atraminių vektorių klasifikatorių, hiperplokštuma konstruojama panašiai, kaip ir maksimalaus tarpo klasifikatoriaus atveju. Tačiau naudojami laisvi kintamieji ξ_i, kur i = 1, 2, ... n. Jų aibė žymima ξ:

$$\xi = \xi_1, \xi_2, \ldots, \xi_\mathrm{n} \tag{20}$$

$$\xi_1 + \xi_2 + \ldots + \xi_\mathrm{n} <= C \tag{21}$$

Laisvas kintamasis ξ_i parodo, kaip gerai klasifikuojamas mėginys x_i. Jei mėginiui x_i $\xi_\mathrm{i} = 0$, tai reiškia, kad mėginys priskirtas gerai klasei. Kai $\xi_\mathrm{i} > 0$, tai reiškia, kad mėginys atsidūrė tarpe M. Kai $\xi_\mathrm{i} > 1$, tada mėginys buvo blogai suklasifikuotas – atsidūrė ne toje hiperplokštumos pusėje Pav. 3 (b).

C raide žymimas pasirenkamas skaičius, konstanta, apibrėžiantis tolerantiškumą klai-

dingam klasifikavimui. Kai C parinktas didelis, leidžiama naudoti daugiau atraminių vektorių ir tarpas M būna platesnis. C parinkus mažą – naudojama mažiau atraminių vektorių. Tada gaunamas mažesnis reikšmių poslinkis ir didesnis variabilumas.

Atraminių vektorių klasifikatoriaus hiperplokštuma apibrėžiama:

$$\beta_0 + y_\mathrm{i}\beta^\mathrm{T}x_\mathrm{i} >= M(1 - \xi_\mathrm{i}) \tag{22}$$

4.3.3 Atraminių vektorių mašina – SVM (angl. *Support Vector Machine*)

Atraminių vektorių klasifikatorius tinka tiesiškai atskiriamiems mėginiams. Kai mėginiai gali būti atskiriami netiesiškai, naudojama atraminių vektorių mašina – atraminių vektorių klasifikatorius, kuriame panaudojami branduoliai (angl. *Kernels*) [ABR64]. Branduolys tarp skirtingų mėginių x_i ir x_j žymimas:

$$K(x_i, x_j) \tag{23}$$

Atraminių vektorių klasifikatoriuje tarp mėginių skaičiuojama jų skaliarinė sandauga (angl. *Inner product*):

$$< x_i, x_j >= x_{\mathrm{i,1}} \times x_{\mathrm{j,1}} + x_{\mathrm{i,2}} \times x_{\mathrm{j,2}} + ... + x_{\mathrm{i,p}} \times x_{\mathrm{j,p}} = \sum_{k=1}^{p} x_{\mathrm{i,k}} x_{\mathrm{j,k}} \tag{24}$$

Tuomet tiesinis atraminių vektorių klasifikatorius aprašomas lygtimi:

$$f(x) = \beta_0 + \sum_{i \in S}^{n} (\alpha_\mathrm{i} < x, x_i >), \text{ kur } S \text{ yra atraminių vektorių aibė} \tag{25}$$

x yra naujas mėginys, $\mathrm{x_i}$ – mėginys iš mokymo aibės. α_i – parametras i-tajam mėginiui. Norint gauti $\alpha_1, \ldots, \alpha_\mathrm{n}$ ir β_0 skaičiuojamos skaliarinės sandaugos $< x_i, x_j >$ tarp $n(n-1)/2$ mėginių porų. α nelygus 0 tik atraminiams vektoriams. Vietoje skaliarinės sandaugos skaičiuojant polinominį arba radialinį (angl. *Radial kernel*) branduolį, kuris įvertina panašumą tarp dviejų mėginių, atraminių vektorių klasifikatorius bus vadinamas atraminių vektorių mašina [JWHT13]:

$$f(x) = \beta_0 + \sum_{i \in S}^{n} (\alpha_\mathrm{i} K(x, x_i)), \text{ kur } S \text{ yra atraminių vektorių aibė} \tag{26}$$

Polinominis branduolys (d – polinomo dydis. d > 1) Pav. 4 (a):

$$K(x_i, x_j) = (1 + \sum_{k=1}^{p} x_{\mathrm{i,k}} x_{\mathrm{j,k}})^\mathrm{d} \tag{27}$$

Radialinis branduolys (γ – konstanta, teigiamas skaičius) Pav. 4 (b):

$$K(x_i, x_j) = exp(-\gamma \sum_{k=1}^{p} (x_{\mathrm{i,k}} - x_{\mathrm{j,k}})^2) \tag{28}$$

4.3.4 RGLM (angl. *Random generalized linear model*) metodas

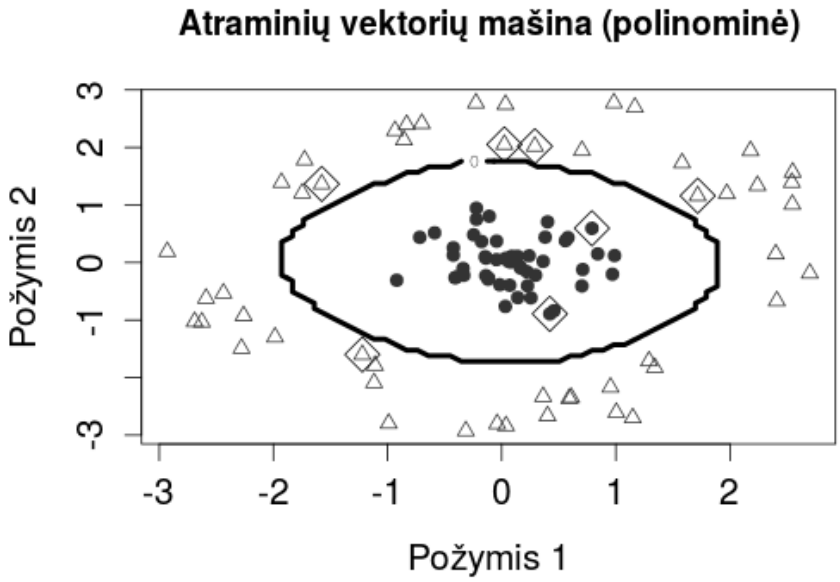
RGLM [SLH13] paremtas savirankos (angl. *Bootstrap aggregation* arba *Bagging*) metodo panaudojimu apibendrintuose tiesiniuose modeliuose. GLM [NW72] apima didžiulę klasę regresijos modelių, pavyzdžiui, tiesinės regresijos modelį pagal normalųjį skirstinį pasiskirsčiusiems duomenims, logistinės regresijos – binariniams duomenims ir kt. RGLM gali būti naudojamas nustatyti binarinius, tęstinius ir kitokius priklausomus kintamuosius, kuriems gali būti apibrėžtas apibendrintas tiesinis modelis. Atsitiktinumą RGLM lemia naudojama neparametrinė savirankos (angl. *Bootstrap*) procedūra, kuri sugeneruoja tokio pat dydžio duomenų aibes kaip originali duomenų aibė, tačiau sugeneruotose duomenų aibėse naudojami keitiniai, t. y. kai atsitiktinai pasirenkami mėginiai, kurie pakeičiami kitais mėginiais iš originalios duomenų aibės. Dėl to kiekvienoje sugeneruotoje duomenų aibėje kai kurie mėginiai kartojasi. Atsitiktinai pasirenkami ir požymių poaibiai.

Vykdant RGLM pirmiausia sukonstruojamos duomenų aibės su keitiniais atitinkamo dydžio kaip tikroji duomenų aibė naudojant neparametrinį savirankos metodą. Tada kiekvienoje aibėje pasirenkamas atsitiktinis požymių poaibis. Atsitiktinai parinkti požymiai kiekvienoje aibėje išrikiuojami pagal kiekvieno ryšį su priklausomu kintamuoju. Kai spėjamas kiekybinis priklausomas kintamasis, požymių svarbai nustatyti gali būti naudojamas koreliacijos koeficientas tarp priklausomojo kintamojo ir kiekvieno požymio. Pasirenkami reikšmingiausi požymiai, kurie toliau bus naudojami daugiamatės regresijos modelyje. Tada kiekvieno daugiamačio modelio spėjimai kiekvienai aibei (angl. *Bag*) apjungiami į galutinį spėjimą, kuris lemia klasės pasirinkimą. Spėjimų apjungimo metodas pasirenkamas priklausomai nuo priklausomojo kintamojo tipo. Tęstiniams duomenims išvedamas gautų spėjimų vidurkis, binariniams taikoma daugumos taisyklė (kuri klasė spėjama dažniausiai), kartais būna naudojamas slenkstis [SLH13].

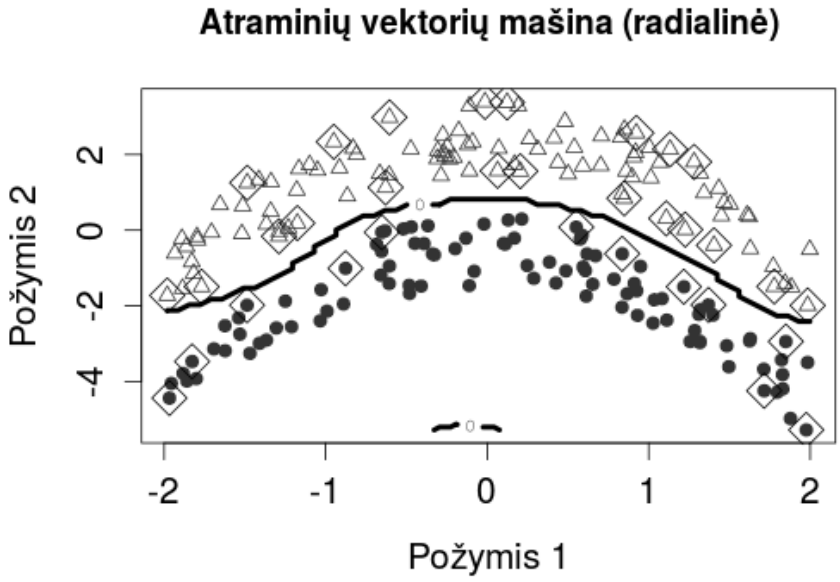
4.3.5 Atsitiktinių miškų (angl. Random Forests) metodas

Random Forests [Bre01] [Ho95] [Ho98] [AG97] metodas naudojamas duomenims klasifikuoti. Algoritmo vykdymo metu pradinė mėginių aibė M atsitiktinai padalinama į S poaibių, iš kurių konstruojami pilni ID3 [Qui86] sprendimų medžiai. Gaunama S skirtingų sprendimų medžių. Testuojamas mėginys klasifikuojamas su visais gautais medžiais. Kiekvienas medis jį priskiria tam tikrai klasei. Apskaičiuojami gautų klasių pasikartojimo dažniai ir mėginys priskiriamas tai klasei, kurios pasikartojimo dažnis buvo didžiausias [JWHT13].

Yra keletas sprendimų medžių algoritmų. *Random Forests* dažnai naudoja ID3 (Algoritmas 1).



(a)



(b)

4 pav.: Mėginių klasifikavimas naudojant branduolius. (a) Mėginiams atskirti naudojamas 4 eilės polinominis branduolys (d = 4). (b) Mėginiai atskiriami naudojant radialinį branduolį.

Algoritmas 1 ID3

```
1: function DALINTI (VIRŠŪNĖ, {MĖGINIŲ AIBĖ})
2:     A ← geriausiai {mėginių aibę} dalinantis požymis (pagal informacijos kiekį)
3:     Sukuriama viršūnė ← A požymiui
4:     Kiekvienai A reikšmei sukuriamos naujos vaikinės viršūnės
5:     Mokymo {mėginių aibė} išdalinama kiekvienai vaikinei viršūnei
6:     for Kiekvienai vaikinei viršūnei – poaibiui do:
7:         if Poaibis „grynas“, t. y. priklauso tik vienai klasei then stop;
8:         else goto Dalinti (viršūnė, {mėginių aibė}).
9:     end
```

Geriausias požymis mėginių aibės dalinimui pasirenkamas skaičiuojant kiekvieno požymio informacijos kiekį (angl. *Information gain*). Informacijos kiekis randamas pagal formulę:

$$Gain(M, A) = H(M) - sum((|S|/M) \times H(S)) \tag{29}$$

Formulėje M yra mėginių aibė, A – požymis, S − mėginių aibės poaibis, H(M) – entropija skaičiuojama:

$$H(M) = -p(+) \times log2(p(+)) - p(-) \times log2(p(-)) \tag{30}$$

Formulėje p(-) yra tikimybė, kad mėginys priklauso tam tikrai klasei.

Informacijos kiekis tarp požymio A ir klasių iš aibės M – tai įvertis reiškiantis, kaip gerai pasirinktas požymis dalina mėginių aibę į skirtingas klases. Pasirenkamas požymis su didžiausiu informacijos kiekio įverčiu [Lav14].

Naudojant sprendimų medžius susiduriama su persimokymo problema. Medis gali labai išaugti. Persimokymo problema sprendžiama naudojant reikšmingumo testus arba apkarpant medį. Medžio apkarpymas vyksta mokymo aibę atsitiktinai padalinus į dvi – skirtą mokymui ir validavimui. Atlikus medžio apmokymą jis testuojamas su validavimo aibe: pašalinus vieną iš jo šakų žiūrima, ar jis vis dar gerai klasifikuoja duomenis. Procesas kartojamas, kol šakos pašalinimas tampa žalingas.

Sprendimų medžių pliusai, kad jie lengvai interpretuojami, gerai susitvarko su triukšmu (jo informacijos kiekis dažnai būna 0), su trūkstamais duomenimis (angl. *Missing data*), yra kompaktiški ir gana greitai veikia. Sprendimų medžių minusas, kad tolydiems duomenims jie naudoja slenksčius, kurie duomenims atskirti „brėžia“ tik horizontalias ir vertikalias linijas dalindami erdvę į zonas. Tai, kai kuriais atvejais, nėra efektyvu, tuomet geriau būtų naudoti tiesinės regresijos modelį, kuris „brėžtų” įstrižą liniją duomenims atskirti ar kitą metodą [Lav14].

4.4 Dimensionalumo mažinimo metodai

4.4.1 Principinių komponenčių analizė

Principinių komponenčių analizė (angl. *Principal components analysis*) [Pea01] [Hot33] – PCA – tai tiesinės duomenų transformacijos metodas, naudojamas duomenų dimensionalumui sumažinti.

Atliekant PCA analizę pirmiausia ieškoma krypties, kuria duomenų variabilumas yra didžiausias. Ta kryptimi „brėžiama" pirmoji principinė komponentė. Tada ieškoma sekančios krypties, kurioje duomenų variabilumas didžiausias ir „brėžiama" antroji principinė komponentė, kuri būtų statmena pirmajai. Jei požymių aibė yra p dydžio, tai iš viso gaunama p principinių komponenčių. Tačiau dažniausiai pasirenkamos pirmosios, tos, kuriose duomenų variabilumas didžiausias.

