Kauno technologijos universitetas

Informatikos fakultetas

Lapų atspindžio spektrų požymių atrankos tyrimas automatiniam trąšų poreikio nustatymui

Baigiamasis magistro studijų projektas

Paulius Mykolaitis

Projekto autorius

doc. dr. Mantas Lukoševičius

Vadovas

Kaunas, 2024

Kauno technologijos universitetas

Informatikos fakultetas

Lapų atspindžio spektrų požymių atrankos tyrimas automatiniam trąšų poreikio nustatymui

Baigiamasis magistro studijų projektas

Programų sistemos (6211BX011)

Paulius Mykolaitis

Projekto autorius

doc. dr. Mantas Lukoševičius

Vadovas

doc. dr. Šarūnas Packevičius

Recenzentas

Kaunas, 2024

Kauno technologijos universitetas

Informatikos fakultetas

Paulius Mykolaitis

Lapų atspindžio spektrų požymių atrankos tyrimas automatiniam trąšų poreikio nustatymui

Akademinio sąžiningumo deklaracija

Patvirtinu, kad mano, Pauliaus Mykolaičio, baigiamasis projektas tema „Lapų atspindžio spektrų požymių atrankos tyrimas automatiniam trąšų poreikio nustatymui“ yra parašytas visiškai savarankiškai ir visi pateikti duomenys ar tyrimų rezultatai yra teisingi ir gauti sąžiningai. Šiame darbe nei viena dalis nėra plagijuota nuo jokių spausdintinių ar internetinių šaltinių, visos kitų šaltinių tiesioginės ir netiesioginės citatos nurodytos literatūros nuorodose. Įstatymų nenumatytų piniginių sumų už šį darbą niekam nesu mokėjęs.

Aš suprantu, kad išaiškėjus nesąžiningumo faktui, man bus taikomos nuobaudos, remiantis Kauno technologijos universitete galiojančia tvarka.

(vardą ir pavardę įrašyti ranka)	(parašas)

Dokumentą elektroniniu parašu
pasirašė PAULIUS,MYKOLAITIS
Data: 2024-05-24 09:04:43

Paulius Mykolaitis. Lapų atspindžio spektrų požymių atrankos tyrimas automatiniam trąšų poreikio nustatymui. Magistro projektinio darbo vadovas doc. dr. Mantas Lukoševičius. Kauno technologijos universitetas, Informatikos fakultetas.

Studijų kryptis ir sritis (studijų krypčių grupė): Programų sistemos; Mašininis mokymas

Raktiniai žodžiai: atraminių vektorių mašina, artimiausių kaimynų algoritmas, požymių išskyrimas, atspindžio spektroskopija, mitybos elementų trūkumai javuose

Kaunas, 2024. 49 p.p.

Santrauka

Pagrindinis darbo tikslas – tarpusavyje palyginti skirtingus požymių išskyrimo metodus atraminių vektorių klasifikavimo algoritmo kontekste ir išskirti svarbiausius požymius, kurie vėliau galėtų būti nagrinėjami srities ekspertų siekiant geriau pažinti augalų sandaros procesus mitybos elementų trūkumo aplinkoje. Darbe yra pateikta augalų mitybos elementų trūkumo atpažinimo, klasifikavimo algoritmų (k-NN, SVM, Random Forest) bei požymių išskyrimo metodų (Laplace, Fisher, ReliefF, χ^2) literatūros apžvalga.

Paulius Mykolaitis. Feature Selection Analysis of Crop Leaf Reflectance Spectra for Automated Nutrient Deficiency Detection. Master's thesis project supervisor Assoc. Prof. Dr. Mantas Lukoševičius; Faculty of Informatics, Kaunas University of Technology.

Research domain: Software engineering; Machine Learning

Keywords: SVM, Support Vector Machines, k-NN, Nearest Neighbor algorithm, feature selection, reflection spectroscopy, crops nutrient deficiency.

Kaunas, 2024. 49 p.p.

Abstract

A literature overview of plant nutrient deficiency diagnostic methods, classification algorithms (k-NN, SVM, Random Forest), and feature selection methods (Laplace, Fisher, ReliefF, χ^2) are provided in the thesis. The main objective of the study is to compare different feature selection methods within the context of Support Vector Machine (SVM) classification algorithm and identify the most important features that can be further analyzed by domain experts in order to better understand the plant physiological processes in plants with nutrient deficiencies.

Turinys

Lentelių sąrašas

Iliustracijų sąrašas

Algoritmų sąrašas

Įžanga

Tikslas ir uždaviniai

Magistrinio darbo projekto tikslas: sukurti ir pritaikyti algoritmus, kurie galėtų būti naudojami inžinerinėje sistemoje, bei sistemos naudotojui pateiktų rekomendacijas augalų mitybos elementų trūkumui mažinti.

Šiame darbe yra analizuojami savybių atrankos metodai radialinių bazinių funkcijų atraminių vektorių, bei artimiausių kaimynų klasifikavimo metodams. Tam tikslui buvo tarpusavyje lyginami klasifikavimo modelio rezultatai panaudojant χ^2, Laplaso, Fišerio, ReliefF filtravimo metodus.

1. Analitinė dalis

1.1. Dalykinė sritis

Tiriamasis darbas yra susijęs su klasifikavimo algoritmų kūrimu analitinėje ekspertinėje sistemoje. Ekspertinė sistema nustato mikroelementų balansą javuose ir teikia rekomendacijas sistemos paslaugų naudotojui, pateikia galimas cheminių preparatų rūšis kompensuosiančias mitybos medžiagų trūkumą pasėliams.

Augalų ligos gali būti diagnozuojamos paprasčiausiai apžiūrint augalus, atliekant tyrimus laboratorijoje (fiziologiniai, biologiniai, serotologiniai ir molekuliniai testai) arba naudojant neinvazinius tyrimo metodus: regimosios šviesos ar infraraudonųjų spindulių spektroskopija (atspindžio, pralaidumo, fluorescensinė, vibracinė, Ramano ir t.t.) arba fotografiniai metodai (augalų audinių fotografijos, aerofotografavimas, multispektrinės nuotraukos.)

Mikro- bei makroelementai skirtingu augalo vegetacijos laikotarpiu taip pat pasižymi skirtingu mobilumu augalo audiniuose. Esant tam tikrų mikroelementų trūkumui, naujuose augalo stiebuose mitybos elementų stygius gali būti a) balansuojamas cheminius elementus augalui perkeliant iš senesnių dalių b) iš dalies perkeliant c) cheminiai elementai gali išlikti nemobilūs senesniuose augalo lapuose.

Nepaisant aplinkos sąlygų, kuriose tiriamas augalas buvo užaugintas, šviesos spektrinis atsakas tarp senų ir naujų augalo stiebų skiriasi, dėl skirtingo fotosintezės proceso intensyvumo, kadangi jaunesniuose augalo stiebuose fotosintezę vykdantys audiniai turi mažiau pigmentų. Todėl kontrolės matavimai yra atliekami skirtingais augimo tarpsniais, skirtingiems augalo stiebams.

1.1.1. Multispektrinės fotografijos metodas

Multispektrinės fotografijos paprastai yra pasėlių ploto aerofotografijos, kurios, priešingai nei RGB fototografijos, pasėlių plotą atvaizduoja įvairiose optinio bei infraraudonųjų spindulių bangos ilgio juostose. Skirtingoms dažnių juostoms atvaizduoti naudojami skirtingi aerofotografijos kanalai (išsaugojamos skirtingos pikselių vertės, įvairiems šviesos bangų ilgiams.)

Vegetacijos indeksai naudojami kartu su multispektrinėmis fotografijomis yra tam tikras išvestinis dydis, priklausantis nuo multispektrinės fotografijos duomenų (kurie yra susiję su konkrečiais bangos ilgiais.) Kiekvienas vegetacijos indeksas palengvina augalų būsenos įvertinimą tam tikru analizės aspektu, kadangi yra pritaikytas tam tikriems augalų biomasės statistiniams rodikliams matuoti, pavyzdžiui, dažniausiai naudojamas normalizuoto skirtumo vegetacijos indeksas (NDVI) padeda įvertinti augalų biomasės tankį ar fotosintezės intensyvumą. Vegetacijos indeksai yra spektrinių duomenų interpretavimo euristikos, kurios nustato spektrinių duomenų išvestinį dydį, paprastai koreliuojantį su kitų matavimų duomenimis, pavyzdžiui: dirvožemio sausringumu, augalų rūšimi, augimo tarpsniu, chemine sandara [3].

Tačiau nėra naudojami vien koreliacijos skaičiavimai vegetacijos indeksui įvertinti. Lapų ploto indeksas gali būti apskaičiuotas tiesiogiai iš skaitmeninių fotografijų, panaudojant mašininio mokymosi algoritmus (arba kitus fotogrametrijos metodus). Šie duomenys toliau naudojami mašininio mokymosi algoritmo, nustatančio sąryšį tarp multispektrinės nuotraukos ir apskaičiuoto

lapų ploto, kurį dengia multispektrinė nuotrauka [3].

Daugelis vegetacijos indeksų sukuriami siekiant tiksliau įvertinti ne vien procesus, statistiškai modeliuojamus tam tikrame augalų plote, bet ir matavimų paklaidas, dėl atmosferoje vykstančių meteorologinių bei termodinaminių reiškinių [3].

Regimosios ir artimųjų bei vidutiniųjų infraraudonųjų spindulių šviesos spektro informacija multispektrinėje fotografijoje yra iš esmės saulės šviesos atspindžio spektras (šviesa yra atspindima nuo augalų lapų paviršiaus,) kuris gali charakterizuoti atspindinčio paviršiaus cheminę sandarą. Tuo tarpu, tolimųjų infraraudonųjų spindulių spektras yra augalų šiluminės spinduliuotės spektras, kuriuo remdamiesi, galime įvertinti termoreguliacinių procesų savybes: augalų transpiraciją, hidratacijos lygį [3].

Pagrindiniai multispektrinės fotografijos metodo trūkumai yra per maža tiek spektrinė (per mažai dažnio juostų, juostos yra per plačios), tiek erdvinė matavimo raiška (kadangi kiekviename fotografjos pikselyje yra užfiksuojamas vidutinis atspindžio spektras, kurį gali sudaryti skirtingų augalų rūšių, dirvožemio, ir t.t. atspindžio spektrai [3].)

1.1.2. Multispektrinės fotografijos metodas tiriamajame darbe

Tiriamajame darbe naudojamą prietaisą sudaro diodo šviesos šaltinis, regimosios šviesos interferometras, bei aktyviojo pikselio jutiklių matrica, kurioje augalo atspindėtos šviesos spektro fotonai paverčiami į elektrinį signalą, kuris analoginio skaitmeninio jutiklio pagalba yra registruojamas prietaiso mikrovaldiklyje.