Pradinių duomenų aibė apibrėžiama tokia pat kaip ir SVM atveju ((3) lygtis), kur duomenų matrica susideda iš n mėginių ir p požymių. Laikoma, kad duomenys centruoti. Tada norint gauti principines komponentes pirmiausia apskaičiuojama kovariacijų matrica. Ji žymima C. Apibrėžiama p tikrinių vektorių e ir tikrinių reikšmių λ. Tikrinės reikšmės apibūdina duomenų variabilumą tikrinio vektoriaus kryptimi. Pirmiausia randamos tikrinės reikšmės, lygtis (31), kur I – vienetinė matrica. Tuomet apskaičiuojami tikriniai vektoriai, lygtis (32). Tikriniai vektoriai surūšiuojami juos atitinkančių tikrinių reikšmių mažėjimo tvarka. Gaunama principinių komponenčių matrica A, kurios stulpelius sudaro išrikiuoti tikriniai vektoriai. Tada pasirinktas i-tasis mėginys, kai i = 1, ..., n, transformuojamas pagal (33) lygtį. Pasirinkus mažiau tikrinių vektorių k, kur $k < p$, principinių komponenčių matrica A sudaroma tik iš k tikrinių vektorių ir duomenys transformuojami į mažesnės dimensijos erdvę [JWHT13].

$$det(C - \lambda I) = 0 \tag{31}$$

$$Ce = \lambda e \tag{32}$$

$$y_i = x_i A \tag{33}$$

4.4.2 Požymių atrinkimo metodai

Duomenų dimensionalumas mažinamas vykdant požymių atranką. Požymių atrinkimui taikomi filtravimo metodai (angl. *Filter methods*), kai požymiai rikiuojami ir pasirenkami svarbiausi – turintys aukščiausią rangą/poziciją. Pavyzdžiui, atsirenkami variabiliausi požymiai, kurie tikimasi turi daugiausia biologinio variabilumo, kurį norima tirti. Požymiai dar gali būti pasirenkami skaičiuojant jų informacijos kiekį ar kt.

Požymių atrinkimui dar naudojami atrinkimo kartu su klasifikatoriumi metodai (angl. *Wrapper methods*). Pavyzdžiui, rekursyvus požymių šalinimas (plačiau aptariamas sekančiame skyriuje), kurį naudojant pasirenkami požymių poaibiai ir žiūrima, su kuriuo jų pavykdavo sukonstruoti geriausią klasifikatorių.

4.4.3 Rekursyvus požymių šalinimas (angl. *Recursive Feature Elimination* – RFE)

RFE [AG97] – rekursyvaus požymių šalinimo kartu su pasirinktu pakartotinio atrinkimo (angl. *Resampling*) metodu algoritmas (Algoritmas 2) naudojamas požymių atrinkimui didelėje požymių aibėje.

Algoritmo vykdymo metu naudojant pakartotinį atrinkimą pradinė mėginių aibė dalinama į mokymui ir testavimui skirtas aibes. Tada pasirinktas modelis (*Random Forests*, SVM ar kt.) apmokomas su visais mokymo aibės požymiais. Gautas modelis testuojamas su testine aibe. Apskaičiuojama kiekvieno požymio svarba/rangas. Toliau apmokymas vyksta naudojant požymių poaibius. Imami, pavyzdžiui, svarbiausi 50, 100 ar daugiau požymių (poaibių dydžiai pasirenkami), su jais apmokomas modelis, kuris vėliau testuojamas su testine aibe. Procesas kartojamas kiekvienam požymių poaibiui kiekvienos pakartotinio atrinkimo iteracijos metu. Kiekvienam poaibiui apskaičiuojama charakteristika, t. y. kaip gerai jam pavykdavo apmokyti klasifikatorių ir su juo klasifikuoti testinę aibę. Pasiliekamas geriausias poaibis, kuris naudojamas galutiniam klasifikavimo modeliui konstruoti [Kuh14].

Algoritmas 2 Rekursyvus požymių šalinimas naudojant pakartotinį atrinkimą

for kiekvienai pakartotinio atrinkimo iteracijai **do**:
    Duomenų aibė dalinama į mokymo ir testinę naudojant pakartotinį atrinkimą
    Modelis apmokomas su mokymo aibe
    Modelis testuojamas su testine aibe
    Skaičiuojama požymių svarba/rangas
    for kiekvienam poaibiui dydžio S_i, i = 1 ... S **do**
        Įsimenami svarbiausi S_i požymiai
        [Nebūtinai] išankstinis duomenų apdorojimas (angl. *Pre-process*)
        Apmokomas modelis naudojant S_i dydžio požymių aibę iš mokymo aibės
        Modelis testuojamas su testine aibe
        [Nebūtinai] Perskaičiuojama svarba/rangas kiekvienam požymiui
    end
end
Apskaičiuojama vykdymo charakteristika kiekvienam S_i dydžio poaibiui naudojant testinę aibę
Nustatomas geriausias požymių aibės dydis
Įvertinamas galutinis požymių sąrašas, kuris bus naudojamas modeliui sudaryti
Galutinis modelis sudaromas naudojant optimalią S_i požymių aibę mokymo aibėje

4.5 Duomenų apdorojimo metodai

4.5.1 Genominių koordinačių perkėlimas

Nustatoma tiksli požymių (angl. *probe*. Trumpa ∼25 bp (bazių porų) viengrandės cirDNR seka) vieta žmogaus genome hg19, t. y. požymių koordinatės perkeliamos (angl. *Lift Over*) iš hg18 į naujausiai anotuotą hg19.

4.5.2 *NA* verčių skaičiavimas

Jei duomenyse yra tuščių *NA* verčių jos apskaičiuojamos kiekvienam požymiui atskirai. Vykdomas ciklas, kurio metu kiekvienam požymui naudojant netuščius DNR metilinimo įverčius apskaičiuojama mediana, kuri įrašoma vietoje tuščių *NA* verčių.

4.5.3 Duomenų transformacija

Mikrogardelių metodu gautų DNR metilinimo duomenų pasiskirstymas dažnai būna nesimetrinis. Tuomet atliekama duomenų log_2 transformacija, po kurios metilinimo įverčių pasiskirstymas įgauna „varpo" formą. Tada galima taikyti pagrindinių charakteristikų, tokių kaip vidurkis skaičiavimą, nes jis nebebūna paslinktas. Esant didesniam vidurkiui paprastai išauga variabilumas, taigi tose grupėse (šiuo atveju sergančiųjų ir sveikųjų), kuriose didesnis vidurkis, padidėja ir variabilumas. log_2 transformacija pataiso šią vidurkio ir variabilumo priklausomybę, dėl to padidėja statistinių testų galia [GT09] ir atsiranda galimybė aptikti esantį tikrąjį skirtumą tarp grupių, t. y. kad kurio nors požymio ekspresija ryškiai padidėjo/sumažėjo lyginant abi grupes.

4.5.4 Normalizavimas (angl. *Quantile normalization*)

Duomenys normalizuojami naudojant kvantilių normalizavimo metodą [BIA03], kuris dažnai taikomas mikrogardelių metodu gautų duomenų analizėje. Jis naudojamas, kai norima kelių eksperimentų duomenis padaryti palyginamus, kad reikšmių pasiskirstymas tarp eksperimentų duomenų būtų toks pat, ir pašalinti esantį techninį variabilumą [GT09]. Po normalizavimo galima taikyti statistinius metodus.

4.5.5 Sisteminių paklaidų pašalinimas

Duomenyse gali būti sisteminių paklaidų (angl. *Batch effects*) [LS07] [CGB+11], t. y. priklausomybių nuo atlikto eksperimento techninių niuansų: laiko, slėgio, temperatūros ir kt. Jeigu tokios priklausomybės nustatytos, jos taip pat turi būti pašalinamos.

Sisteminėms paklaidoms nustatyti gali būti panaudota principinių komponenčių analizė. Po jos imamos ir atvaizduojamos pirmosios principinės komponentės. Jei mėginiai jose atsiskiria į grupes, bet ne pagal sergančius/sveikus mėginius ar pagal vėžio formą, tuomet galima manyti, kad duomenyse yra sisteminės paklaidos. Pavyzdžiui, duomenyse gali būti stiprus lyties efektas, kai mėginiai pirmose principinėse komponentėse labai gerai atskiriami pagal lytį, bet ne pagal diagnozę. Tai gali trukdyti tolesniame duomenų apdorojime, klasifikatoriaus konstravime. Jei žinoma, dėl ko atsiranda tokie efektai, pavyzdžiui, dėl požymių susijusių su lytinėmis chromosomomis, pašalinus tokius požymius efektas išnyksta.

4.5.6 Mėginių atsiskyrėlių šalinimas

Mėginiai atsiskyrėliai iš duomenų pašalinami naudojant principinių komponenčių metodą. Imamos pirmosios 3 principinės komponentės. Kiekvienai jų apskaičiuojamas duomenų

vidurkis ir du standartiniai nuokrypiai (2 Std). Mėginiai nutolę bent vienoje principinėje komponentėje per du standartinius nuokrypius nuo vidurkio yra pašalinami. Mėginiai atsiskyrėliai gali apsunkinti klasifikatoriaus konstravimą, todėl tolesnėje duomenų analizėje nenaudojami.

4.5.7 Signalo neturinčių požymių pašalinimas

Iš duomenų pašalinami požymiai neturintys signalo. Arčiau nei per ∼200 bp (bazių porų) nuo jų cirDNR sekų nėra „ACGT", „CCGG", „GCGC" motyvų.

4.5.8 Požymių glodinimas

Perskaičiuojamos požymių vertės pagal arčiausiai esančius požymius naudojant glodinimo metodą (angl. *Kernel Regression Smoother*).

5 Rezultatai

Toliau aprašomi rezultatai gauti po duomenų apdorojimo ir analizės pritaikius mašininio mokymosi metodus.

5.1 Tiesiosios žarnos vėžio duomenų tyrimas

Pradinius duomenis sudaro 393 mėginiai Lent. 1, 393 pacientų DNR metilinimo duomenys, ir 6553600 požymių – DNR metilinimo įverčiai įvairiose žmogaus genomo (hg19) vietose. Buvo atlikti du eksperimentai. Pirmojo eksperimento (angl. *Batch 1*) duomenys sudaryti iš 193 skirtingų pacientų mėginių, antrojo (angl. *Batch 2*) iš 200 skirtingų pacientų mėginių. Pacientai priklauso sergančiųjų, 193 pacientai (CRC), ir sveikųjų (kontrolinė – CTR), 200 pacientų, grupėms.

1 lentelė: Pradiniai duomenys.

	Batch 1		***Batch 2***		
	Vyrai	**Moterys**	**Vyrai**	**Moterys**	**Iš viso:**
Kontrolė	50	50	50	50	200
Sergantys	46	47	65	35	193
	96	97	115	85	393
	193		200		

5.1.1 Duomenų tyrimo seka

Mikrogardelių metodu gautų duomenų tyrimas daug žingsnių turintis uždavinys, todėl pirmiausia apibrėžiama duomenų apdorojimo strategija tiesiosios žarnos vėžio cirDNR metilinimo duomenims tirti Pav. 5.

Pirmiausia perkeliamos požymių koordinatės iš hg18 į naujausiai anotuotą hg19.

CirDNR metilinimo duomenų pasiskirstymas nėra simetrinis, todėl atliekama duomenų log_2 transformacija.

Pirmojo eksperimento (angl. *Batch 1*) duomenys naudojami apmokymui, antrojo (angl. *Batch 2*) – testavimui. Apmokymo aibės duomenys normalizuojami kvantilių normalizavimo metodu. Testavimo aibės duomenys normalizuojami pagal apmokymo aibės duomenis, panaudojant normalizuotų apmokymo aibės duomenų reikšmių pasiskirstymą. Gaunamas reikšmių pasiskirstymas panašus abiejose duomenų aibėse, duomenys palyginami.

Mėginiai atsiskyrėliai iš duomenų pašalinami apmokymo ir testavimo aibėms atskirai. Analizuojant apmokymo aibės duomenis gauti 2 mėginiai atsiskyrėliai, o testavimo – 4 mėginiai atsiskyrėliai. Pav. 6 vaizduoja mėginių atsiskyrėlių išsidėstymą naudojant apmokymo aibės duomenis (*Batch 1*) gautose principinėse komponentėse. Punktyrinėmis linijomis kairėje ir dešinėje pažymėti kiekvienos principinės komponentės 2 standartiniai nuokrypiai nuo vidurkio, kuris pažymėtas vidurine tiese.

Sekančiame žingsnyje pašalinami požymiai neturintys signalo.

Tada perskaičiuojami požymių įverčiai naudojant glodinimo (angl. *Smoothing*) metodą.

Atlikus duomenų apdorojimą prasideda jų analizė. Analizės metu mažinamas požymių skaičius vykdant požymių atranką. Naudojami požymių filtravimo ir požymių atrinkimo kartu su klasifikatoriumi metodai. Požymių atranka atliekama, kadangi naudojant visą požymių aibę klasifikatorius gali persimokyti, dalis požymių nėra reikšmingi ir reikalingi, naudojant visą aibę ilgėja klasifikatoriaus apsimokymo laikas, norima rasti svarbiausius požymius, kurie būtų reikšmingi biologiškai.

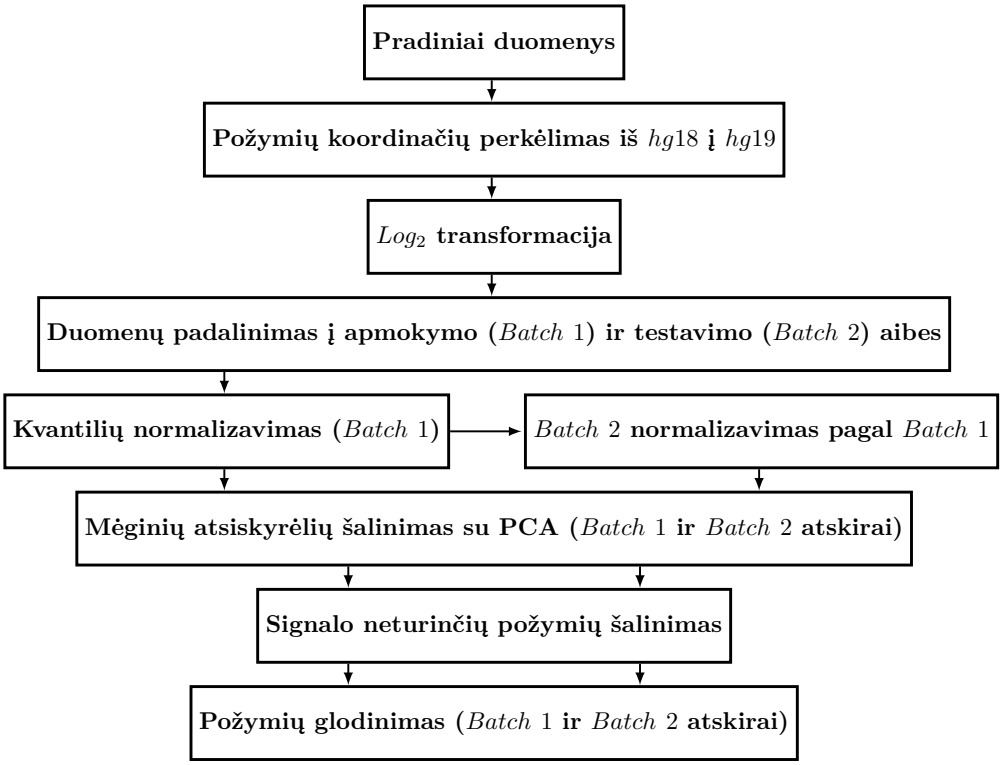
Atrinkus požymius konstruojamas klasifikatorius naudojant pasirinktus mašininio mokymosi algoritmus. Išbandoma keletas mašininio mokymosi algoritmų, kadangi duomenų savybės nėra žinomos. Gauti rezultatai palyginami, sprendžiama, kuris klasifikatorius tinkamiausias.

Analizuojant duomenis vykdoma kryžminė validacija CV (angl. *Cross validation*), kai mėginių aibė padalinama į pasirinktą skaičių dalių, pavyzdžiui, 10 lygių dalių. Vykdoma 10 iteracijų, kiekvienos metu 9 dalys naudojamos apmokymui ir 1 testavimui. Dešimties kartų kryžminės validacijos metu išmokstama, kiek geriausia naudoti požymių galutinio klasifikatoriaus konstravimui po jų atrinkimo, ar kokius geriausia naudoti klasifikatoriaus parametrus.

5.1.2 Požymių atranka naudojant t testą

Pradinių požymių labai daug, dalis jų triukšmas ir tik kai kurie reikalingi klasifikatoriaus konstravimui. Palikus tik signalą turinčius požymius iš 6553600 lieka 2200552, bet vis dar aktuali $p \gg N$ problema. Kadangi požymių kur kas daugiau negu tiriamų mėginių, naudojami įvairūs požymių atrinkimo metodai duomenų požymių aibei sumažinti.

Statistiniai filtravimo metodai naudojami atrinkti požymiams, kurių raiška reikšmingai pakitusi vėžiu sergančiųjų grupėje lyginant su kontroline grupe. Atlikus testą gaunamos p



5 pav.: Duomenų apdorojimo schema.

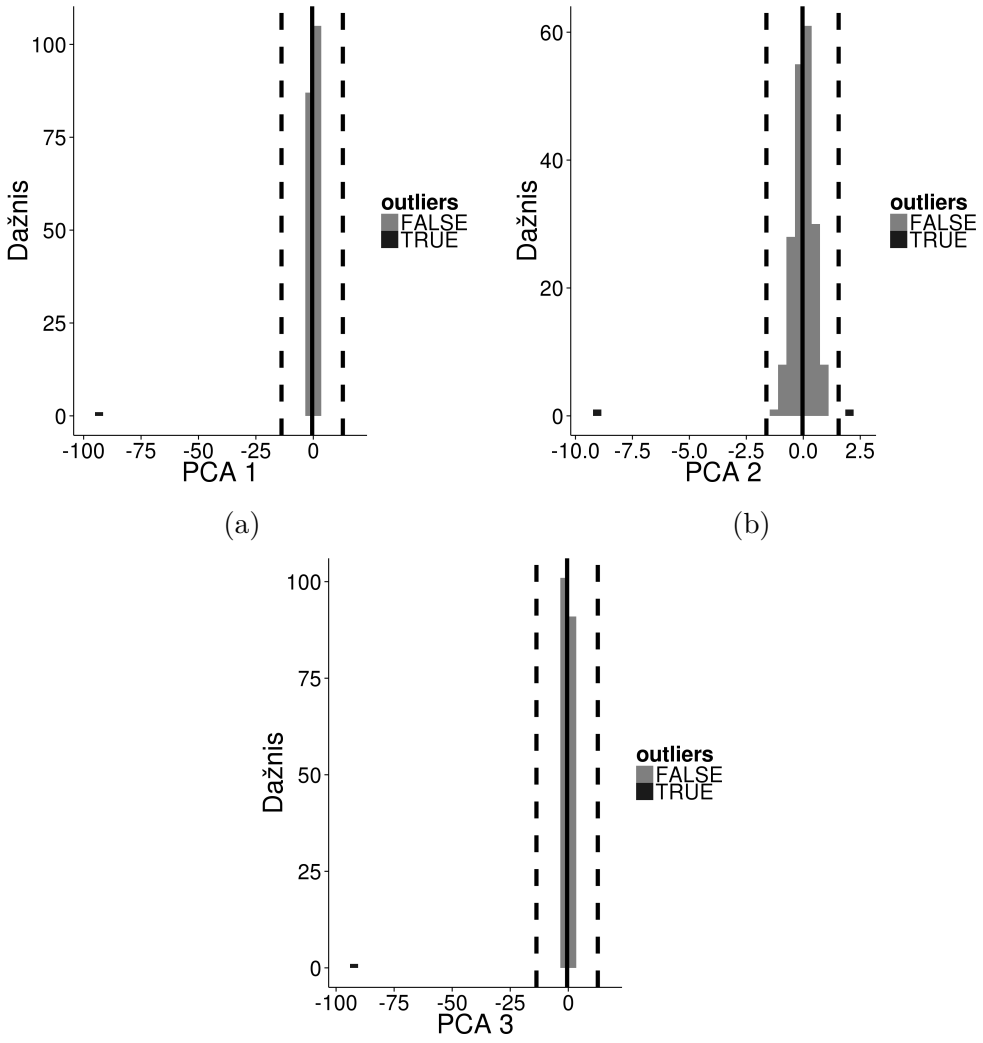
vertės ir pasirenkami tik tie požymiai, kurių p vertės, pavyzdžiui, mažesnės ir lygios 0.01, tai reiškia, kad tie požymiai statistiškai reikšmingi ir klaidos tikimybė, kad juose reikšmingo pokyčio iš tikrųjų nėra, labai maža. Naudojant atrinktus reikšmingai pakitusius požymius toliau galima konstruoti klasifikatorių.

Požymiai atrenkami naudojant t testą (angl. *Welch's t test*). Apskaičiuojamos p vertės ir pasirenkami požymiai, kurių p vertės < 0.01. Apmokymo aibės (*Batch 1*) duomenyse esantį reikšmingą pokytį galima pastebėti p verčių pasiskirstymo histogramoje Pav. 7 (a). Joje p vertės pasiskirsčiusios nevienodai, daug mažų p verčių. Testavimo aibės (*Batch 2*) duomenyse yra daugiau reikšmingų požymių Pav. 7 (b).

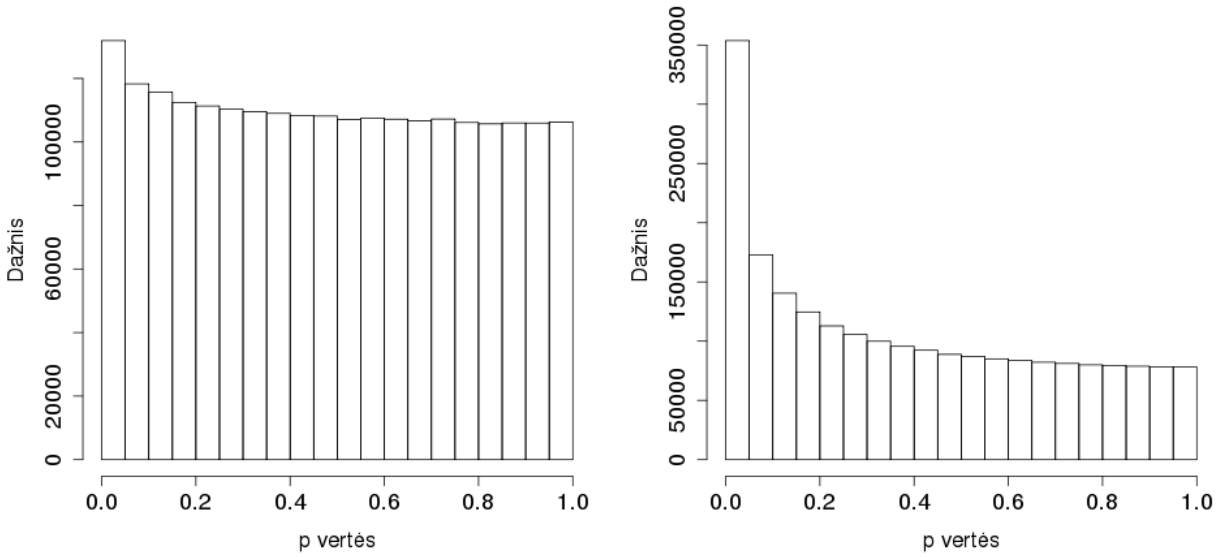
5.1.3 Požymių atranka naudojant pokytį tarp grupių vidurkių (angl. *Fold Change*)

Kartu su p vertėmis požymių atrinkimui naudojamas ir pokytis tarp grupių vidurkių FC (angl. *Fold Change*). Imami požymiai, kurių pokytis nutolęs per du standartinius nuokrypius nuo bendro vidurkio. Pav. 8 vaizduoja pokyčių tarp grupių vidurkių pasiskirstymus apmokymo (a) ir testavimo (b) duomenyse, reikšmingi požymiai nutolę į kairę nuo pirmosios vertikalios punktyrinės linijos ir nutolę į dešinę nuo antrosios vertikalios punktyrinės linijos.

Klasifikatoriaus konstravimui pasirenkami tie požymiai, kurių maža p vertė (< 0.01) ir pokytis tarp grupių vidurkių nutolęs per du standartinius nuokrypius nuo bendro pokyčių vidurkio. Tokie reikšmingi požymiai pažymėti „Volcano" grafike Pav. 9 (a), kurio x ašyje skirtumas tarp požymių vidurkių sergančiųjų ir kontrolinėje grupėse, o y ašyje atitinkamai kiekvieno požymio neigiama p vertė logaritmu pagrindu 10. Grafike taip pat pažymėti



6 pav.: Apmokymo aibės (*Batch 1*) mėginiai atsiskyrėliai pirmoje (a), antroje (b) ir trečioje (c) principinėse komponentėse. Mėginiai esantys už punktyrinių linijų, už dviejų standartinių nuokrypių nuo vidurkio, yra mėginiai atsiskyrėliai.



(a) (b)

7 pav.: P verčių pasiskirstymai apmokymo (a) ir testavimo (b) aibėse.

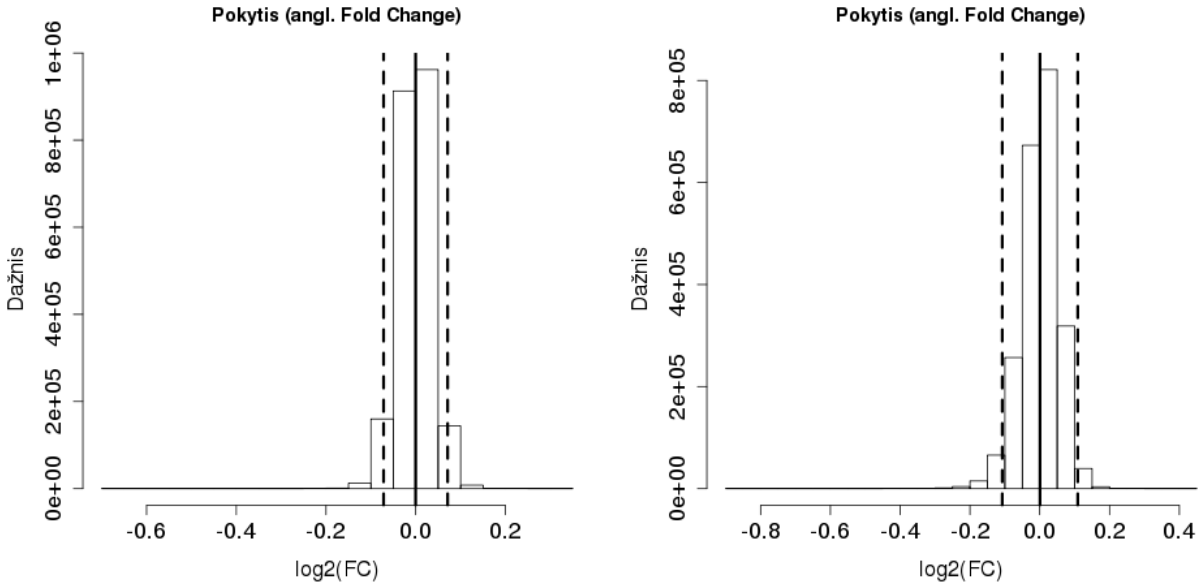
požymiai, kurių reikšminga tik p vertė (< 0.01), ir kuriems reikšmingas tik pokytis tarp grupių vidurkių. Verti dėmesio požymiai nutolę nuo vidurio x ašyje ir esantys aukštai y ašyje, t. y. kurių didelis skirtumas tarp grupių vidurkių ir labai maža p vertė.

Tokiu būdu atrinkti 27789 požymiai iš apmokymo aibės duomenų, kurie naudojami klasifikatoriaus konstravime.

5.1.4 Klasifikatoriaus konstravimas naudojant atsitiktinius miškus ir atraminių vektorių mašinas

Atrinkus požymius prieš tai aprašytu filtravimo metodu, imami požymiai, kurių reikšminga p vertė ir pokytis tarp grupių vidurkių. Pav. 10 kairėje parodyti veiksmai atliekami su apmokymo aibe (*Batch 1*), dešinėje su testavimo aibe (*Batch 2*).

Naudojant mokymo aibę apmokomi atsitiktiniai miškai ir atraminių vektorių mašinos, kurie testuojami su testine aibe. Pav. 11 rodo kaip pavyko apmokyti (a) klasifikatorius ir juos testuoti (b). Lent. 2 pavaizduoti klasifikatorių testavimo rezultatai. Pirmame stulpelyje – klasifikatoriaus pavadinimas, antrame – testavimo tikslumas ir „AUC" vertė (angl. *The Area Under an ROC Curve*. Klasifikatoriaus tikslumo įvertis) atitinkamai trečiame. *Stage 0* stulpelyje – kokia dalis sveikų pacientų buvo suklasifikuota teisingai. *Stage I* – kokia dalis pacientų sergančių pirmos stadijos vėžiu suklasifikuota teisingai. Atitinkamai *Stage IIA* ir *Stage IIB* (sergančių II stadijos B vėžio forma *Batch 2* duomenyse nėra). Visi klasifikatoriai testavo panašiai (tikslumas apie 0.68).



(a) (b)

8 pav.: Apmokymo (a) ir testavimo (b) duomenų pokyčių tarp grupių vidurkių pasiskirstymai. Vidurinė linija žymi vidurkį, kairėje ir dešinėje esančios punktyrinės linijos žymi du standartinius nuokrypius nuo vidurkio.

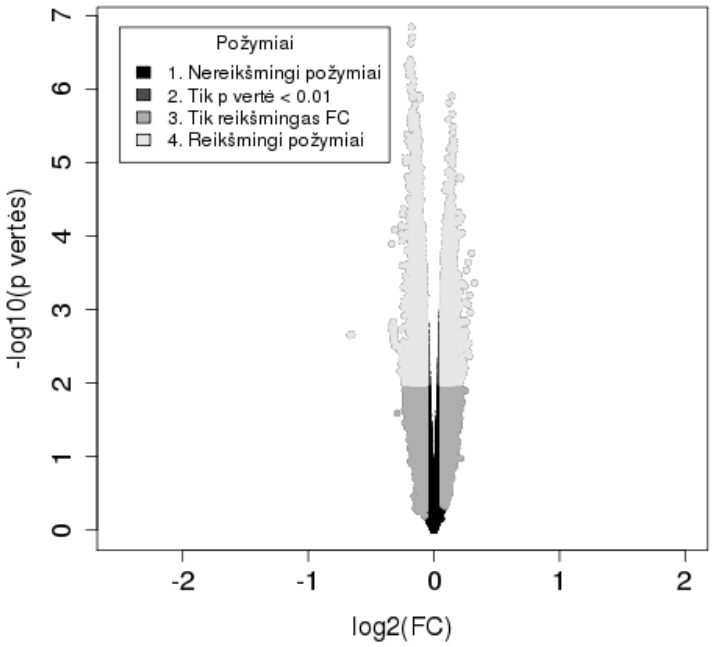
2 lentelė: Atsitiktinių miškų ir atraminių vektorių mašinų klasifikatorių testavimo rezultatai.