Skirtingi mitybos elementai turi skirtingą mobilumą augalo audiniuose. Mikroelementų trūkumas sukelia lapo sandaros pokyčius, kuriuos srities ekspertai gali vizualiai nustatyti lygindami skirtingo amžiaus lapų požymius. Tiriamojo darbo metodas pateikia rekomendacijas registruoti keletą, skirtingo amžiaus, lapų matavimų. Surinktiems keleto lapų spektriniams matavimams yra atliekama klaidų analizė, nulio hipotezių testavimas. Sekančio etapo metu yra pritaikomi mašininio mokymosi klasifikavimo algoritmai. Tyrimo metu atsižvelgiama į įvairių aplinkos veiksnių poveikį augalų spektrų matavimams, bei papildomus matavimo kintamuosius, tokius kaip: augalo rūšis, veislė, augimo tarpsnis, drėgnumo sąlygos, dirvožemis, aplinkos temperatūra, matavimo paros laikas, cheminių preparatų naudojimas.

1.2. Trąšų trūkumo nustatymas pagal morfologinius augalo simptomus

Mikroelementų trūkumas pasėliuose gali būti nustatomas stebint augalo sandaros pakitimus (spalvų, tekstūros, augimo anomalijos, vytimas ir t.t.), kurie paprastai būna būdingi tam tikriems pasėlių ligų požymiams (mikroelementų trūkumui, pertekliui, stresui dėl aplinkos sąlygų ar kenkėjų poveikiui.) Priešingai nei makroelementai, mikroelementai yra cheminiai elementai kurių santykinai nedidelis kiekis reikalingas organizmo optimaliam vystymuisi. Požymiai, būdingi vienam augalo kultivarui, dėl tam tikro mitybos elemento trūkumo, kitoms augalo veislėms gali pasireikšti skirtingai. Mitybos elementų trūkumas ar perteklius sutrikdo augalo medžiagų apytakos sistemą [4].

Apžiūros metodas remiasi Lybicho minimumo dėsnio (angl. *Liebig law of the minimum*) prielaida: jei nepaisoma klimato, dirvožemio ir t.t. sąlygų tik optimalus mikroelementų kiekis leis

pasiekti didžiausią įmanomą javų derlių. Jei bent vieno mikroelemento trūksta, derlius žymiai sumažėja [5].

Kviečiui augti yra reikalinga 14 mitybos elementų (N, P, K, Ca, Mg, S, Fe, Mn, B, Zn, Cu, Mo, Ni, Cl; kitoms javų kultūroms – apie 10) tam, kad augimo sąlygos būtų optimalios. Kiti cheminiai elementai gali turėti toksinį poveikį augalui, pavyzdžiui: Al, Se, Cd,– nėra susiję su jokiais medžiagų apytakos procesais augale ir yra taršos elementai, inhibitoriai, bloginantys kitų mikroelementų biocheminius procesus.

1.3. Mašininio mokymosi algoritmai

Mašininio mokymosi algoritmai yra taikomosios statistikos algoritmai, nustatantys statistinius duomenų sąryšius, tačiau paprastai neįvertinantys pasikliovimo intervalų nustatytiems funkciniams sąryšiams, kurie yra dažniau naudojami kitose statistikos šakose [6].

1.1 lentelė. Mašininio mokymosi algoritmų pritaikymo galimybės [2]

Algoritmų klasė / **Savybė**	**Neuroniniai tinklai**	**SVM**[1]	**Sprendimų medžiai**	**MARS**[2]	**k-NN**
Kategorinių bei tolydžiųjų kintamųjų atskyrimas	Reikalingas	Reikalingas	Nereikalingas	Nereikalingas	Reikalingas
Papildomi skaičiavimai trūkstamoms požymių vektorių ypatybėms nustatyti	Reikalingi	Reikalingi	Nereikalingi	Nereikalingi	Nereikalingi
Atsparumas požymių vektorių riktams (angl. *outliers*)	Mažesnis	Mažesnis	Didesnis	Mažesnis	Didesnis
Jautrumas monotoninėms požymių vektorių transformacijoms	Didesnis	Didesnis	Mažesnis	Didesnis	Didesnis

Algoritmų klasė / Savybė	Neuroniniai tinklai	SVM[1]	Sprendimų medžiai	MARS[2]	k-NN
Skaičiavimų kompleksiškumas, didinant požymių vektorių skaičių	Prastesnis	Prastesnis	Geresnis	Geresnis	Prastesnis
Svarbiausių ypatybių išskyrimas	Prastesnis	Prastesnis	Geresnis	Geresnis	Prastesnis
Priklausomybių tarp požymių vektorių sudarymas	Geresnis	Geresnis	Prastesnis	Prastesnis	Prastas
Modelio suprantamumas	Prastesnis	Prastesnis	Geresnis	Geresnis	Prastesnis
Mokymosi geba	Didesnė	Didesnė	Mažesnė	Vidutinė	Didesnė

1.3.1. Klasifikavimas ir regresija

Klasifikavimo uždavinys matematiškai yra išreiškiamas formule $f : \mathbb{R}^n \to 1, \cdots, k$, jei $y = f(\vec{x})$, ($\vec{x} - -$ požymių vektorius, angl. *feature vector*,) funkcija f kiekvienam duomenų pavyzdžiui priskiria vieną iš duomenų kategorijų k. Klasifikavimo uždavinyje mašininio mokymosi algoritmas apskaičiuoja sąryšį f tarp duomenų pavyzdžių ir jų kategorijų (arba tikimybių duomenims būti priskirtiems tam tikroms kategorijoms.)

Jei klasifikavimo uždavinyje kiekvieno duomenų pavyzdžio požymių vektorius yra tiksliai apibrėžtas (visos požymių vektoriaus komponentės yra žinomos,) klasifikavimo uždaviniui užtenka rasti vieną funkcinį sąryšį f, nustatantį priklausomybę tarp duomenų pavyzdžio $\vec{x}$ ir jo kategorijos $y \in \{1, \cdots, k\}$. Priešingu atveju, algoritmas iš pradžių turi išmokti požymių vektoriaus komponenčių tikimybės tankio funkciją ir, sekančiame etape apskaičiuoti trūkstamas vertes požymių vektoriui, prieš atlikdamas įprastą klasifikavimo uždavinį.

Regresijos užduotis yra apibendrintai išreiškiama formule: $f : \mathbb{R}^n \to \mathbb{R}$,– kiekvienam požymio vektoriui yra nuspėjama labiausiai tikėtina tolydžiojo dydžio reikšmė.

[1]SVM – atraminio vektoriaus mašinos
[2]MARS – daugiamatė adaptyvi regresijos glaudžioji kreivė

1.2 lentelė. Klasifikavimo matrica

		Klasifikatoriaus skaičiavimų rezultatas	
		Teigiama vertė	Neigiama vertė
Tikroji vertė	Teigiama vertė	Teisingai teigiama	Klaidingai neigiama
	Neigiama vertė	Klaidingai teigiama	Teisingai neigiama

Klasifikavimo bei regresijos užduotis atliekančių mašininio mokymosi algoritmų tikslumui įvertinti dažniausiai naudojama modelio tikslumo metrika (angl. *accuracy*). Tikslumo metrika – tai santykinis skaičius teisingai klasifikuojamų verčių. Pagal (1.2) lentelės duomenis, modelio tikslumas įvertinamas teisingai klasifikuojamų pavyzdžių skaičių (klasifikavimo matricos įstrižainėje) padalinant iš bendro klasifikuojamų pavyzdžių skaičiaus. Taip pat, yra naudojama atvirkštinė metrika – klaidingumas (angl. *error rate*.) Metrikos, vertinančios klasifikavimo ar regresijos modelio tinkamumą, yra sudaromos testavimo duomenų rinkiniui. Paprastai yra daroma prielaida, kad testavimo rinkinio duomenys statistiškai yra tame pačiame duomenų skirstinyje kaip ir mašininio mokymo rinkinys. Todėl tikslumo metrika yra naudojama su naujais, nežinomais duomenimis – siekiama įvertinti generalizacijos paklaidą (angl. *generalization error*) tam tikram klasifikavimo uždaviniui, darant prielaidą, kad tam uždaviniui egzistuoja tam tikras bendras įvesties duomenų pasiskirstymas [6].

Klasifikavimo uždavinys gali būti sprendžiamas panaudojant prižiūrimojo (angl. *supervised*) ar neprižiūrimojo mokymosi algoritmus. Neprižiūrimojo mokymosi algoritmai gali priklausyti ne tik nuo uždavinio pobūdžio, bet ir nuo dalykinės srities, kurioje šie algoritmai yra taikomi, kadangi vadovaujamasi euristikomis optimaliausiam sprendiniui rasti. Ieškant optimaliausio sprendinio, neprižiūrimųjų algoritmų algoritmai dažnai gali būti keičiami, modifikuojami:

- taikomos įvairios nuostolių skaičiavimo funkcijos,
- skirtingos duomenų klasterizavimo strategijos (hierarchinis, grafinis, duomenų siejimo, mišriojo funkcijų skirstinio, radialinės bazės funkcijų klasterizavimas ir t.t.,)
- naudojamos įvairios metrikos požymių vektorių atstumui nuo klasterių centrų įvertinti,
- bei skirtingos papildomos algoritmo optimizavimo modifikacijos, pavyzdžiui, k-means algoritmui [7]:

  - strategijos optimizuojančios klasterių inicijavimą (k-means++, hierarchinis, hipersferos inicijavimas ir t.t.;)
  - tiesinis klasterizavimo algoritmas arba netiesinės modifikacijos, tokios kaip, kernel

k-means, COLL modelis ir t.t.;
- klasterių glaudumo (angl. *compactness*) bei santykinės požymių vektorių komponenčių svarbos derinimas, (fuzzy k-means, SYNCLUS, Wk-means, PROCLUS, EWk-means)
- automatinis klasterių skaičiaus optimizavimas, X-means.

Neprižiūrimųjų algoritmų klasei nėra priskiriami vien klasterizavimo algoritmai. Šioje apžvalgoje, jei lygintume vien neprižiūrimuosius mokymosi algoritmus, didesnis dėmesys būtų skiriamas klasterizavimo algoritmams (jei su tam tikrais požymių vektoriais galime atlikti klasterizavimą, tai pavyks atlikti ir klasifikavimo uždavinį.)

Prižiūrimojo ir neprižiūrimojo mokymosi sąvokos neturi tikslaus matematinio apibrėžimo. Gali būti sukurti hibridiniai algoritmai, naudojantys abejas mašininio mokymosi rūšis. Jei lygintume tarpusavyje šias mokymosi rūšis abstrakčiai, kaip skaičiavimus apytiksliai įvertinančius įvesties duomenų statistinius pasiskirstymus, tai esminis skirtumas tarp prižiūrimųjų mašininio mokymosi algoritmų $\hat{y} = p(y|\vec{x})$ ir neprižiūrimųjų $\hat{y} = p(\vec{x}),$ būtų tai, kad prižiūrimieji algoritmai apytiksliai įvertina sąlyginio statistinio pasiskirstymo funkciją (priklausančią nuo $\vec{x}$) sužymėtoms duomenų vertėms (angl. *labels*,) o neprižiūrimieji algoritmai apskaičiuoja įvesties duomenų pasiskirstymo funkciją tiesiogiai [6].