	Tikslumas	**AUC**	**Stage 0**	**Stage I**	**Stage IIA**	**Stage IIB**
rf	0.67	0.78	51/97 (0.53)	68/83 (0.82)	14/16 (0.88)	NA
svmLinear	0.67	0.73	50/97 (0.52)	68/83 (0.82)	15/16 (0.94)	NA
svmRadial	0.68	0.77	44/97 (0.45)	75/83 (0.90)	15/16 (0.94)	NA
svmPoly	0.68	0.73	52/97 (0.54)	67/83 (0.81)	15/16 (0.94)	NA

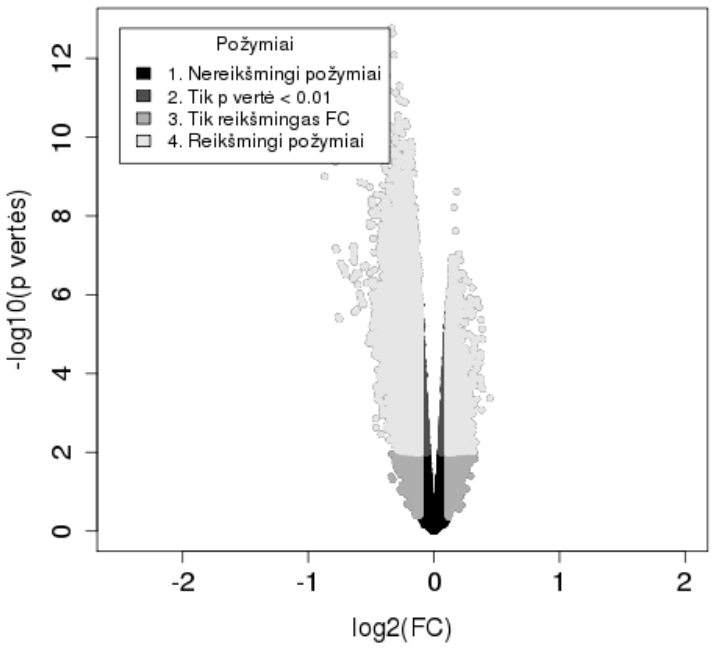
5.1.5 Požymių atranka naudojant atsitiktinių miškų metodą

Dėl nepakankamai gerų prieš tai gautų klasifikatorių testavimo rezultatų, požymių aibę nutarta sumažinti naudojant požymių atrinkimą kartu su klasifikatoriumi. Pav. 12 kairėje veiksmai atliekami su apmokymo aibe (*Batch 1*), dešinėje su testavimo aibe (*Batch 2*).

Pirmiausia atliekamas požymių filtravimas. Kaip ir prieš tai, imami požymiai iš pirmojo eksperimento duomenų, kuriems reikšminga p vertė (< 0.01) ir pokytis tarp grupių vidurkių nutolęs per du standartinius nuokrypius nuo bendro pokyčių vidurkio. Tada požymiai atrenkami dar kartą panaudojant požymių atrinkimą kartu su klasifikatoriumi – atsitiktinių miškų metodu [UdA06]. Po atrinkimo gauti 106 požymiai. Apmokymui ir testavimui imami tik atrinkti požymiai. Tada dar kartą apmokomi atsitiktiniai miškai ir atraminių vektorių mašinos. Pav. 13 matoma kaip kiekvienam klasifikatoriui sekėsi apsimokyti (a) ir kaip gautiems klasifikatoriams pavyko testuoti (b). Lent. 3 pavaizduoti gauti testavimo rezultatai. Pirmame stulpelyje – klasifikatoriaus pavadinimas, antrame – testavimo tikslumas, trečiame stulpelyje „AUC" vertė. *Stage 0* stulpelyje – kokia dalis sveikų pacientų buvo suklasifikuota teisingai. *Stage I* – kokia dalis pacientų sergančių pirmos stadijos vėžiu suklasifikuota

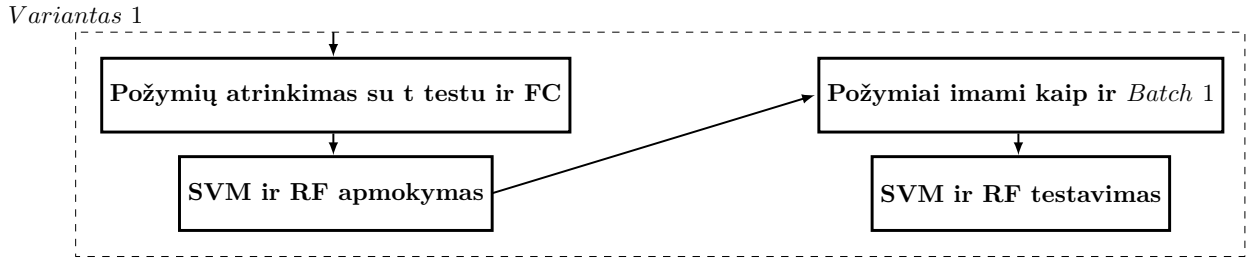


(a)



(b)

9 pav.: Apmokymo (a) ir testavimo (b) duomenų „Volcano” grafikai.



10 pav.: Duomenų analizės schema.

teisingai. Atitinkamai *Stage IIA* ir *Stage IIB* (sergančių II stadijos B vėžio forma *Batch 2* duomenyse nėra). Geriausiai testuoti sekėsi atsitiktinių miškų klasifikatoriui (tikslumas 0.74).

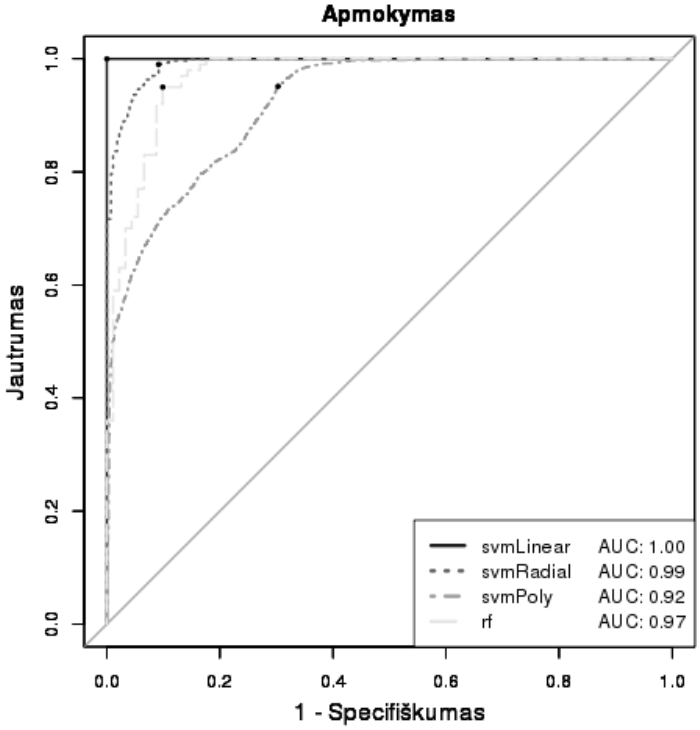
3 lentelė: Atsitiktinių miškų ir atraminių vektorių mašinų klasifikatorių testavimo rezultatai.

	Tikslumas	**AUC**	**Stage 0**	**Stage I**	**Stage IIA**	**Stage IIB**
rf	0.74	0.82	64/97 (0.66)	69/83 (0.83)	13/16 (0.81)	NA
svmLinear	0.61	0.68	43/97 (0.44)	65/83 (0.78)	13/16 (0.81)	NA
svmRadial	0.66	0.75	50/97 (0.52)	68/83 (0.82)	13/16 (0.81)	NA
svmPoly	0.65	0.74	54/97 (0.56)	64/83 (0.77)	11/16 (0.69)	NA

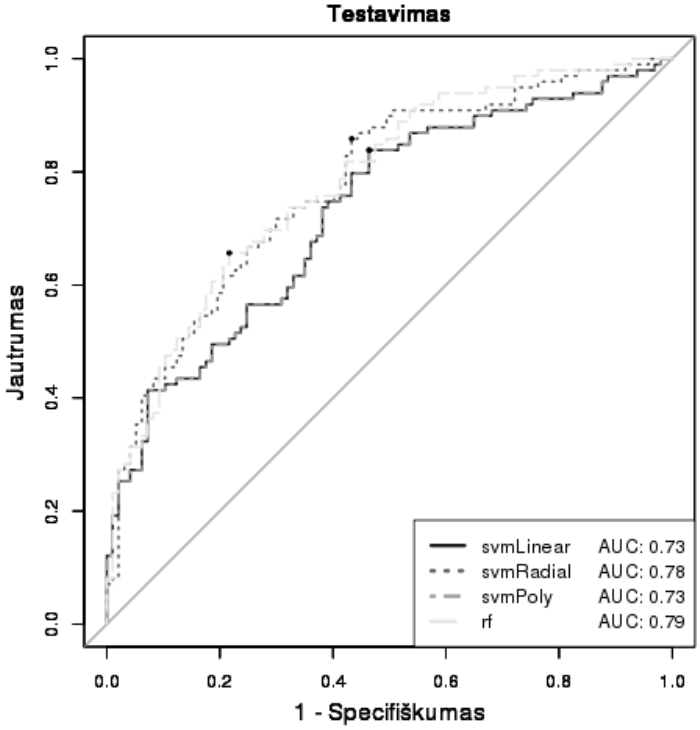
5.1.6 Požymių atranka imant variabiliausius požymius

Pav. 14 pavaizduota atliktos analizės schema, kairėje veiksmai atliekami su apmokymo aibe (*Batch 1*), dešinėje su testavimo aibe (*Batch 2*). Apmokymo aibė dalinama į 10 poaibių (angl. *Folds*). 9 poaibiai naudoti atrenkant 1000 variabiliausių požymių. Tada sudaromi k dydžio poaibiai: 100, 300, 500, 700 ir 1000 pirmųjų variabiliausių požymių. Vykdomas ciklas, kurio metu su kiekvienu k dydžio poaibiu apmokomas atsitiktinių miškų klasifikatorius, kuris testuojamas su atitinkamai k dydžio testavimui skirtu poaibiu, gautu iš likusio 1 poaibio padalinus apmokymo aibę. Po ciklo įvykdymo grąžinamas geriausias k, t. y. testuojant su k dydžio požymių poaibiu buvo gautas didžiausias tikslumas. Po 10 kryžminės validacijos iteracijų gaunama 10 geriausių k, iš kurių išvedamas vidurkis.

Po 10 kartų kryžminės validacijos iteracijų išmokus kiek variabiliausių požymių naudoti geriausia (vidutinis $k = 240$), konstruojant galutinį klasifikatorių iš apmokymo aibės duomenų (*Batch 1*) ir jį testuojant su testine aibe (*Batch 2*) naudojamas tik išmoktas skaičius variabiliausių požymių. Su mokymo aibe apmokomi galutiniai atsitiktinių miškų ir atraminių vektorių mašinų klasifikatoriai, kurie testuojami su testine aibe. Gautose „Roc" kreivėse Pav. 15 matoma, kaip klasifikatoriams sekėsi apsimokyti ir testuoti. Lent. 4 pavaizduoti testavimo rezultatai. Pirmuosiuose stulpeliuose – klasifikatoriaus pavadinimas, testavimo tikslumas ir „AUC" vertė. *Stage 0* stulpelyje – kokia dalis sveikų pacientų buvo suklasifikuota teisingai. *Stage I* – kokia dalis pacientų sergančių pirmos stadijos vėžiu suklasifikuota teisingai. Atitinkamai *Stage IIA* ir *Stage IIB* (sergančių II stadijos B vėžio forma *Batch 2* duomenyse nėra). Visų klasifikatorių testavimo tikslumas mažas (apie 0.6).

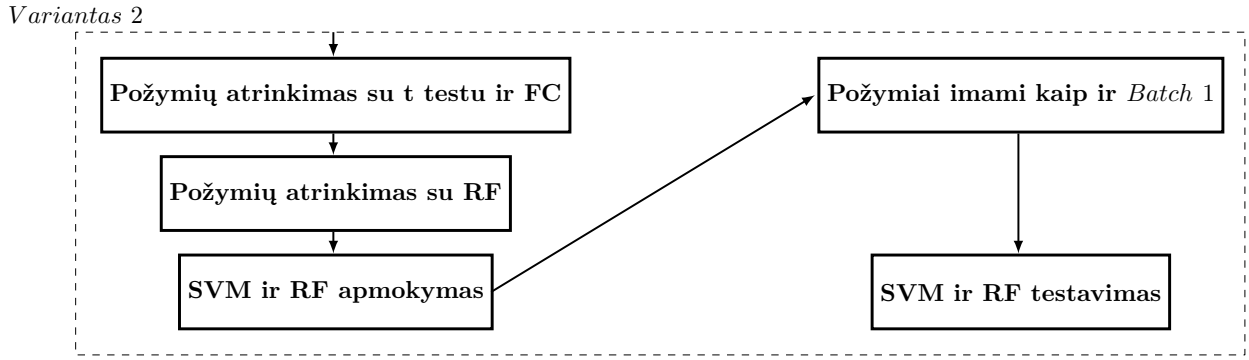


(a)



(b)

11 pav.: Po apmokymo (a) ir testavimo (b) gautos „ROC" kreivės atraminių vektorių mašinų ir atsitiktinių miškų klasifikatoriams, kai požymiai analizei buvo atrinkti imant statistiškai reikšmingiausius ir kuriems reikšmingas pokytis tarp grupių vidurkių.



12 pav.: Duomenų analizės schema.

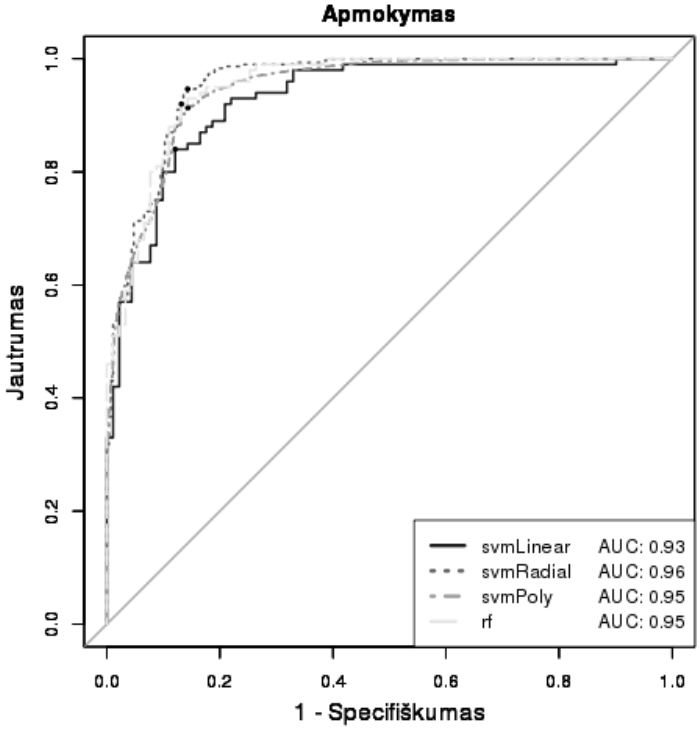
4 lentelė: Atsitiktinių miškų ir atraminių vektorių mašinų klasifikatorių testavimo rezultatai.

	Tikslumas	**AUC**	**Stage 0**	**Stage I**	**Stage IIA**	**Stage IIB**
rf	0.61	0.64	34/97 (0.35)	74/83 (0.89)	12/16 (0.75)	NA
svmLinear	0.55	0.58	65/97 (0.67)	37/83 (0.45)	6/16 (0.38)	NA
svmRadial	0.51	0.51	76/97 (0.78)	17/83 (0.20)	8/16 (0.50)	NA
svmPoly	0.58	0.63	62/97 (0.64)	45/83 (0.54)	8/16 (0.50)	NA

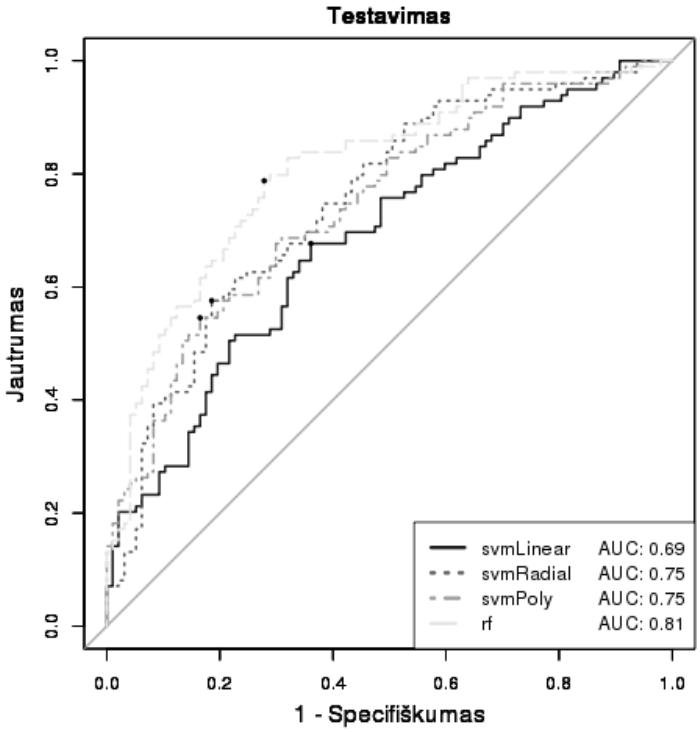
5.1.7 Požymių atranka imant statistiškai reikšmingiausius požymius

Atlikta panaši analizė, kaip ir prieš tai su variabiliausiais požymiais. Analizės schema Pav. 14, kairėje veiksmai atliekami su apmokymo aibe (*Batch 1*), dešinėje su testavimo aibe (*Batch 2*). Apmokymo aibės mėginiai dalinami į 10 poaibių. Vykdant 10 kartų kryžminę validaciją 9 poaibiai naudojami atrenkant 1000 statistiškai reikšmingiausių požymių (požymiai rikiuojami pagal p vertes didėjimo tvarka. Imama 1000 pirmųjų). Tada sudaromi k dydžio poaibiai: 100, 300, 500, 700 ir 1000 pirmųjų reikšmingiausių požymių. Vykdomas ciklas, kurio metu su kiekvienu k dydžio poaibiu apmokomas atsitiktinių miškų klasifikatorius, kuris testuojamas su atitinkamai k dydžio testiniu poaibiu, likusiu 1 mėginių poaibiu skirtu testavimui. Po ciklo įvykdymo grąžinamas geriausias k, t. y. testuojant su k dydžio požymių poaibiu gautas didžiausias tikslumas. Po 10 kryžminės validacijos iteracijų gaunama 10 geriausių k, išvedamas jų vidurkis.

Po 10 kryžminės validacijos iteracijų išmokus, kiek statistiškai reikšmingiausių požymių naudoti geriausia (vidutinis $k = 340$), galutinio klasifikatoriaus apmokymui ir testavimui imamas išmoktas apmokymo aibės (*Batch 1*) ir testavimo aibės (*Batch 2*) požymių skaičius. Su mokymo aibe apmokomi galutiniai atsitiktinių miškų ir atraminių vektorių mašinų klasifikatoriai, kurie testuojami su testine aibe. Rezultatai pavaizduoti Lent. 5. Pirmame stulpelyje – klasifikatoriaus pavadinimas, antrame – testavimo tikslumas ir „AUC" vertė atitinkamai trečiame. *Stage 0* stulpelyje – kokia dalis sveikų pacientų buvo suklasifikuota teisingai. *Stage I* – kokia dalis pacientų sergančių pirmos stadijos vėžiu suklasifikuota teisingai. Atitinkamai *Stage IIA* ir *Stage IIB* (sergančių II stadijos B vėžio forma *Batch 2* duomenyse nėra). Geriausias testavimo tikslumas 0.73. Pav. 16 (a) rodo kaip pavyko apmokyti klasifikatorius, o (b) kaip gerai jie testavo.

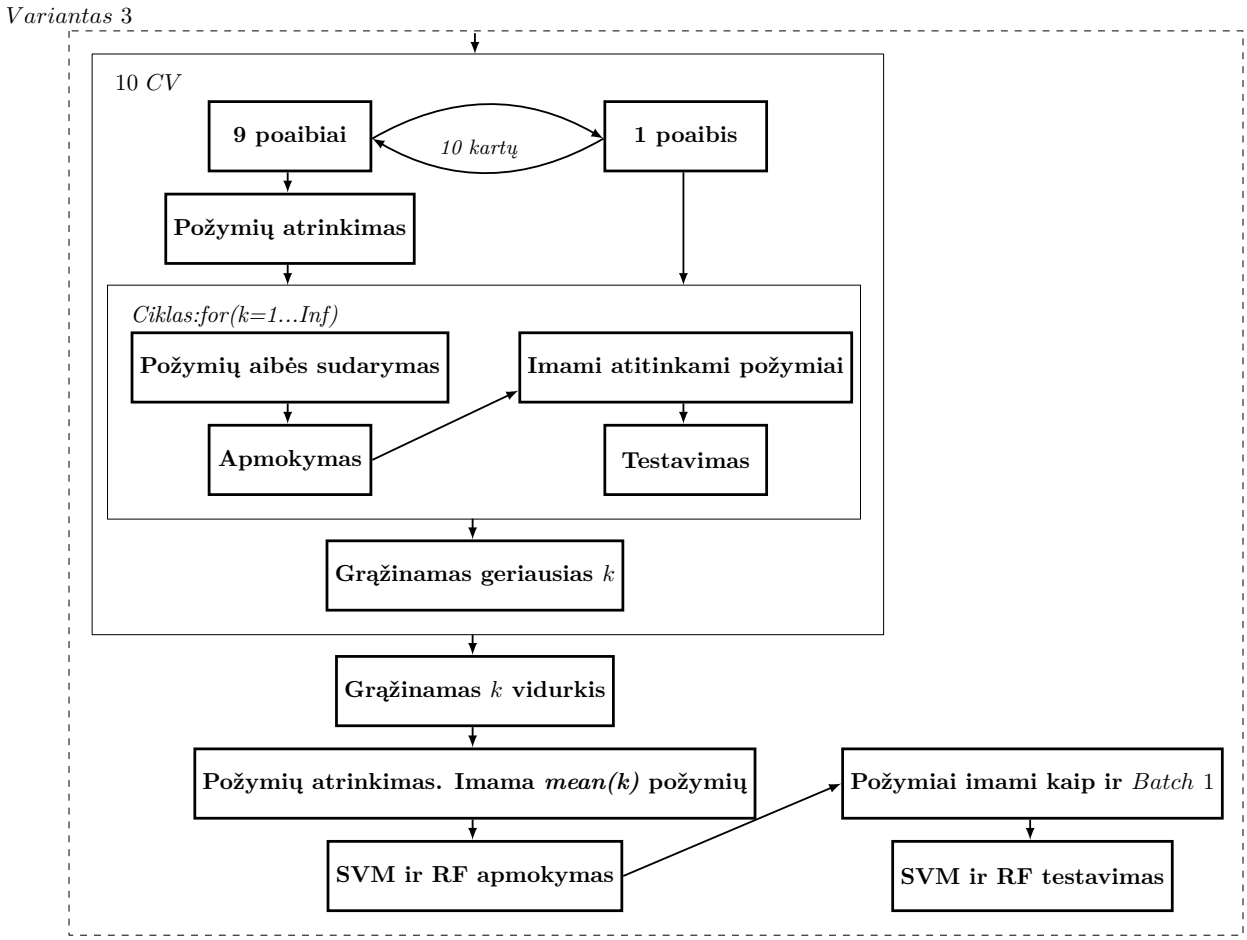


(a)

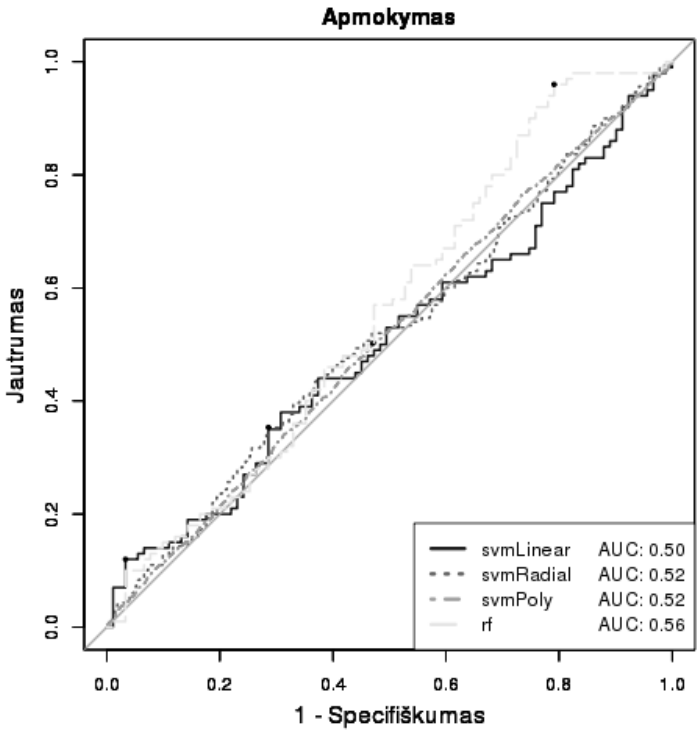


(b)

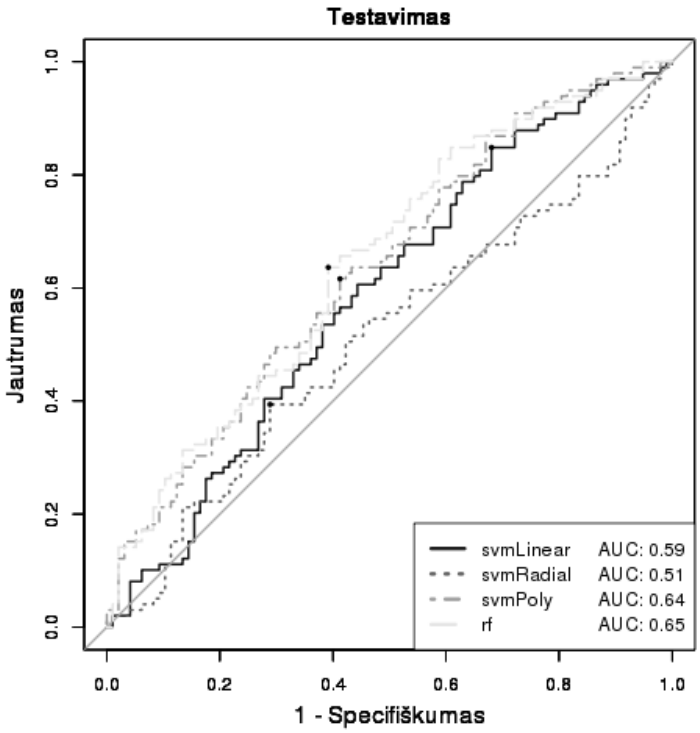
13 pav.: Po apmokymo (a) ir testavimo (b) gautos „ROC" kreivės atraminių vektorių mašinų ir atsitiktinių miškų klasifikatoriams, kai požymiai analizei buvo atrinkti pirmiausia imant statistiškai reikšmingiausius ir kuriems reikšmingas pokytis tarp grupių vidurkių (FC) ir tada atlikus požymių atranką kartu su atsitiktinių miškų klasifikatoriumi.



14 pav.: Duomenų analizės schema.

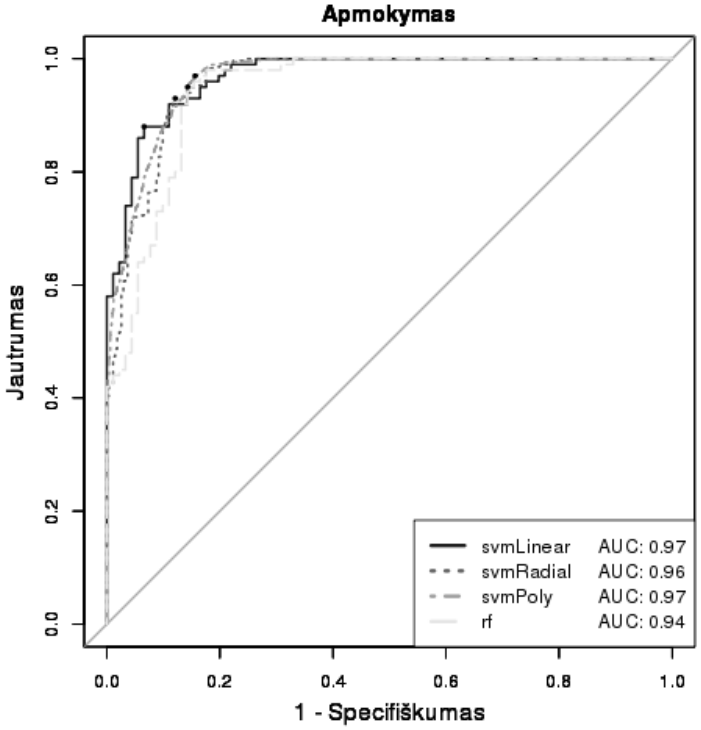


(a)

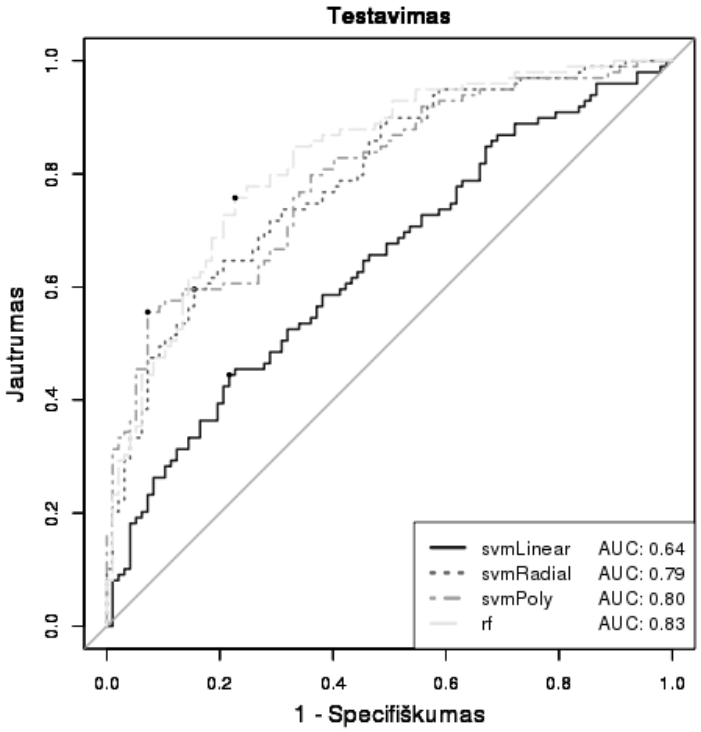


(b)

15 pav.: Po apmokymo (a) ir testavimo (b) gautos „ROC" kreivės atraminių vektorių mašinų ir atsitiktinių miškų klasifikatoriams. Naudojami atrinkti vidutinis $k = 240$ variabiliausių požymių.