1.3.1.1 K artimiausių kaimynų K artimiausių kaimynų (k-NN) mašininio mokymosi algoritmas yra prižiūrimas (patys požymių vektoriai yra naudojami kaip klasifikavimo ar regresijos uždavinio sprendiniai, jiems priskiriant tam tikras vertes y_i) ir neparametrinis (algoritmo parametrų skaičius nepriklauso nuo sprendinio; sprendinys gali būti tam tikra įvesties vektorių neparametrizuota funkcija.) Algoritmas veikia pagal principą, kad požymių vektorius galima klasifikuoti pagal panašumą, jei egzistuoja metrika, pagal kurią jie gali būti tarpusavyje palyginami. Algoritmo generalizacijos paklaida, požymių vektorių skaičiui didėjant, konverguoja į Bajeso paklaidą (angl. *Bayes error*) su sąlyga, jei klasifikavimo ar regresijos sprendiniai nustatomi su papildomu balsavimu (pavyzdžiui, naujam vektoriui priskiriama ta artimiausių kaimynų klasė, kuri yra dažniausia tarp artimiausių kaimynų [6].)

1 Algoritmas Apibendrintas k artimiausių kaimynų algoritmas [1]

Require: $\mathcal{D} = \{(\vec{x}_1, y_1), (\vec{x}_2, y_2), ..., (\vec{x}_k, y_k)\}$ – apmokymo duomenų rinkinys
Require: $z = (\vec{x}', y')$ - požymių vektorius, kuriam yra atliekamas spėjimas
1: **visiems** $(\vec{x}_i, y_i) \in \mathcal{D}$ **kartoti**
2: Suskaičiuoti atstumą $d(\vec{x}_i, \vec{x}')$ tarp z ir kortežo $(\vec{x}_i, y_i)$ požymių vektorių
3: Rasti $\mathcal{D}_z \subseteq \mathcal{D}$, pagal k mažiausias $d(\vec{x}_i, \vec{x}')$ reikšmes
4: Grąžinti rezultatą: $\arg\max_v \sum_{(\vec{x}_i, y_i) \in \mathcal{D}_z} \mathcal{I}(v = y_i)$ – dažniausią klasę tarp k artimiausių kaimynų

Algoritmo trūkumai:

- didelis skaičiavimų bei atminties panaudojimo kompleksiškumas
- visoms požymių vektoriaus komponentėms (ypatybėms) priskiriama vienoda svarba (algoritmas negali išskirti esminių požymių vektorių bruožų, tačiau ir prieš algoritmo

apmokymą požymių vektorius reikia transformuoti taip, kad ypatybės vektorių erdvėje būtų pasiskirsčiusios tolygiai)

Regresijos uždavinys išsprendžiamas apskaičiuojant naujo požymių vektoriaus atstumo nuo k kaimynų (kurie pagal atstumą surandami apmokymo vektorių rinkinyje) vidurkį. Pagal vektorių atstumą gali būti kartu vidurkinamos ir vektoriams priskirtos vertės y.

Klasifikavimo uždavinyje y vertės, susijusios su požymių vektoriais yra užkoduojamos unitariniu kodu (angl. *one-hot encoding*) ir, suskaičiuojant vidurkius nuo k artimiausių kaimynų, kartu vidurkinamas ir šių vektorių vertes žymintis unitarinis kodas. Tokiu būdu, nustatoma kiekybinė proporcija naujiems požymių vektoriams būti priskirtiems tam tikroje klasėje. Apskaičiuojant didžiausią komponentės vertę $\arg\max_i \vec{x} = \{x_1, x_2, ..., x_i, ..., x_n\}$ suvidurkintame unitariniame kode gali būti apskaičiuojama klasifikuojamo vektoriaus klasė i.

Didesnis artimiausių kaimynų skaičius didina modelio atsparumą triukšmui, tačiau mažina gebėjimą atskirti požymių vektorių klases (arba didina vidurkinimo paklaidą regresijos uždavinyje.) Optimalus artimiausių kaimynų skaičius yra nustatomas empiriškai, kaip ir atstumų skaičiavimo metrika (požymių vektorių artimumas turėtų atitikti vektoriaus savybę būti priskirtam tam tikrai klasei) [8].

Yra įvairių k-NN algoritmo modifikacijų:

- svorinio koeficiento (dažnai turinčio funkcinę priklausomybę nuo atstumo) priskyrimas regresijos ar klasifikavimo žymėms y_i [1]
- svorinio koeficiento priskyrimas tam tikriems požymių vektoriams atskleidžiant išankstinę nuostatą apie tam tikrų požymių vektorių didesnį patikimumą nei likusiųjų (PEBLS) [1]
- modelyje saugomi požymių vektoriai gali būti šalinami, dėl atminties ir skaičiavimų spartinimo – kondensavimas (angl. *condensing*) – arba, dėl atsparumo triukšmui didinimo – redagavimas (angl. *editing*)
- neraiškiosios logikos artimiausių kaimynų metodas (angl. *fuzzy k-NN*) [9]
- sprendimo taisyklių pritaikymas pagal uždavinio specifiką (kitos, nei artimiausių kaimynių, strategijos, pavyzdžiui įtakos stačiakampiai, angl. *rectangle-of-influence*) [10]

1.3.1.2 Atraminių vektorių mašina Atraminių vektorių mašina yra prižiūrimojo mokymosi metodas, kuris gali būti pritaikomas klasifikavimo ir regresijos uždaviniams. Optimizavimo metu yra maksimizuojamas klasifikavimo tikslumas ir minimizuojama reguliarizacijos paklaida. Atraminių vektorių mašina yra tiesinis klasifikavimo metodas. Paprasčiausiu atveju, jei nepriklausomojo kintamojo $\vec{x}$ komponenčių plokštumos projekcijose, priklausomojo kintamojo y vertes (klases) galime atskirti tiesėmis, klasifikavimo uždavinį galime modeliuoti tiesinių plokštumų hipererdvėje. Bet kurio taško priklausomybę klasei C galime įvertinti apskaičiavę tiesinės plokštumos vektoriaus projekciją $\vec{w}$ šioje ribinėje hiperplokštumoje $C(\vec{x}, \vec{w}, b) = sgn(\vec{w} \cdot \vec{x} - b)$. Ribinei hiperplokštumai požymių vektorių rinkinyje sudaryti sprendžiamas tiesinio optimizavimo uždavinys – randamas hiperparametrų $\vec{w}$ rinkinys maksimizuojantis požymių vektorių rinkinio $\mathbb{D}$

projekcijų sumą [11]:

$$\arg\min_{\vec{x}\in\mathbb{D}} \frac{|\vec{x}\cdot\vec{w}+b|}{\sqrt{\sum_{i=1}^{d} w_i^2}} \tag{1}$$

Optimizavimo uždavinyje daromos prielaidos [12]:

1. optimizuojama funkcija yra išgaubtoji tiriamojoje požymių vektorių erdvėje, t.y., jei pasirinktume bet kuriuos du požymių vektorius $\vec{x}_1$, $\vec{x}_2$ ir šių vektorių santykio daugiklį q, $0 \leq q \leq 1$, tada požymių vektorių erdvėje optimizuojama išgaubtoji funkcija turi savybę $f(q\vec{x}_1 + (1-q)\vec{x}_2) \leq qf(\vec{x}_1) + (1-q)f(\vec{x}_2)$.

2. optimizuojamoms funkcijoms taikoma Sleiterio sąlyga: egzistuoja bent vienas požymių vektorius, tenkinantis sąlygą $sgn(\vec{x}\cdot\vec{w}+b) < 0$

Kadangi ieškome paviršiaus, maksimaliai atskiriančio vektorių požymių klases y_i, optimizavimo lygtyje galime taikyti apribojimą $|\vec{x}\cdot\vec{w}+b| \geq 1$ (optimizavimo tikslas yra $|\vec{x}\cdot\vec{w}+b| >> 0$), t.y.

$$\begin{cases} \vec{x}_i\cdot\vec{w}+b \geq 1, \text{jei požymių vektoriaus klasė yra } C, \text{ arba kitaip } y_i = 1 \\ \vec{x}_i\cdot\vec{w}+b \leq 1, \text{jei požymių vektoriaus nepriklauso klasei } C, \text{ arba kitaip } y_i = -1 \end{cases} \tag{2}$$

Šią lygčių sistemą glaustai galime užrašyti pavidalu $y_i(\vec{x}_i\cdot\vec{w}+b) \geq 1$.

Sudarome Lagranžo lygtį:

$$L(\vec{w}, b, \vec{\alpha}) = \frac{|\vec{w}|^2}{2} - \sum_{i=1}^{n} \alpha_i[y_i(\vec{w}\cdot\vec{x}_i+b) - 1], \alpha_i \geq 0, \tag{3}$$

Lagranžo lygtį galime suprastinti kintamųjų atžvilgiu. Apskaičiuojame dalines išvestines parametrams $\vec{w}$, bei b, randame balno taško vertes $w_j = \sum_i \alpha_{ij}x_{ij}y_j$, $b = y_k - \vec{w}\cdot\vec{x}_k$, k – bet kuris požymių rinkinio vektorius. Iš balno taško sąlygos, dalinės parametro b išvestinės lygties, nustatome apribojimą $-\sum_{i=1}^{n}\alpha_i y_i = 0$. Šiuos sprendinius panaudoję toje pačioje Lagranžo lygtyje, turime optimizavimo lygtį vien tik Lagranžo daugikliams [12]:

$$L(\vec{\alpha}) = \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{j=1}^{n}\sum_{i=1}^{n}\alpha_i\alpha_j y_i y_j \vec{x}_i\cdot\vec{x}_j \tag{4}$$

Matrica $\mathbb{D} = \vec{x}_i\cdot\vec{x}_j y_i y_j$ yra simetrinė. Matome, jog optimizavimo uždavinio kompleksiškumas yra proporcingas $O(n^2k)$ (n - požymių vektorių skaičius, vektorių ilgis - k). Požymių vektoriai, kurių Lagranžo daugiklio reikšmės $\alpha_i > 0$ yra vadinami atraminiais vektoriais

Tiesiškai neatskiriamus lygties sprendinius galime nustatyti panaudodami strategijas [12] :

- sukurdami papildomą laisvojo kintamojo apribojimą Lagranžo lygtyje, leidžiantį padidinti tam tikrų požymių vektorių klasifikavimo paklaidą $y_i(\vec{w}\cdot\vec{x}_i+b) \geq 1-\xi_i$
- pritaikydami Hilberto erdvės transformaciją požymių vektoriams: transformuoti požymių vektoriai $\vec{x} \rightarrow \phi(\vec{x})$ atvaizduojami požymių erdvėje $\mathbb{F}$ taip, jog skaliarinių funkcijų

sandauga Hilberto erdvėje $\kappa(\vec{x}_1, \vec{x}_2) = \langle \phi(\vec{x}_1), \phi(\vec{x}_2) \rangle$ yra neneigiama, pusiau apibrėžtinė ir simetrinė bet kurių dviejų požymių vektorių atžvilgiu. Pusiau apibrėžtinės skaliarinės funkcijų sandaugos branduolio pavyzdžiai yra skaliarinės vektorių sandaugos polinomas $(1 + \vec{x}_1 \cdot \vec{x}_2)^p$, radialinės bazės funkcija (Gauso branduolys) – $\exp -\frac{|\vec{x}_1 - \vec{x}_2|^2}{2\sigma^2}$.