(a)



(b)

16 pav.: Po apmokymo (a) ir testavimo (b) gautos „ROC" kreivės atraminių vektorių mašinų ir atsitiktinių miškų klasifikatoriams. Naudojamas vidutinis $k = 340$ atrinktų statistiškai reikšmingiausių požymių, kurių mažiausios p vertės.

5 lentelė: Atsitiktinių miškų ir atraminių vektorių mašinų klasifikatorių testavimo rezultatai.

	Tikslumas	**AUC**	**Stage 0**	**Stage I**	**Stage IIA**	**Stage IIB**
rf	0.73	0.82	60/97 (0.62)	70/83 (0.84)	15/16 (0.94)	NA
svmLinear	0.58	0.64	44/97 (0.45)	59/83 (0.71)	11/16 (0.69)	NA
svmRadial	0.68	0.79	50/97 (0.52)	71/83 (0.86)	14/16 (0.88)	NA
svmPoly	0.70	0.79	58/97 (0.60)	67/83 (0.81)	14/16 (0.88)	NA

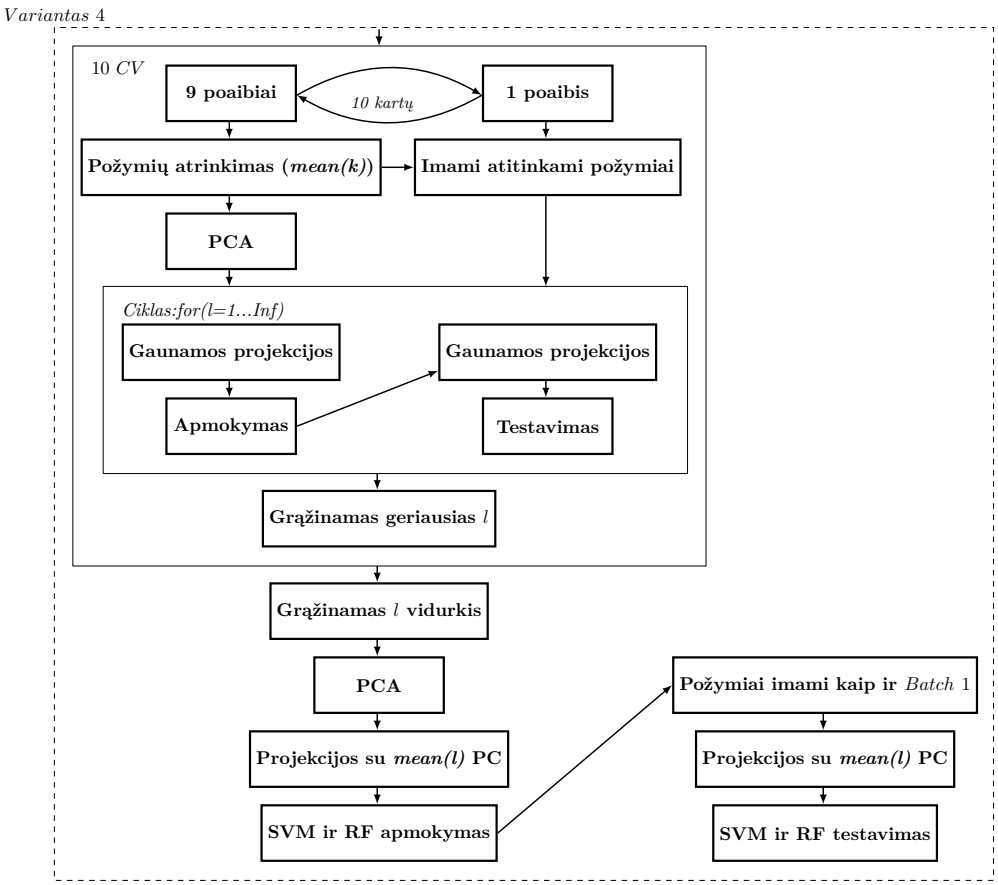
5.1.8 Dimensionalumo mažinimas su PCA, principinių komponenčių atranka

Pav. 17 pavaizduota atliktos analizės schema, kairėje veiksmai atliekami su apmokymo aibe (*Batch 1*), dešinėje su testavimo aibe (*Batch 2*). Apmokymo aibės mėginiai dalinami į 10 poaibių. Naudojant 9 poaibius imama vidutinis $k = 340$ statistiškai reikšmingiausių požymių (požymiai rikiuojami pagal p vertes didėjimo tvarka. Imama $k = 340$ pirmųjų). Atliekama principinių komponenčių analizė. Sudaromi l dydžio poaibiai: 5, 15, 25, 35, 45, 55, 65, 75, 85, 95, 105, 115 ir 125 pirmųjų principinių komponenčių. Vykdomas ciklas, kurio metu su kiekvienu l dydžio principinių komponenčių skaičiumi gaunamos 9 apmokymo aibės poaibių projekcijos, su jomis apmokomas atsitiktinių miškų klasifikatorius. Atitinkamai gaunamos testavimui skirto poaibio projekcijos. Tuomet gautas klasifikatorius testuojamas. Po ciklo įvykdymo grąžinamas geriausias l, t. y. principinių komponenčių skaičius, kurį naudojant gautas didžiausias testavimo tikslumas. Po 10 kryžminės validacijos iteracijų gaunama 10 geriausių l, išvedamas jų vidurkis.

Imant vidutinis $k = 340$ statistiškai reikšmingiausių apmokymo aibės (*Batch1*) ir testavimo aibės (*Batch 2*) požymių atliekama principinių komponenčių analizė su apmokymo aibės duomenimis. Gaunamos apmokymo aibės duomenų projekcijos naudojant išmoktą vidutinis $l = 25$ pirmųjų principinių komponenčių. Atitinkamai gaunamos ir testavimo aibės projekcijos. Su mokymo aibės projekcijomis apmokomi galutiniai atsitiktinių miškų ir atraminių vektorių mašinų klasifikatoriai, kurie testuojami su testinės aibės duomenų projekcijomis. Rezultatai pavaizduoti Lent. 6. Pirmame stulpelyje – klasifikatoriaus pavadinimas, antrame – testavimo tikslumas, trečiame – „AUC" vertė. *Stage 0* stulpelyje – kokia dalis sveikų pacientų buvo suklasifikuota teisingai. *Stage I* – kokia dalis pacientų sergančių pirmos stadijos vėžiu suklasifikuota teisingai. Atitinkamai *Stage IIA* ir *Stage IIB* (sergančių II stadijos B vėžio forma *Batch 2* duomenyse nėra). Pav. 16 (a) rodo kaip pavyko apmokyti klasifikatorius, o (b) kaip gerai jie testavo.

6 lentelė: Atsitiktinių miškų ir atraminių vektorių mašinų klasifikatorių testavimo rezultatai.

	Tikslumas	**AUC**	**Stage 0**	**Stage I**	**Stage IIA**	**Stage IIB**
rf	0.79	0.80	75/97 (0.77)	68/83 (0.82)	13/16 (0.81)	NA
svmLinear	0.64	0.69	60/97 (0.62)	56/83 (0.67)	11/16 (0.69)	NA
svmRadial	0.66	0.72	63/97 (0.65)	57/83 (0.69)	11/16 (0.69)	NA
svmPoly	0.67	0.73	54/97 (0.56)	66/83 (0.80)	12/16 (0.75)	NA



17 pav.: Duomenų analizės schema.

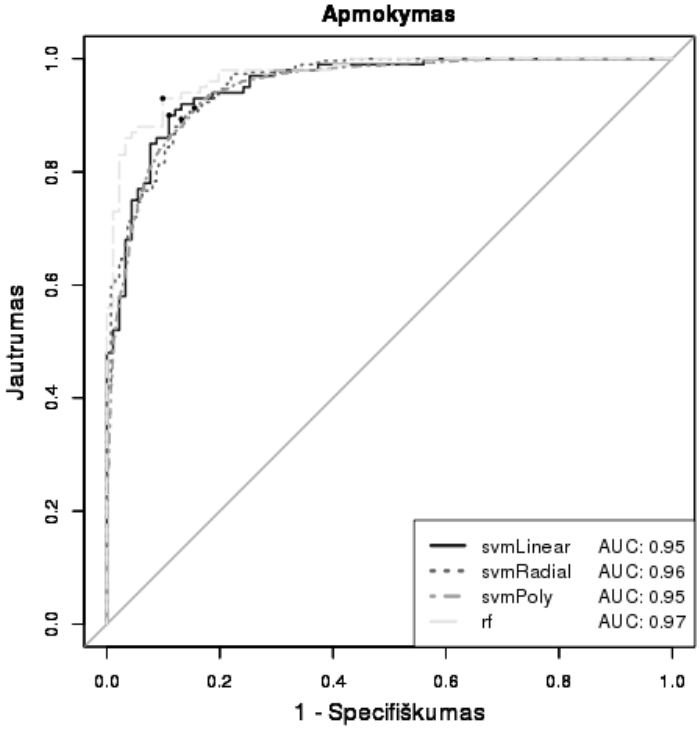
5.2 Kitų vėžio formų duomenų tyrimas

Tiriami duomenys gauti iš 3 skirtingų eksperimentų [LHA+14] (duomenys imti iš *Gene Expression Omnibus (GEO)* internetinės duomenų bazės), kuriuose buvo tirtas kraujyje esančių leukocitų DNR metilinimas šlapimo pūslės vėžio atvejais (angl. *Bladder cancer*) (GEO id numeris: GSE50409), galvos ir kaklo suragėjusių ląstelių karcinomos atvejais (angl. *Head and neck squamous cell carcinoma (HNSCC)*) (GEO id numeris: GSE30229) ir kiaušidžių vėžio atvejais (angl. *Ovarian cancer*) (GEO id numeris: GSE19711). Visuose eksperimentuose periferinio kraujo mėginiai tirti mikrogardelių metodu naudojant *Illumina Infinium 27k Human DNA methylation Beadchip v1.2* gardeles, kurios padengia apie 27,000 CpG lokusų žmogaus genome. Šlapimo pūslės vėžio eksperimente 223 mėginiai sudaro sergančiųjų grupę, 205 mėginiai – kontrolinę grupę. HNSCC eksperimente 92 mėginiai sudaro sergančiųjų ir 92 – kontrolinę grupes. Paskutiniajame kiaušidžių vėžio eksperimente 266 pacientų mėginiai sudaro sergančiųjų, 274 – kontrolinę grupes.

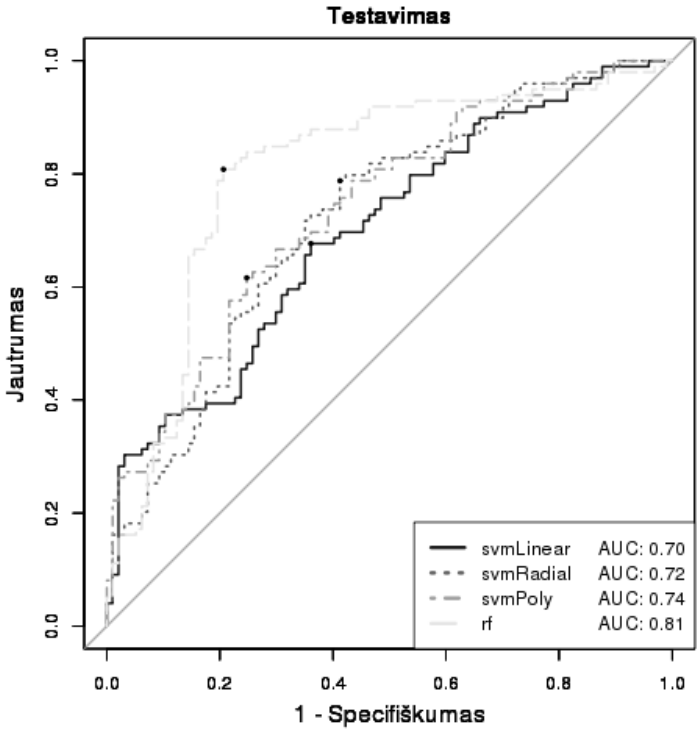
Duomenų aibė susideda iš trijų lentelių: *key* – informacijos apie mėginius, *gmap* – informacijos apie požymius, įvairias žmogaus genomo vietas, ir *beta* verčių matricos. *beta* vertė tam tikro mėginio vienam požymiui – pagal formulę gaunamas įvertis:

$$beta = M/(M + U + alpha), 0 \leq beta \leq 1 \tag{34}$$

M ir U – metilinto ir nemetilinto signalų intensyvumai. *alpha* vertė (naudota 100) skirta

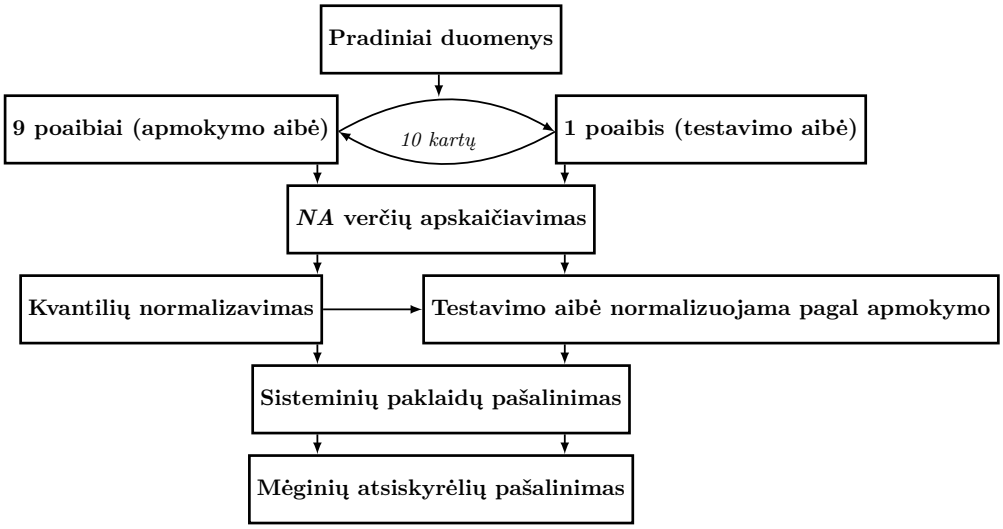


(a)



(b)

18 pav.: Po apmokymo (a) ir testavimo (b) gautos „ROC" kreivės atraminių vektorių mašinų ir atsitiktinių miškų klasifikatoriams, kai požymiai tiriant cirkuliuojančios DNR metilinimo duomenis buvo atrinkti imant vidutinis $k = 340$ statistiškai reikšmingiausių požymių, tada imant vidutinis $l = 25$ pirmųjų principinių komponenčių.