Atraminių vektorių mašinos privalumas – tiesinio globaliai optimalaus klasifikavimo sprendinio optimizavimas.

Algoritmo trūkumai:

- algoritmas sprendžia tik tiesinio optimizavimo uždavinius; sudėtingesnio pobūdžio uždaviniuose reikalinga tam tikra, nežinoma netiesinės Hilberto erdvės transformacijos funkcija optimizavimo rezultatams gerinti.
- algoritmą sudėtinga pritaikyti dideliems požymių vektorių rinkiniams; gali būti naudojamos įvairios specifinės optimizavimo strategijos, kaip, pavyzdžiui, reti masyvai ar netiesinių transformacijų aproksimacijos funkcijos, kurios gali sumažinti klasifikavimo rezultatų tikslumą

1.3.1.3 k-means klasterizavimas k-means algoritmas yra neprižiūrimasis duomenų klasterizavimo algoritmas požymių vektorius $\mathbb{X} = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_n\}$ suskirstantis į grupes k. Nėra vieno absoliučiai teisingo būdo suskirstyti požymių vektorius į tam tikrą skaičių grupių k. Todėl klasterizavimo rezultatai yra prasmingi tiek, kiek yra prasmingas požymio vektoriaus ypatybių tarpusavio palyginimas. Duomenų klasterizavimo metu, tarpusavyje palyginami daugiamačiai kintamieji, jiems priskiriant artimiausio klasterio skaičių [6].

2 Algoritmas Apibendrintas k-means algoritmas [2]

Require: $k \in \mathbb{N}$ – klasterių skaičius
Require: $\mathcal{D} = \{\vec{x}_1, \vec{x}_2, ..., \vec{x}_m\}$ – apmokymo vektorių aibė
Require: ϵ – klasterizavimo tikslumo parametras

1: Inicijuoti klasterių centrus: atsitiktinai klasterių centrams priskirti vektorius $\vec{x}_i$ (k-means++; arba naudoti kitus metodus) $\mathbb{C} = \{\mu_1, \mu_2, ..., \mu_k\}$
2: **kartoti**
3: **jei** $\mathbb{C}_2$ buvo apskaičiuotas **tada**
4: Priskiriame $\mathbb{C}_2$ koordinates matricai $\mathbb{C}$
5: Kiekvienam vektoriui $\vec{x}_i$ priskiriame artimiausią klasterį μ
6: Perskaičiuojame klasterių centrų koordinates pagal vidurkį vektorių, priskirtų klasterių centrams, rezultatą išsaugojame matricoje $\mathbb{C}_2$
7: **until** Skirtumas tarp $\mathbb{C}$ ir $\mathbb{C}_2$ klasterių centrų yra didesnis už ϵ

k-means algoritmas taip pat gali būti įvairiai modifikuojamas (žr. skyriuje Klasifikavimas ir regresija, 15 puslapyje.) Jei nėra taisyklės, pagal kurią būtų galima perskaičiuoti klasterių centroides, sudaroma vektorių atstumo matrica ir pritaikomas k-medoidžių (angl. *k-medoids*) algoritmas, vietoje perskaičiuojamų klasterių centrų priskiriantis naujus požymių vektorius iš tos pačios klasterizuojamų vektorių aibės.

1.3 lentelė. Augalų ligų klasifikavimo metodai, analizuojantys lapų fotografijas [13]

Segmentavimas	**Išskirti požymiai**	**Klasifikavimas**	**Modelio tikslumas**
Ribinis spalvų kanalų (angl. *thresholding*) segmentavimas	Forma, tekstūra, dispersija, pilkumo transformacijos, spalvų histogramos	SVM	93.1% [14]
k-means segmentavimas	nuotraukos tekstūros požymiai	neuroniniai tinklai	94% [15]
Morfologinės operacijos	nuotraukos tekstūros požymiai	Bajeso neuroniniai tinklai	88.59% [17]
k-means segmentavimas	nuotraukos spalvų ir tekstūros požymiai	neuroniniai tinklai	93% [16]
Otsu segmentavimas	Diskrečiosios kosinusų transformacijos ir Haaro transformacijos spektras	SVM	94.45% [13]

1.4. Panašūs tyrimai

Augalų ligų požymiai gali būti nustatomi mašininio mokymosi algoritmais apdorojant augalų lapų fotografijas. [13] šaltinyje pateiktas metodas duomenų analizę atlieka trimis etapais: 1) pažeistų audinių fotografijos yra segmentuojamos, 2) nustatytoms nuotraukos sritims išskiriami požymiai, 3) išskirtos nuotraukos sritys yra klasifikuojamos. Yra atliktos kelios tokio metodo studijos, kuriose naudojami įvairūs algoritmai trims modelio pakopoms (duomenų rinkiniai nėra išsamiai aprašyti [13, 14]; duomenų rinkinio dydžiai šaltiniuose: 40, 2 klasės [13], 200, 6 klasės [15, 16], 100, 3 klasės [13])

Pažangesniuose ligų atpažinimo metoduose yra pritaikomi sąsūkos neuroniniai tinklai, automatiškai nustatantys augalų fotografijos sritis, būdingas klasifikuojamiems ligų požymiams [18]. šaltinyje algoritmas 99.35% tikslumu (F_1 įverčio metrika, InceptionNet neuroninių tinklų architektūra) klasifikavo 14 augalų rūšių, bei 26 skirtingas ligų klases (iš viso surinkta 54000 nuotraukų, Plant Village duomenų rinkinys, 38 skirtingoms klasėms.) Tačiau, ekspertams, internetinės paieškos sistemos pagalba, sudarius nuotraukų rinkinį ir ištestavus pastarąjį modelį, metodas teturėjo 31.4% tikslumą (palyginkime, 2.63% tikslumas yra įmanomas klases parenkant atsitiktinio spėjimo būdu, testavimo rinkinyje esant 38 klasėms.) Sąsūkos neuroninių tinklų metodo analizė (neuroninių tinklų parametrų studija,) yra detaliau nagrinėjama [19] šaltinyje. Nors pažangesni metodai geba geriau apibendrinti etaloninį, metodų palyginimo studijose dažnai naudojamą duomenų rinkinį, mašininio mokymosi modelis, praktiniu požiūriu, yra naudingas jei sudaromas panaudojant gausų ir kokybiškai sudarytą duomenų rinkinį, reprezentuojantį įvairius probleminės srities

atvejus.

1.5. Požymių išrinkimo metodai

Sprendžiant klasifikavimo uždavinius, tenka optimizuoti gausius klasifikavimo modelių hiperparametrų rinkinius, todėl požymių vektorių rinkinį yra tikslinga optimizuoti tam, kad: a) sumažintume skaičiavimo laiką klasifikavimo modeliui derinti; b) sumažintume riziką sudaryti modelį, pernelyg priderintą (angl. *overfitted*) mokymo duomenų rinkiniui, bet ne bendro pobūdžio problemai spręsti. Požymių vektorių rinkinio optimizavimo strategijos yra skirstomos į: požymių ištraukimo (angl. *feature extraction*) – rinkinio požymius keičiant dydžiais, apskaičiuojamais požymių tarpusavio funkcinės priklausomybių sąryšiais, ir požymių išskyrimo (angl. *feature selection*) – duomenų rinkinį sudarant iš požymių, tenkinančių tam tikras matematines sąlygas arba optimalumo kriterijus (požymių svarbą) tam sukurtose pagalbinėse euristikose ar procedūrose, kurios skaičiavimo sąnaudų požiūriu yra paprastesnės, nei sprendinio paieška klasifikavimo uždavinyje.

Paprasčiausiais požymių svarbos pavyzdys yra požymių vektorių rinkinys, sudarytas iš tam tikro skaičiaus tarpusavyje nekoreliuojančių požymių. Kaip ir mašininio mokymosi algoritmai, taip ir požymių optimizavimo algoritmai yra skirstomi į prižiūrimųjų, neprižiūrimųjų ir pusiau prižiūrimųjų algoritmų klases.

Optimalus požymių rinkinys pagerina mašininio mokymosi laiką, sumažina modelio parametrų skaičių, modelio parametrai geriau apibendrina įvesties vektorius.

Požymių išskyrimo kriterijai:

- Atvejo vertinimas (angl. *individual evaluation*) – požymiams priskiriamas rangas (požymio svarba)
- Poaibio vertinimas (angl. *subset evaluation*) – naudojama paieškos strategija optimaliausiam požymių rinkiniui rasti.

Požymių išskyrimo procedūra:

1. Poaibio sudarymas – paieškos euristika, kurioje sudaromas požymių poaibis (nusakomas paieškos būsena.) Pagrindinės paieškos problemos yra pradinės paieškos būsenos nustatymas, ir paieškos organizavimas

2. Poaibio įvertinimas – nepriklausomieji kriterijai: naudoja esmines apmokymo rinkinio savybes, požymių rinkiniui nėra reikalingi (mining) algoritmai.

3. Paieškos stabdymo kriterijai

4. Rezultatų validavimas – statistiniai testai naujai sudarytiems duomenų rinkiniams įvertinti

Požymių išskyrimo metodai [20]:

1. Filtravimas – naudojamos tiesioginės duomenų rinkinio vertinimo metrikos, nepriklausančios nuo klasifikavimo uždavinio modelio. Kiekvienam požymiui, paprastai, yra apskaičiuojamas

rangas. Rangas yra įvertinamas požymių vektoriams atskirai arba požymių vektorių grupėms (angl. *batch.*)

2. Karkaso metodas (angl. *Wrapper*) – Naudojamas klasifikavimo algoritmas, kuris iteraciniu būdu įvertina išskirtų požymių kokybę. Iteracija susideda iš dviejų etapų: paieškos euristikos ir surasto požymių vektorių rinkinio įvertinimo – klasifikavimo uždavinio metrikos. Klasifikavimo modeliui yra taikomas juodosios dėžės principas. Viena iš karkaso metodų kategorijų yra retųjų požymių mokymosi modeliai (angl. *Sparse learning based models.*)Šie metodai mažina klasifikavimo modelio variacinę paklaidą, mažiau reikšmingus požymius transformuodami standartinėmis reguliarizacijos strategijomis (L_2, L_1 norma.) Vienas iš karkaso metodo trūkumų yra per sudėtingas ir per ilgai trunkantis nuostolių funkcijos optimizavimas klasifikavimo uždavinyje.

3. Įterptinis metodas (angl. *Embedded*) – naudojamas mašininio mokymosi algoritmas (arba algoritmo dalis) optimaliausiam parametrų rinkiniui rasti.

4. Hibridinis metodas, pritaikantis visas įmanomas optimalių požymių paieškos strategijas. Šio metodo tikslas – rasti statistiškai bendriausius požymius, siekiant padidinti išrinktų požymių universalumą, sumažinti rezultatų nestabilumą mažiems, daugelio požymių vektorių rinkiniams.

Egzistuoja skirtingos požymių filtravimo teorijos, pagrindinės iš jų yra [20]:

- Požymių panašumo (angl. *Similarity based*) – vektoriai skirstomi pagal panašumo metrikas, vektorių panašumas apskaičiuojamas, remiantis vektorių klasėmis arba tam tikromis vektorių atstumo metrikomis (koreliacijos, Euklido, kosinusų, ...) Esminis metodo trūkumas yra tai, kad reikšmingai tarpusavyje koreliuojantys požymiai nėra pašalinami, todėl požymių rinkinys dažnai yra perteklinis.
- Informacijos teorija (angl. *Information theory*) – algoritmai, sudaryti pagal euristinius filtravimo kriterijus, maksimizuojančius požymių svarbą, bei minimizuojančius pasikartojančių požymių skaičių naujame požymių rinkinyje. Paprastai ši teorija yra taikoma diskretiniams požymiams klasifikavimo uždavinyje.
- Statistiniai metodai (angl. *Statistical methods*) – remiamasi statistiškai apskaičiuojamais dydžiais individualiems požymiams. Neatsižvelgiama į požymių tarpusavio priklausomybes, todėl požymių rinkinys gali būti sudarytas iš tarpusavyje koreliuojančių požymių. Metodų privalumas – geras skaičiavimų efektyvumas.

Jei nežinoma, kurie požymiai duomenų rinkinyje yra optimalūs yra sudėtinga lyginti tarpusavyje skirtingus požymių išskyrimo metodus, dėl bendro duomenų rinkinio dydžio, triukšmo duomenų rinkinyje, didelio požymių vektoriaus dimensijų skaičiaus. Požymių paieškos metodo tikslumas negali būti geresnis nei mašininio mokymosi metodo, kadangi klasifikavimo algoritmas yra naudojamas požymių išskyrimo metodo validavime.

Požymių atrankos metodai gali būti gerinami [21]:

- Metodų ansamblių

- Papildomai naudojant ansamblių medžius ar požymių ištraukimą
- Pritaikant specifines tiriamosios ekspertinės srities euristikas
- Sugretinant keletą skirtingų požymių atrankos metodų

Šio tyrimo metu buvo pritaikyti požymių išskyrimo metodai [20]:

1. **Laplaso įvertis** – požymių panašumo metrika. Požymių panašumo metrikos uždaviniai yra apibendrinami lygtimi

$$\max_S U(S) = \max_S \sum_{f \in S} U(f) = \max_S \sum_{\vec{f} \in S} \hat{\vec{f}}' \hat{\mathbb{S}} \hat{\vec{f}}, \tag{5}$$

kurioje $U(f)$ yra naudingumo funkcija požymiui f, $\hat{\vec{f}}$ yra standartizuoti ir normalizuoti požymių vektoriai, $\hat{\mathbb{S}}$ – vektorių panašumo matrica, sudaroma pritaikant įvairias, požymiams derančias, atstumo metrikas.

Laplaso įvertis yra grafų teorijos algoritmas (remiamasi Laplaso tikrinio atvaizdavimo teorija). Sprendžiama tikrinių verčių lygtis Laplaso matricai. Daroma prielaida, kad požymių multidimensinius sąryšius (vietinę daugdaros struktūrą) galima išsaugoti sugrupuojant vektorius (sudarant požymių grafą,) remiantis vektorių vietinės metrikos reikšmėmis (grafo svoriniais koeficientais [22].) Vietinės metrikos gali būti Euklido, koreliacijos, kosinusų atstumas tarp N artimiausių požymių ir t.t. Gauso branduolio transformacijos pagalba yra tiksliau atsižvelgiama į atstumus tarp gretimų požymių Euklido erdvėje, t.y. panašumo matricos elementus apskaičiuojame pagal formulę:

$$\begin{cases} \mathbb{S}(i,j) = \exp -\frac{\|\hat{\vec{f}}_i - \hat{\vec{f}}_j\|_2}{t}, \quad \hat{\vec{f}}_i, \hat{\vec{f}}_j - \text{artimiausių kaimynų vektoriai} \\ \mathbb{S}(i,j) = 0, \text{visais kitais atvejais} \\ \mathbb{D}(i,i) = \sum_{j=1}^{N} \mathbb{S}(i,j), - \text{diagonalieji panašumo matricos elementai} \end{cases} \tag{6}$$

$t \in [0; +\infty)$ - yra globalus parametras, kurį galime pasirinkti savo nuožiūra.

Uždavinys yra išsprendžiamas atskirai lygtimis sujungtiesiems grafo komponentams, kurie buvo sudaryti pastarosios procedūros metu:

$$\mathbb{L}\hat{\vec{f}} = \lambda \mathbb{D}\hat{\vec{f}} \tag{7}$$

$\hat{\vec{f}}$, λ yra tikriniai vektoriai ir jų vertės. Laplaso matrica (arba Laplaso operatorius, apibendrinta prasme,) $\mathbb{L} = \mathbb{D} - \mathbb{S}$. Algoritmas reikalauja mažiau skaičiavimo išteklių nei kiti netiesiniai optimizavimo metodai, kadangi sprendžiamas simetrinių tikrinių verčių uždavinys retiems požymiams. Požymių artimumas nustatomas pradinėje požymių vektorių erdvėje. Algoritmas yra mažiau jautrus triukšmui bei riktams [23].

2. **ReliefF įvertis** – metodas yra tinkamas duomenų rinkiniui, turinčiam stipriai tarpusavyje koreliuojančius požymius. Požymių svarbumas ($\vec{w}$, atitinkantis požymių vektoriaus dydį) yra nustatomas pradinę vektoriaus svarbumo reikšmę, lygią nuliui ($\vec{w} = \vec{0}$,) tam tikro skaičiaus

iteracijų metu didinant dėmeniu

$$w_p = w_p - \sum_{j=1}^{k} \frac{(f_{rp} - f_{hj})}{mk(\max \mathbb{F}_p - \min \mathbb{F}_p)} + \sum_{c \neq klas(\vec{f}_r)} \frac{P(c)}{1 - P(klas(\vec{f}_r))} \sum_{j=1}^{k} \frac{f_{rp} - f_{cjp}}{mk(\max \mathbb{F}_p - \min \mathbb{F}_p)}, \tag{8}$$

$\mathbb{F}_p$ – visų požymio p reikšmių aibė. f_{rp} – p-tasis, kiekvienos iteracijos metu parinkto, atsitiktinio vektoriaus požymis. $\mathbb{H}$ yra vektoriaus $\vec{f}_r$ k- artimiausių kaimynų aibė. $\mathbb{M}_j$ yra k-artimiausių kaimynų vektoriaus $\vec{f}_r$ kiekvienoje j-tojoje klasėje, skirtingoje, nei yra pastarasis vektorius. Įvertis didėja jei atsitiktinio proceso metu yra mažinami tos pačios klasės požymio atstumo skirtumai ir didinami atstumai nuo artimiausių požymių, esančių skirtingose, nei $\vec{f}_r$ požymis klasėse [24].

3. **Fišerio įvertis** – Požymių išskyrimo metodas, nustatantis didžiausius įverčius požymiams, esantiems mažiausiai nutolusiems nuo savosios klasės pavyzdžių, bei toliausiai – nuo likusių klasių pavyzdžių.

$$FS(f_i) = \frac{\sum_{j=1}^{c} n_j(\mu_{ij} - \mu_i)^2}{\sum_{j=1}^{c} \sigma_{ij}} \tag{9}$$

n_j – požymių skaičius klasėje, μ_i – požymio vidutinė reikšmė duomenų rinkinyje, μ_{ij} – požymio vidutinė reikšmė duomenų klasėje, σ_{ij} – požymio vertės standartinis nuokrypis klasėje.

4. χ^2 **įvertis** – metodas yra sudarytas, remiantis statistinio testo nulio hipoteze, teigiančia, kad vidutinio nepriklausomų atsitiktinių kintamųjų ir jų teorinio skaičiaus skirtumas klasėse yra būdingas χ^2 skirstiniui [25]. Juo didesnis skirtumas tarp atsitiktiniu būdu parenkamų kintamųjų skaičiaus ir skaičiaus būdingo teoriškai nepriklausomiems kintamiesiems klasėse tuo didesnė tikimybė, jog šių klasių kintamieji tarpusavyje koreliuoja. χ^2 įvertis apskaičiuojamas

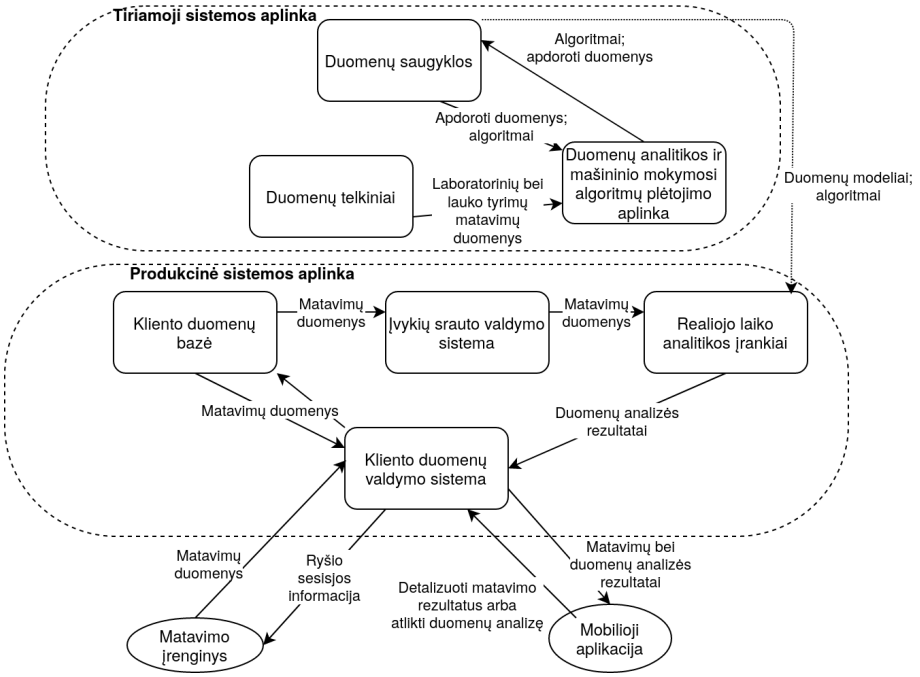
$$CS(f_i) = \sum_{j=1}^{r} \sum_{s=1}^{c} \frac{(n_{js} - \mu_{js})^2}{\mu_{js}} \tag{10}$$