19 pav.: Duomenų apdorojimo schema.

stabilizuoti *beta* vertėms.

Siekiant įvertinti klasifikatoriaus patikimumą atliekama 10 kartų kryžminė validacija, kurios metu visa duomenų aibė padalinama į 10 dalių. Devynios naudojamos apmokymui, 1 dalis paliekama testavimui.

5.2.1 Duomenų tyrimo seka

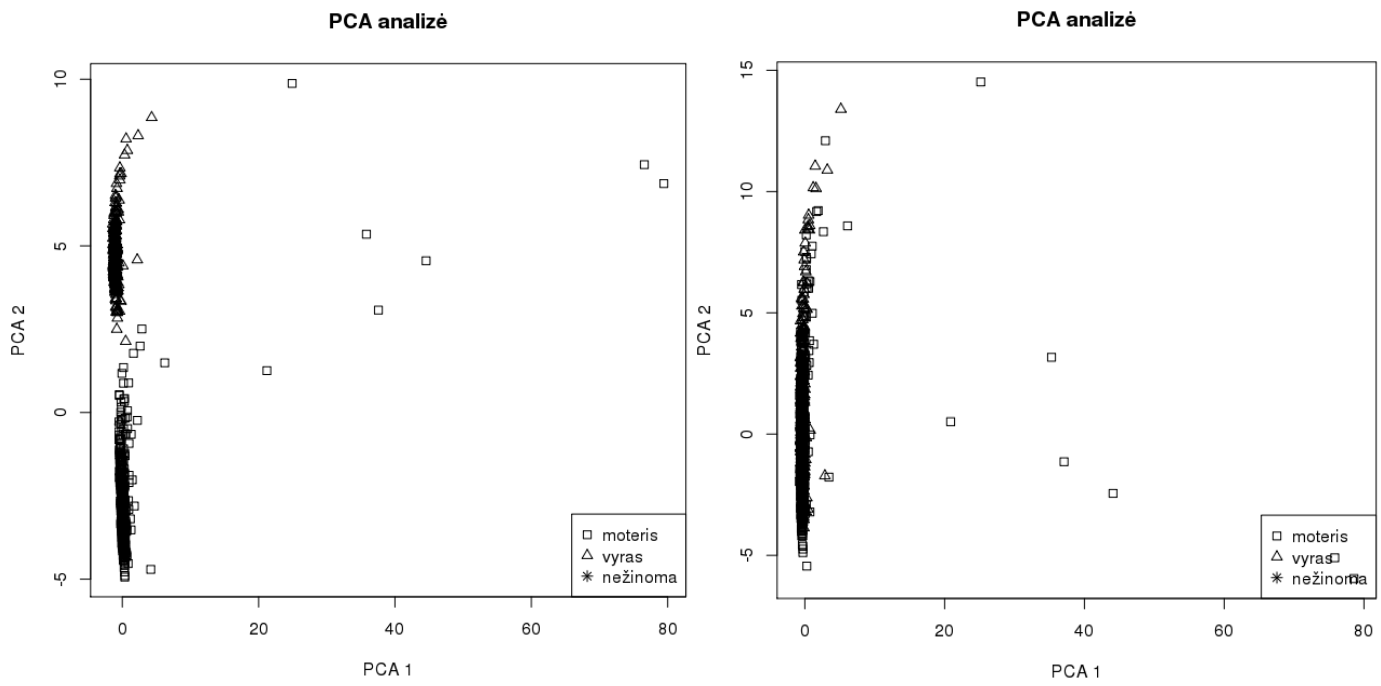
Pirmiausia apibrėžiama duomenų apdorojimo strategija leukocitų DNR metilinimo duomenims tirti Pav. 19.

Visų trijų eksperimentų duomenys apjungiami į vieną duomenų aibę (Pav. 19 „Pradiniai duomenys"). Vykdoma 10 kartų kryžminė validacija, kurios metu 9 visos duomenų aibės poaibiai naudojami apmokymui (apmokymo aibė), likęs poaibis – testavimui (testavimo aibė). Pirmiausia apskaičiuojamos duomenyse esančios tuščios *NA* vertės sudarytoms apmokymo ir testavimo aibėms atskirai. Tada duomenys normalizuojami. Sekančiame žingsnyje iš duomenų pašalinamas sisteminis variabilumas – ryškus lyties efektas, kai mėginiai geriausiai atskiriami pagal lytį, o ne pagal vėžio formą ar diagnozę (sveikas/sergantis). Pav. 20 pavaizduota kaip duomenys atrodo prieš (a) ir po (b) lyties efekto pašalinimo. Lyties efektas panaikinamas pašalinant požymius susijusius su lytinėmis chromosomis „X" ir „Y". Tuomet iš duomenų pašalinami mėginiai atsiskyrėliai.

Atlikus duomenų apdorojimą prasideda jų analizė. Analizės metu mažinamas požymių skaičius. Atrinkus požymius konstruojamas klasifikatorius naudojant pasirinktus mašininio mokymosi algoritmus. Išbandžius keletą mašininio mokymosi algoritmų gauti rezultatai palyginami, sprendžiama, kuris klasifikatorius tinkamiausias, ar įmanoma duomenis klasifikuoti.

5.2.2 Analizė su PCA ir LDA

Po duomenų apdorojimo konstruojamas klasifikatorius: atliekama principinių komponenčių analizė, imamas tam tikras principinių komponenčių skaičius, gaunamos projekcijos



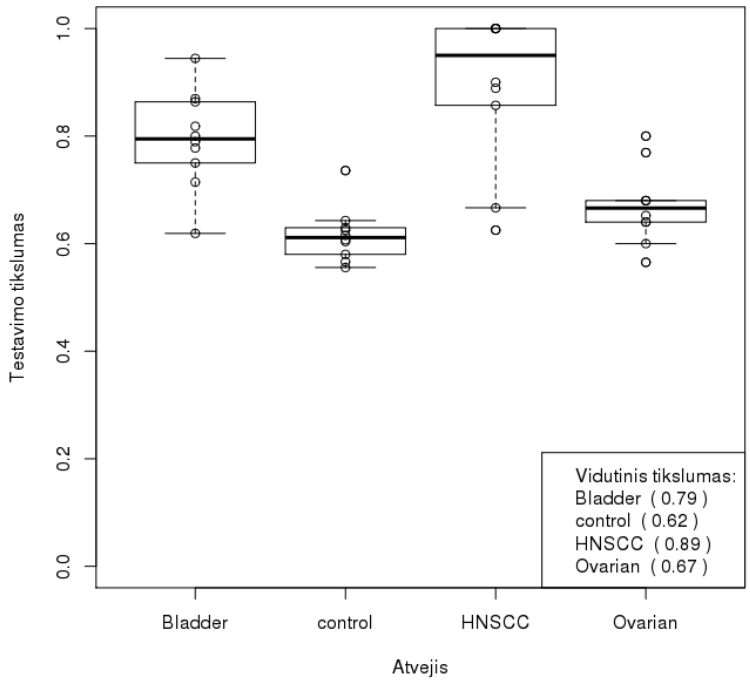
(a) (b)

20 pav.: Pirmosios dvi principinės komponentės prieš (a) ir po (b) su lytimi susijusių požymių pašalinimo.

ir atliekama tiesinė diskriminantinė analizė (angl. *Linear discriminant analysis (LDA)*). Duomenų analizės schema pavaizduota Pav. 23. Apdorojus apmokymo ir testavimo aibės duomenis jie analizuojami. Vykdoma dar viena 10 kartų kryžminė validacija siekiant išmokti modelio parametrą: išsiaiškinti kiek pirmųjų principinių komponenčių naudoti geriausia. Antrosios 10 kartų kryžminės validacijos metu apmokymui skirtas duomenų poaibis iš pirmosios 10 kartų kryžminės validacijos padalinamas į 10 dalių. Su 9 dalimis imant skirtingą principinių komponenčių skaičių gaunamos projekcijos, tada atliekama LDA, gautas LDA klasifikatorius testuojamas su likusia 1 dalimi. Po vienos iteracijos grąžinamas principinių komponenčių skaičius, su kuriuo gautas geriausias testavimo tikslumas. Po 10 tokių antrosios kryžminės validacijos iteracijų, gaunama 10 grąžintų principinių komponenčių skaičių, su kuriais gauti geriausi testavimo rezultatai. Išvedamas jų vidurkis. Galutinis klasifikatorius konstruojamas imant visą apmokymui skirtą duomenų poaibį iš pirmosios 10 kartų kryžminės validacijos, atliekama principinių komponenčių analizė ir imamas prieš tai išmoktas pirmųjų principinių komponenčių skaičius (dešimties grąžintų principinių komponenčių skaičių vidurkis). Gavus projekcijas konstruojamas LDA klasifikatorius, kuris testuojamas su testinio poaibio atitinkamomis projekcijomis.

Gautas vidutinis apmokymo tikslumas 0.81, o testavimo 0.68. Pav. 21 pavaizduotas testavimo tikslumas kiekvienai klasei atskirai. Per 10 iteracijų sukonstruoti klasifikatoriai vidutiniškai geriausiai klasifikavo šlapimo pūslės vėžio ir HNSCC vėžio atvejus, prasčiau atskirdavo kontrolę ir kiaušidžių vėžio atvejus.

Kadangi duomenys sudaryti iš trijų skirtingų eksperimentų, manyta, kad klasifikatorius išmoksta ne skirtumus tarp klasių (vėžio atvejų ir kontrolės), bet skirtumus tarp skirtingų



21 pav.: Testavimo rezultatai per visas 10 iteracijų kiekvienai klasei atskirai.

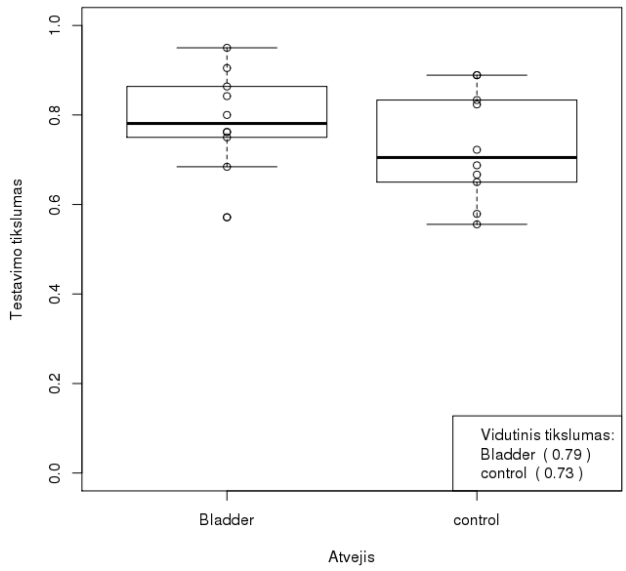
eksperimentų. Todėl prieš tai aprašyta analizė pakartota kiekvieno eksperimento duomenims atskirai ir pažiūrėta, ar įmanoma atskirti vėžio atvejus nuo kontrolės. Pav. 22 pavaizduoti gauti rezultatai. Analizuotant šlapimo pūslės vėžio eksperimento duomenis (a) gauti geriausi rezultatai, šiuose duomenyse geriausiai atskiriami vėžio atvejai ir kontroliniai mėginiai. Kitų eksperimentų vidutinis vėžio atvejų klasifikavimo tikslumas panašus, tačiau yra didesnis testavimo tikslumo variabilumas, tai reiškia, kad vienos kryžminės validacijos iteracijos metu vėžio atvejai klasifikatoriaus buvo priskirti teisingai klasei didesniu tikslumu nei kitos iteracijos metu.

5.2.3 Požymių atranka naudojant atsitiktinių miškų metodą

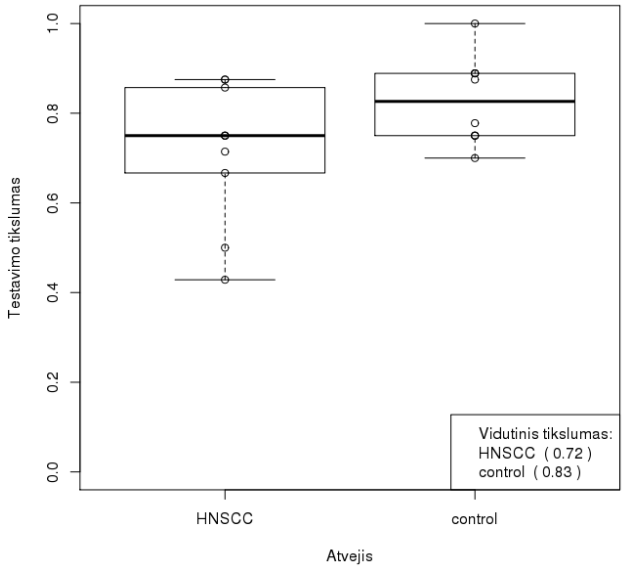
Duomenų analizėje išbandoma kita strategija. Požymių aibei sumažinti panaudojamas kitas metodas, ne PCA analizė, o požymiai atrenkami panaudojant požymių atrinkimą kartu su klasifikatoriumi. Pav. 25 kairėje veiksmai atliekami su apmokymo aibe, dešinėje su testavimo aibe. Prieš tai duomenys apdorojami.

Požymiai atrenkami naudojant požymių atrinkimą kartu su klasifikatoriumi – atsitiktinių miškų metodu [UdA06]. Apmokymui ir testavimui imami tik atrinkti požymiai. Tada dar kartą apmokomi atsitiktiniai miškai, atraminių vektorių mašinos ir LDA.

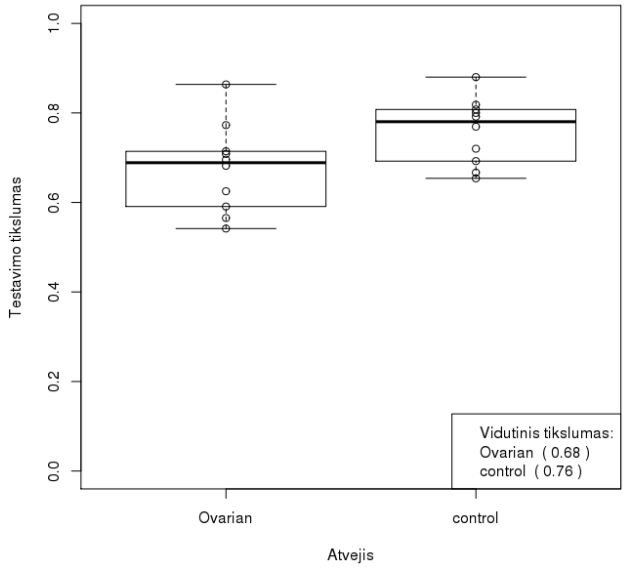
Gauti rezultatai pavaizduoti Pav. 24. Vidutiniškai geriausiai testavo atsitiktinių miškų klasifikatorius 0.69. Gauti panašūs rezultatai kaip ir po PCA – LDA analizės, kai bendras testavimo tikslumas 0.68.