2. Projektinė dalis

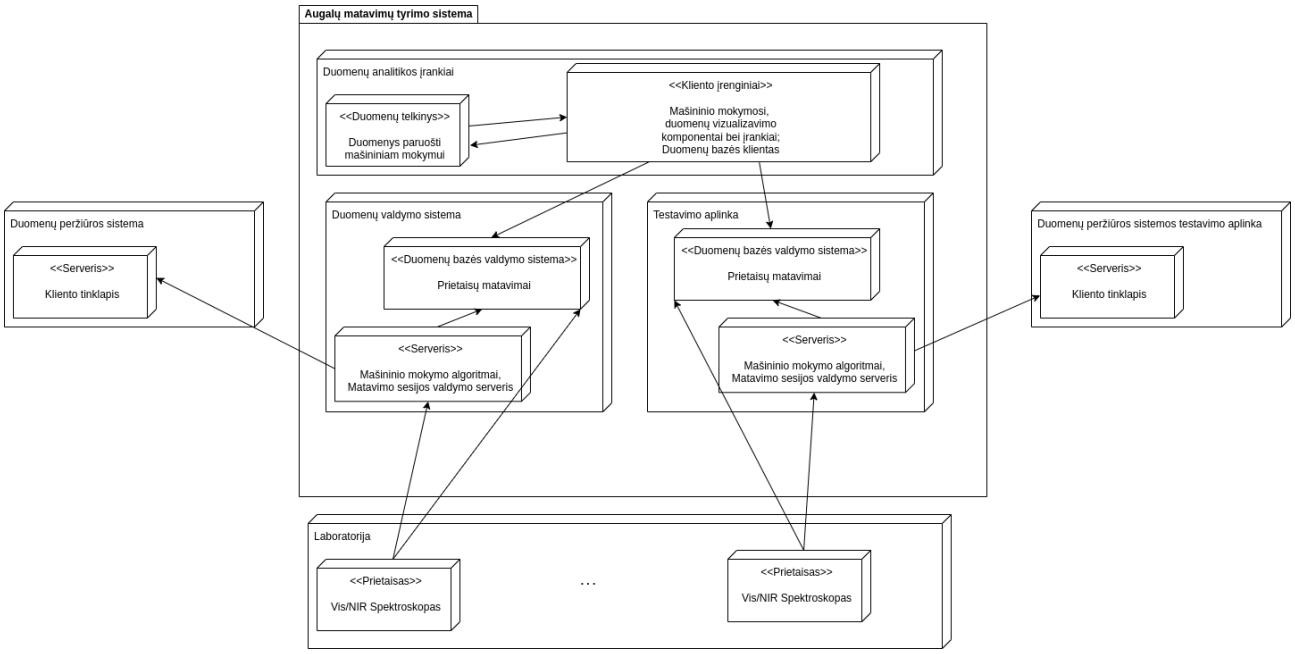
2.1. Projekto programinė sistema

Automatizuojama veiklos sritis yra laboratorijos bei lauko spektriniai matavimai, bei jų paruošimas ekspertinės sistemos mašininio mokymosi algoritmų prototipams. Laboratorijos tyrimų matavimus, bei augalų matavimo aplinkos sąlygas buvo registruojamos tyrimų aplinkos sistemoje (2.1 pav.). Priešingai, nei gamybinė, tyrimų aplinkos sistema yra pritaikyta sistemos analitikams, laborantams, matavimo duomenų analizei, mašininio mokymosi algoritmų plėtojimui, rezultatų vizualizavimui atlikti. Šios programinės įrangos pagrindiniai duomenų šaltiniai yra reliacinės duomenų bazės bei vietinės infrastruktūros duomenų telkiniai. Sistemą sudaro bendri programinės įrangos moduliai, kurių pagrindinės funkcijos yra duomenų bazės sąsaja, matavimo prietaisų parametrų interpretavimas, spektrinių matavimų kalibravimo algoritmai, matavimų agregavimas, įvairūs požymių vektorių formavimo metodai, matavimo aplinkos duomenų integravimas, bei augalų cheminių tyrimų rezultatų statistinė analizė. Tyrimų aplinkos sistemos tikslas – sukurti klasifikavimo algoritmus, kurių pagalba galėtume įvertinti optimalias augalų tręšimo sąlygas. Klasifikavimo algoritmai identifikuoja mitybos elementų trūkumo atvejus augaluose.

Kliento duomenų valdymo sistemoje, naudotojas, remdamasis pateiktais matavimo rezultatais, dirvožemio informacija sistemoje, gali priimti sprendimus, kokius cheminius preparatus naudoti būtų optimalu ar tikslinga. Įvykių srauto valdymo sistemoje yra vykdomos matavimo rezultatų klasifikavimo užduotys, kurių rezultatai yra siunčiami į kliento duomenų valdymo sistemą. Klasifikavimo algoritmai yra konfigūruojami kliento duomenų valdymo sistemoje (galime pritaikyti skirtingus klasifikavimo algoritmus, atlikti pakartotinius skaičiavimus, pritaikyti algoritmus atsižvelgiant į kliento pateiktas augalų veisles, augimo tarpsnius.)



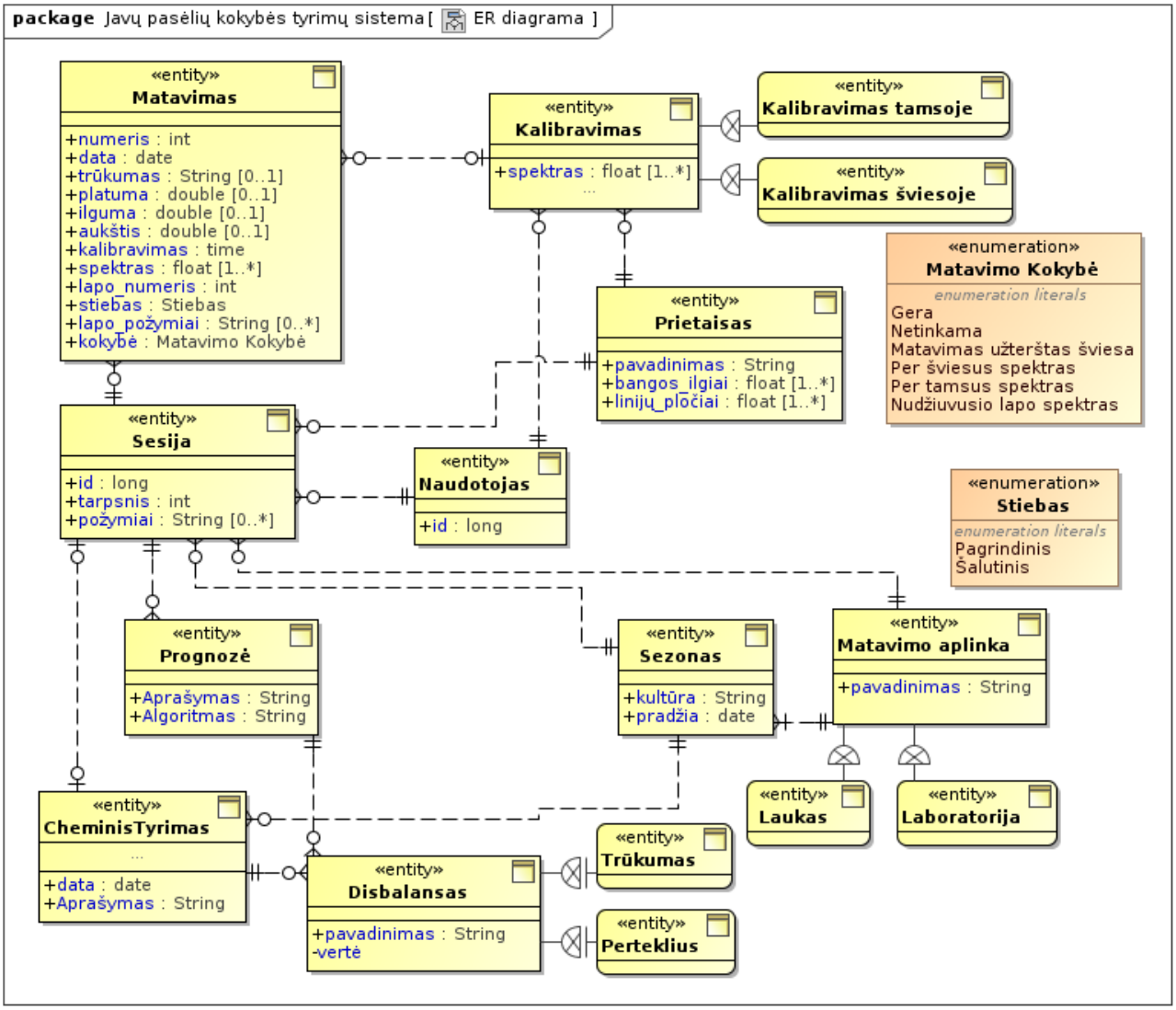
2.1 pav. Programinės sistemos architektūros diagrama



2.2 pav. Programinės sistemos diegimo diagrama

Augalų matavimų tyrimų programinę sistemą (2.2 pav.) sudaro:

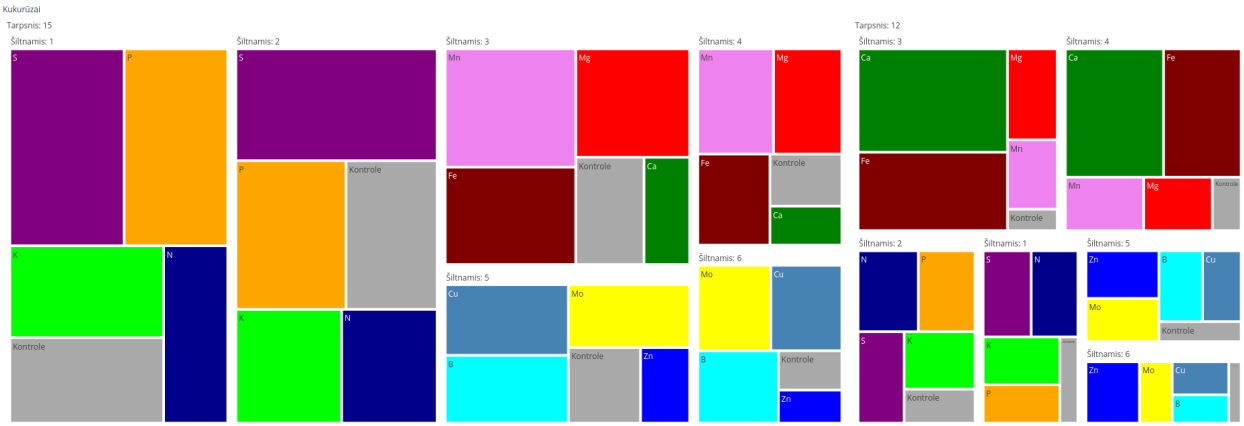
- Duomenų valdymo sistemos serveris paslaugos naudotojams teikiantis automatinio ekspertinio vertinimo duomenis, generuojamus statistinių modelių. Pagrindiniai sistemos naudotojai yra specializuoto įrenginio, atliekančio augalų spektrinius matavimus naudotojai, bei duomenų analitikai.
- Reliacinė duomenų bazė, kurios paskirtis yra registruoti įvairių spektrometrų matavimus, bei matavimo aplinkos parametrus registruojamus prietaisuose (GPS informacija, naudotojo įvedami parametrai.)
- Duomenų peržiūros įrankiai, kuriuos duomenų analitikas naudoja peržiūrėti surinktus spektrinius matavimus naudojant įvairius filtravimo parametrus. Mašininiam mokymui tinkami matavimai yra išsaugomi duomenų telkiniuose (failų sistemose, atskirose duomenų bazėse.) Optimaliausi algoritmai yra įdiegiami duomenų valdymo sistemos serveryje, kuriame galutiniam sistemos naudotojui pateikus spektrinius matavimus, duomenų peržiūros platformoje yra rodomos matavimo rekomendacijos.



2.3 pav. Duomenų bazės esybių ryšių diagrama

Duomenų bazės esybių ryšių diagramoje (2.3 pav.) yra pateikti tyrimo metu surinktų duomenų sąryšiai. Matavimai yra grupuojami atskiroje matavimo sesijose. Kalibravimo bei augalų matavimai yra kaupiami atskirose modelio esybėse. Augalų matavimai yra kalibruojami vėliau, išanalizavus kalibravimo bei matavimų duomenis. Šiuos duomenis vėliau peržiūrint gali būti taikomi skirtingi kalibravimo metodai.

2.2. Duomenų rinkinys požymių vektorių atrankos tyrimui

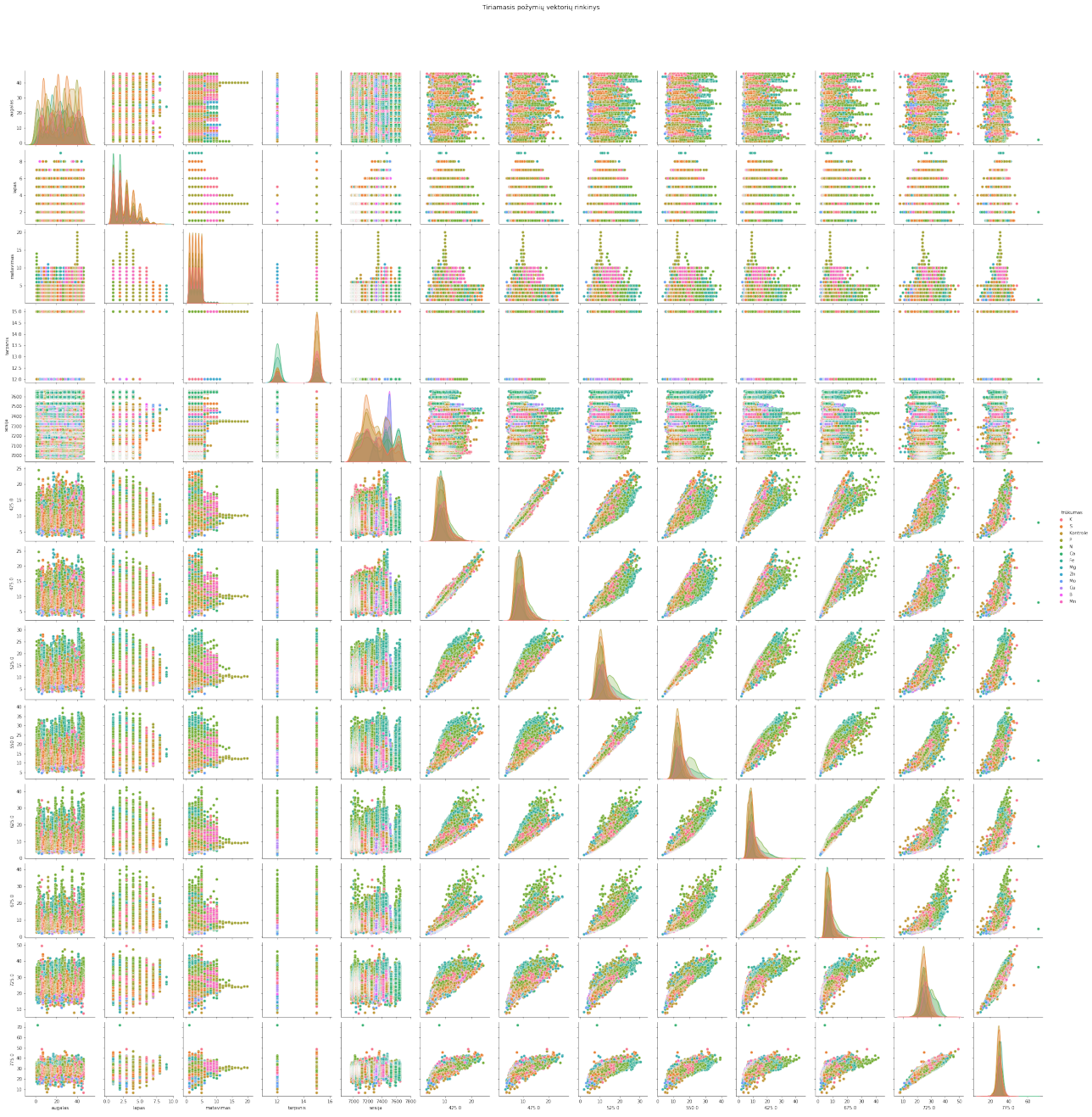


2.4 pav. Požymių vektorių skaičius duomenų rinkinio klasėse

Bendrą duomenų rinkinį sudaro laboratorijoje užaugintų augalų spektrų matavimai: 13 klasių, 270 augalų, 83125 požymių vektorių. Šiltnamiuose išmatuotų požymių vektorių klasių santykiniai dydžiai yra pateikti (2.4 pav.) iliustracijoje. Požymių vektorių rinkinys yra nesubalansuotas klasifikavimo uždaviniui, todėl optimizavimo metu buvo pritaikyti stratifikuoto kryžminio validavimo metodai.

Dėl sisteminių kalibravimo paklaidų, skirtingų prietaisų spektriniai matavimai buvo koreguoti tiesinio poslinkio koeficientu, apskaičiuotu remiantis mažiausiųjų kvadratų metodu. Iš duomenų rinkinio buvo pašalinti klaidingi matavimai bei kontrolės matavimai, esantys mikroelementų trūkumo klasėse (augalų matavimai, atlikti anksčiau kaip savaitė po trūkumo eksperimento pradžios.) Duomenų rinkinys buvo standartizuojamas, normalizuojamas, transformuojamas tokiu būdu, kad to paties augalo matavimai, tos pačios sesijos metu sudarytų naują požymių vektorių. Šį požymių vektorių sudarė skirtingų lapų vidutinės matavimų reikšmės, kombinaciniai skirtumai tarp šių reikšmių, bei augalo augimo tarpsnis. Augalo matavimo vektoriai buvo pakartotinai standartizuojami bei normalizuojami.

2.5 pav. yra pateiktos sklaidos diagramos, apibūdinančios pradinio vektorių rinkinio požymių tarpusavio priklausomybes. Iš sklaidos diagramų matome, jog daugelis spektrinių matavimų verčių yra koreliuojantys požymiai.



2.5 pav. Požymių tarpusavio priklausomybė tiriamajame duomenų rinkinyje

3. Tiriamoji dalis

3.1. Naudoti požymiai

Požymių vektoriai buvo sudaromi panaudojant pirmų aštuonių augalų lapų matavimų vidurkius (požymiai L1, L2, ... L8), bei šių matavimų skirtumų kombinacijas (L1-2, L1-3, ..., L7-8.) Požymių vektoriams sudaryti taip pat buvo naudojamas augalo augimo tarpsnio (T) kintamasis. Kiekvieno augalo lapo spektrinis matavimas yra sudarytas iš aštuonių, 50nm pločio, spektrinių juostų matavimų, kurių centriniai bangos ilgiai yra 425nm, 475nm, 525nm, 550nm, 625nm, 675nm, 725nm, 775nm.

Mašininio mokymosi uždavinys buvo sprendžiamas kartojant optimizavimą kryžminio validavimo rinkiniui 10 kartų – naudojama vidutinė 10-k kryžminio validavimo tikslumo (požymių vektorių validavimo rinkinyje) metrika. Kryžminio validavimo metu buvo stengiamasi spręsti balansuotą klasifikavimo uždavinį – skirtingų klasių požymių vektorių rinkinius sudarė tiek požymių vektorių, kiek jų turėjome mažiausio dydžio klasėje. Likusių klasių požymių vektoriai buvo parenkami atsitiktiniu būdu.

3.2. Požymių atranka

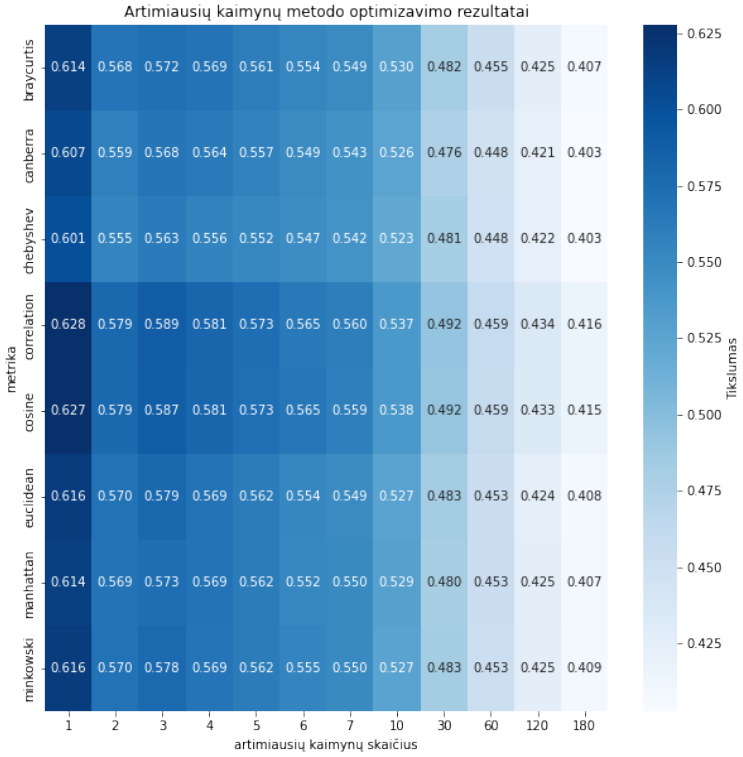
Svarbiausi požymiai buvo nustatyti remiantis filtravimo metodais: ReliefF, Fišerio, Laplaso bei χ^2 įverčiu (3.4). Požymių filtravimo metodo stabilumui įvertinti buvo pasirinkta optimizuoti keletą hiperparametrų optimizavimo paviršiaus taškų ir tris, skirtingo dydžio požymių vektorių poaibius (10%, 12.5% ir 20% bendrojo požymių rinkinio). Požymių rinkiniai buvo palyginami remiantis SVM RBF klasifikavimo tikslumo metrika. Išskirti požymiai buvo taip pat lyginami su atsitiktiniais to paties dydžio požymių aibės vektoriais. Atsitiktinio metodo klasifikavimas buvo kartojamas 10 kartų, naudojant 10-k kryžminį validavimą. Požymių išskyrimo metodų palyginimui taip pat buvo naudojami principinės komponentų bei tiesinio diskriminanto analizės metodai.