(a)

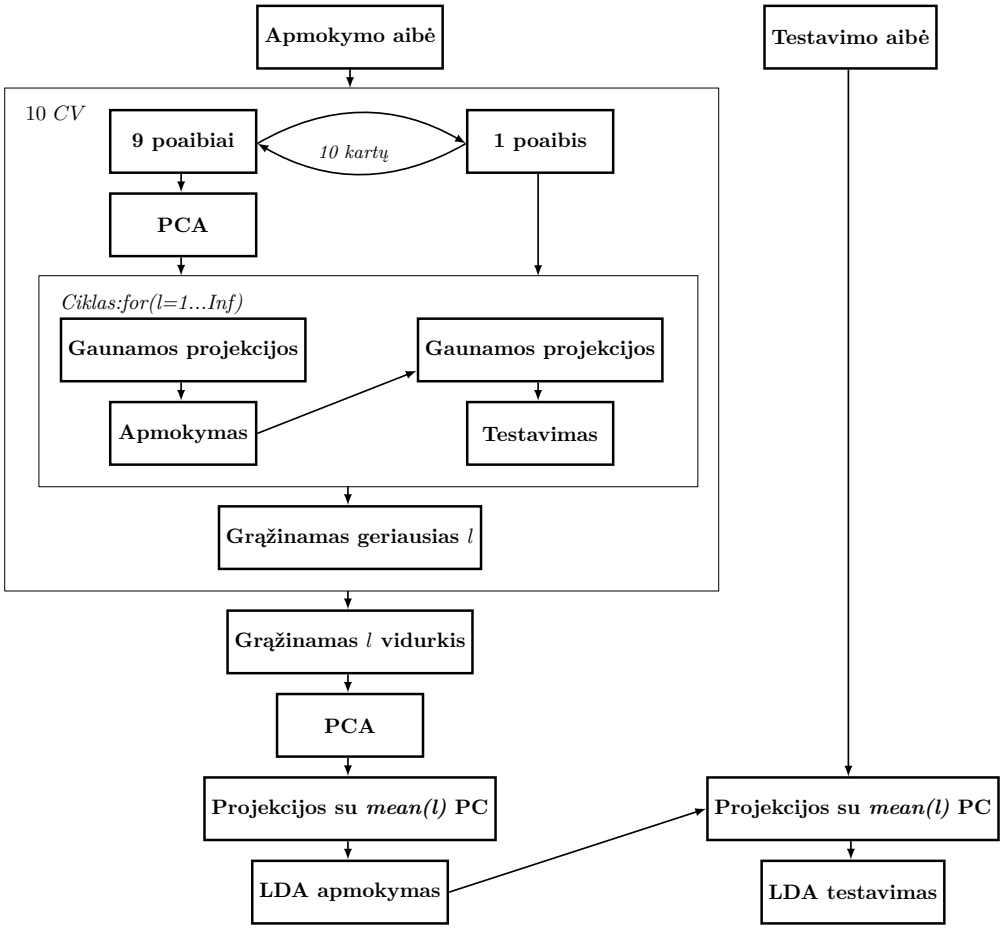


(b)

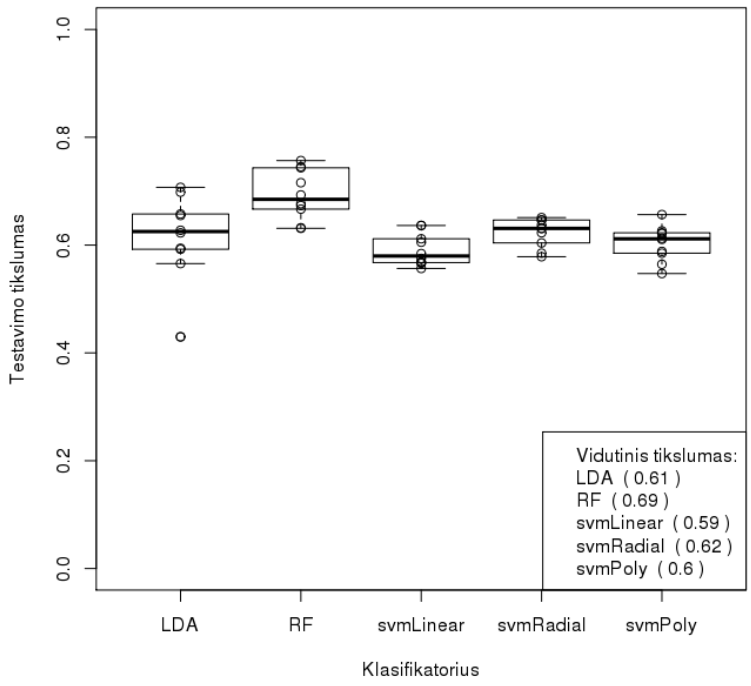


(c)

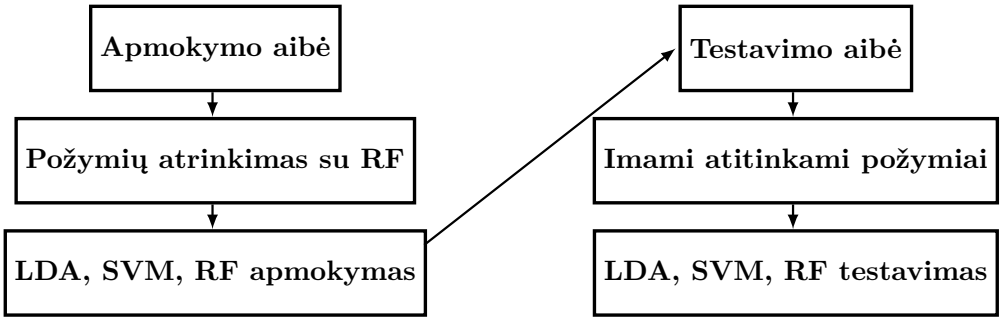
22 pav.: Siekiant įsitikinti, ar tikrai iš kraujo ląstelių DNR metilinimo galima nustatyti vėžį, prieš tai aprašyta analizė su PCA ir LDA pakartota kiekvieno eksperimento duomenims atskirai. (a) tirti šlapimo pūslės vėžio atvejai, (b) HNSCC atvejai ir (c) kiaušidžių vėžio atvejai.



23 pav.: Duomenų analizės schema. Dešimties kartų kryžminės validacijos metu išmokstamas modelio parametras, kiek geriausia naudoti principinių komponenčių. Tada naudojant apmokymo aibės duomenis atliekama PCA, imamas išmoktas principinių komponenčių skaičius. Gavus projekcijas konstruojamas LDA klasifikatorius, kuris testuojamas su testine aibe.



24 pav.: LDA, atsitiktinių miškų ir atraminių vektorių mašinų testavimo rezultatai. Testavimo tikslumo pasiskirstymai kiekvienam klasifikatoriui po 10 kryžminės validacijos iteracijų.



25 pav.: Duomenų analizės schema. Atrenkami požymiai tuomet apmokomi ir testuojami LDA, SVM ir RF klasifikatoriai.

Išvados

Literatūros dalyje išsiaiškinus keletą mašininio mokymosi algoritmų, duomenų analizės būdų ir juos pritaikius tiriant tiesiosios žarnos vėžio cirDNR metilinimo duomenis ir leukocitų DNR metilinimo duomenis gautus iš trijų skirtingų eksperimentų prieita keletas išvadų:

1. Tiriant cirDNR metilinimo duomenis geriausi rezultatai gauti imant išmoktą skaičių statistiškai reikšmingiausių požymių, pritaikius PCA, imant išmoktą pirmųjų PC skaičių ir apmokius atsitiktinių miškų klasifikatorių. Gautas testavimo tikslumas 0.79 rodo, kad cirDNR metilinimo duomenys galėtų būti naudojami vėžio diagnostikoje.

2. Geriausi rezultatai analizuojant leukocitų DNR metilinimo duomenis gauti po duomenų apdorojimo pritaikius PCA, ėmus išmoktą pirmųjų PC skaičių ir pritaikius LDA klasifikatorių. Gautas klasifikatorius atskyrė įvairius vėžio atvejus ir kontrolę su vidutiniu 0.7 testavimo tikslumu.

3. Tiriant skirtingų eksperimentų leukocitų DNR metilinimo duomenis, pastebėta, kad šlapimo pūslės vėžio duomenyse signalas stipresnis (geriau atskiriami vėžio atvejai nuo kontrolės), nei HNSCC ir kiaušidžių vėžio duomenyse.

Hipotezę, jog cirDNR metilinimo duomenyse esantis signalas yra ne tik iš vėžinių ląstelių DNR, bet ir iš žuvusių kraujo ląstelių DNR, patikrinti būtų galima atliekant tolimesnius tyrimus.

Literatūra

[ABR64] Mark A. Aizerman, Emmanuel M. Braverman, and Lev I. Rozonoer. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[AG97] Yali Amit and Donald Geman. Shape quantization and recognition with randomized trees. *Neural Computation*, 9(7):1545–1588, 1997.

[ANKA14] Arpit Agarwal, Harikrishna Narasimhan, Shivaram Kalyanakrishnan, and Shivani Agarwal. GEV-Canonical Regression for Accurate Binary Class Probability Estimation when One Class is Rare. *International Conference on Machine Learning*, 32, 2014.

[BGV92] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik. A Training Algorithm for Optimal Margin Classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pages 144–152, New York, NY, USA, 1992. ACM.

[BIA03] B. M. Bolstad, R. A. Irizarry, and M. Astrand. A comparison of normalization methods for high density oligonucleotide array data based on variance and bias. *Bioinformatics*, 19:185–193, 2003.

[Bre01] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[CB09] Howard Cedar and Yehudit Bergman. Linking DNA methylation and histone modification: patterns and paradigm. *Nature Reviews Genetics*, 10:295–304, May 2009.

[CGB+11] Chao Chen, Kay Grennan, Judith Badner, Dandan Zhang, Elliot Gershon, Li Jin, and Chunyu Liu. Removing Batch Effects in Analysis of Expression Microarray Data An Evaluation of Six Batch Adjustment Methods. *PLoS*, 19, 2011.

[CKL+12] Rene Cortese, Andrew Kwan, Emilie Lalonde, Olga Bryzgunova, Anna Bondar, Ying Wu, Juozas Gordevicius, Mina Park, Gabriel Oh, Zachary Kaminsky, Justina Tverkuviene, Arvydas Laurinavicius, Feliksas Jankevicius, Dorota H.S. Sendorek, Syed Haider, Sun-Chong Wang, Sonata Jarmalaite, Pavel Laktionov, Paul C. Boutros, and Arturas Petronis. Epigenetic markers of prostate cancer in plasma circulating DNA. *Human Molecular Genetics*, 21(16):3619–3631, 2012.

[Est07] M Esteller. Cancer epigenomics: DNA methylomes and histone-modification maps. *Nat Rev Genet*, 8(4):286–98, April 2007.

[Fle09] Tristan Fletcher. *Support Vector Machines Explained*. 2009.

[GBD92] Stuart Geman, E. Bienenstock, and R. Doursat. Neural networks and the biasvariance dilemma. *Neural Computation*, 4:1–58, 1992.

[GMT13] R. Gaidelytė, N. Madeikytė, and D. Tendziagolskytė. Lietuvos sveikatos statistika 2012. 2013.

[GT09] Hinrich Gohlmann and Willem Talloen. *Gene Expression Studies Using Affymetrix Microarrays*. Chapman & HallCRC, New York, NY, USA, 2009.

[Ho95] Tin Kam Ho. Random decision forest. *Proceedings of the 3rd International Conference on Document Analysis and Recognition, Montreal, QC*, 45(1):278–282, August 1995.

[Ho98] Tin Kam Ho. The random subspace method for constructing decision forests. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(8):832–844, 1998.

[Hot33] H. Hotelling. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology*, (24):417–441,498–520, 1933.

[HTF09] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning. Data Mining, Inference, and Prediction (Second Edition)*. Springer, 2009.

[JWHT13] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An Introduction to Statistical Learning with Applications in R*. Springer Texts in Statistics, 2013.

[KLS10] Myoung Sook Kim, Juna Lee, and David Sidransky. DNA methylation markers in Colorectal cancer. *Cancer Metastasis Rev*, 29:181–206, 2010.

[Kuh14] Max Kuhn. The caret Package. [Žiūrėta 2015 04 26] Prieiga per internetą: `http://topepo.github.io/caret/index.html`, 2014.

[Lav14] Victor Lavrenko. Decision tree, video lectures. [Žiūrėta 2015 01 19] Prieiga per internetą: `https://www.youtube.com/playlist?list=PLBv09BD7ez_4temBw7vLA19p3tdQH6FYO`, 2014.

[LHA+14] Scott M Langevin, E Andres Houseman, William P Accomando, Devin C Koestler, Brock C Christensen, Heather H Nelson, Margaret R Karagas, Carmen J Marsit, John K Wiencke, and Karl T Kelsey. Leukocyte-adjusted epigenome-wide association studies of blood from solid tumor patients. *Epigenetics*, 9(6):884–895, 2014.

[LS07] JT. Leek and JD. Storey. Capturing heterogeneity in gene expression studies by surrogate variable analysis. *PLoS Genetics*, 3, 2007.

[NW72] John Nelder and Robert Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society*, 135(3):370–384, 1972.

[Pea01] K. Pearson. On Lines and Planes of Closest Fit to Systems of Points in Space. *Philosophical Magazine*, 11(2):559–572, 1901.

[PH13] Leek JT Parker HS, Bravo HC. Removing batch effects for prediction problems with frozen surrogate variable analysis. arXiv:1301.3947, 2013.

[Qui86] J R Quinlan. Induction of Decision Trees. *Mach. Learn.*, 1(1):81–106, 1986.

[SLH13] Lin Song, Peter Langfelder, and Steve Horvath. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC Bioinformatics*, pages 1471–2105, 2013.

[SOST14] Shinya Suzumura, Kohei Ogawa, Masashi Sugiyama, and Ichiro Takeuchi. Outlier Path: A Homotopy Algorithm for Robust SVM. *International Conference on Machine Learning*, 32, 2014.

[ST14] Friedhelm Schwenker and Edmondo Trentin. Pattern classification and clustering: A review of partially supervised learning approaches. *Pattern Recognition Letters*, 37:4–14, 2014.

[UdA06] Ramón Díaz Uriarte and Sara Alvarez de Andrés. Gene selection and classification of microarray data using random forest. *BMC Bioinformatics*, 7(3):1471–2105, 2006.

[Vap95] Vladimir N. Vapnik. *The Nature of Statistical Learning Theory*. Springer-Verlag New York, Inc., New York, NY, USA, 1995.

[WE05] Gregory Grant Warren Ewens. *Statisical Methods in Bioinformatics: An Introduction (Second Edition)*. Springer, 2005.

[WHO14] World Health Organization. Cancer. [Žiūrėta 2015 01 19] Prieiga per internetą: `http://www.who.int/mediacentre/factsheets/fs297/en/`, 2014.

[WS15] Kristina Warton and Goli Samimi. Methylation of cell-free circulating DNA in the diagnosis of cancer. *Frontiers in Molecular Biosciences*, 2(13), 2015.

[WYG14] Junfeng Wen, Chun-Nam Yu, and Russell Greiner. Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification. *International Conference on Machine Learning*, 32, 2014.

[YZW+14] Jinfeng Yi, Lijun Zhang, Jun Wang, Rong Jin, and Anil K. Jain. A Single-Pass Algorithm for Efficiently Recovering Sparse Cluster Centers of High-dimensional Data. *International Conference on Machine Learning*, 32, 2014.