Optimizavimo rezultatai yra pateikti 1 priedo lentelėse. Kiekvienoje lentelėje yra kryžminio validavimo metu nustatytas klasifikavimo algoritmo tikslumas validavimo rinkinyje, varijuojant modelio hiperparametrus ir požymių rinkinio dydį. SVM RBF hiperparametrų rinkinį sudarė trys reguliarizacijos C ir radialinės bazės funkcijų rinkinio γ parametrų paviršiaus taškai, pasirinkti iš hiperparametrų paviršiaus, sudaryto įvairaus dydžio požymių vektoriaus poaibiams (iki 12.5%.) Optimaliausi sprendiniai buvo nustatyti parametrams $\gamma \in [0.1; 0.01]$, $C \in [10^4; 10^5]$ (parametrai - $(\gamma, C) = (0.01, 10^4)$ ir $(\gamma, C) = (0.1, 10^5)$.) Taip pat buvo pasirinktas vienas mažiau optimalus parametrų rinkinys $(\gamma, C) = (0.001, 10^3)$. 1 priedo lentelėse matyti, jog klasifikavimo tikslumas monotoniškai didėja, didėjant optimalių požymių aibėms, tačiau skiriasi filtravimo metodų savybės išskirti santykinai sparčiau konverguojančius požymius. Tai yra paaiškinama skirtingomis požymių išskyrimo strategijomis, atsižvelgiant į požymių vektorių koreliaciją ir atstumus tarp išorinių ir vidinių klasės vektorių.

3.3. Klasifikavimas su visais požymiais

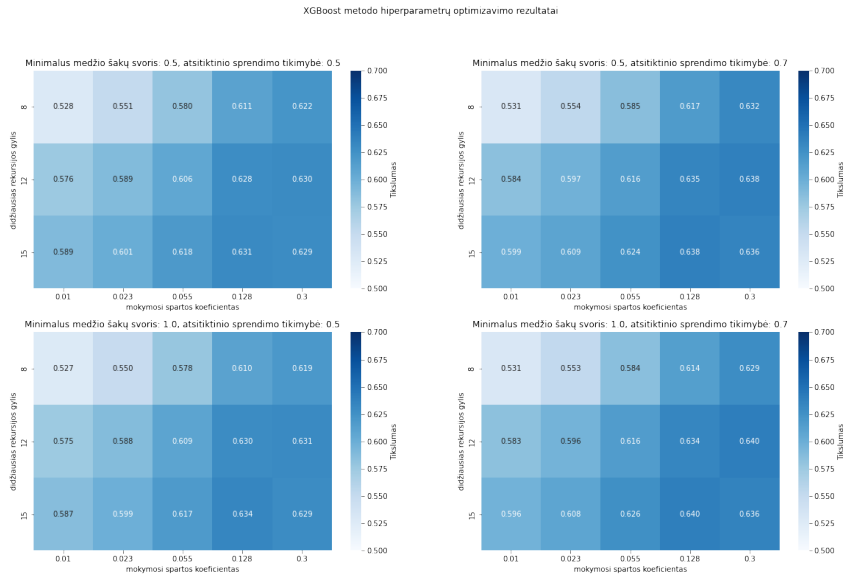
Artimiausių kaimynų metodo hiperparametrų optimizavimas buvo atliekamas panaudojant visą požymių vektorių rinkinį. Optimizavimo rezultatai yra pateikti 3.1 pav. Tiksliausi rezultatai buvo nustatyti naudojant vieno artimiausio kaimyno parametrą, tačiau, tokiu atveju, sudaromas modelis

yra mažiau atsparus klaidingiems matavimams mokymo rinkinyje, todėl paprastai yra naudojamas trijų ar daugiau artimiausių kaimynų modelis, siekiant sumažinti variacijos paklaidą validavimo rinkinyje.

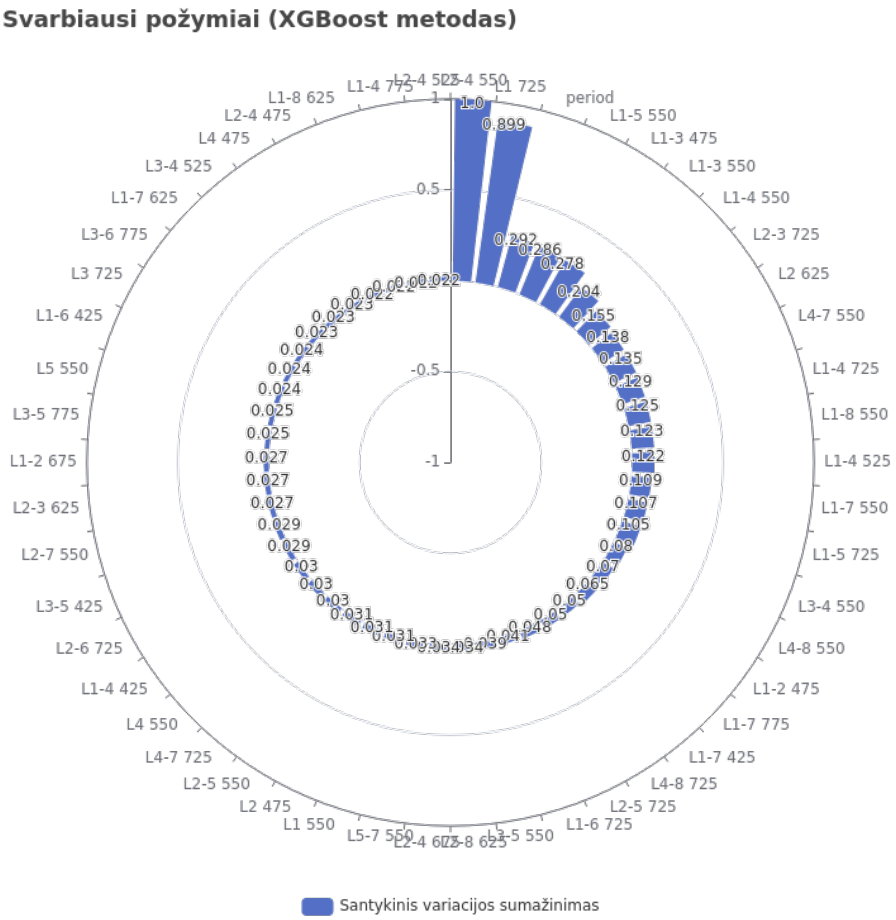


3.1 pav. k- artimiausių kaimynų metodo hiperparametrų optimizavimo rezultatai

Atsitiktinio miško (XGBoost) metodo hiperparametrai buvo optimizuojami panaudojant bendrą požymių vektorių rinkinį (3.2 pav.) Požymiai, daugiausiai sumažinę optimizuojamo duomenų rinkinio variaciją sprendimų medžiuose, yra pateikti 3.3 paveiksle. Didžiausias įvertis buvo nustatytas antrojo ir ketvirtojo lapo 550 nm atspindžio intensyvumo skirtume („L2-4 550"). Šis požymis statistiškai daugiausiai sumažino mokymosi duomenų rinkinio variaciją, sprendimų medžių sudarymo metu.



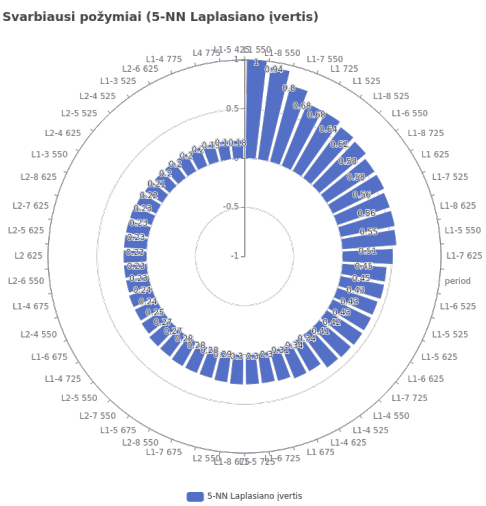
3.2 pav. XGBoost metodo hiperparametrų optimizavimo rezultatai



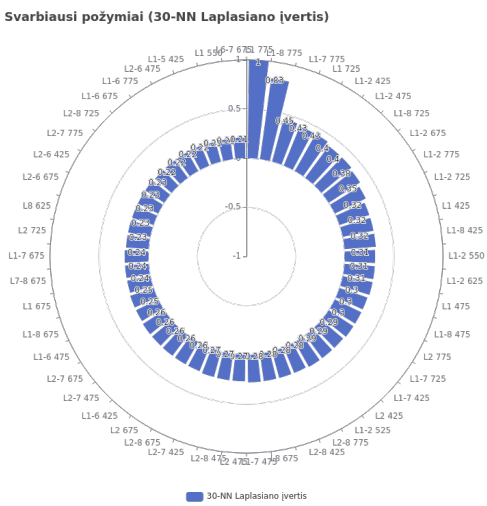
3.3 pav. Svarbiausi požymiai XGBoost klasifikavimo modelyje

Klasifikavimo uždavinio optimizavimo rezultatus visame požymių vektorių rinkinyje, panaudojant dalinius požymius, atrinktus skirtingų požymių filtravimo metodų, palyginome su klasifikavimo rezultatais nustatytais, įvertinant visus vektoriaus požymius:

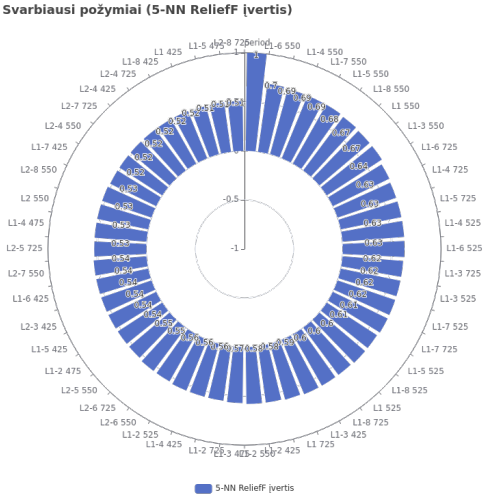
- XGboost mokymosi spartos koeficientas (eta=0.3), maksimalus rekursijos gylis 12, atsitiktinio sprendimo tikimybė 0.7, minimalus šakų svoris 1.0 – 64(1.4)% tikslumas
- SVC RBF C=1e4, gamma=0.01 (visi požymiai) – 58.4(0.7)% tikslumas
- SVC RBF C=1e4, gamma=0.01 (50 χ^2 įverčio požymių) – 49.9(0.7)% tikslumas
- SVC RBF C=1e4, gamma=0.01 (50 ReliefF 5-NN įverčio požymių) – 47.6(0.5)% tikslumas
- SVC RBF C=1e4, gamma=0.01 (50 ReliefF 30-NN įverčio požymių) – 47(0.5)% tikslumas
- SVC RBF C=1e4, gamma=0.01 (50 Laplaso 30-NN įverčio požymių) – 42.4(0.8)% tikslumas
- SVC RBF C=1e4, gamma=0.01 (50 Laplaso 5-NN įverčio požymių) – 36(0.4)% tikslumas
- SVC RBF C=1e4, gamma=0.01 (50 Fišerio įverčio požymių) – 48.8(0.7)% tikslumas
- 3-NN koreliacijos atstumo metrika ((visi požymiai) – 58.9(0.7)% tikslumas
- 3-NN koreliacijos atstumo metrika (50 χ^2 įverčio požymių) – 53.9(1.2)% tikslumas
- 3-NN koreliacijos atstumo metrika (50 ReliefF 5-NN įverčio požymių) – 57.4(0.9)% tikslumas
- 3-NN koreliacijos atstumo metrika (50 ReliefF 30-NN įverčio požymių) – 56.2(1.0)% tikslumas
- 3-NN koreliacijos atstumo metrika (50 Laplaso 30-NN įverčio požymių) – 48.6(0.8)% tikslumas
- 3-NN koreliacijos atstumo metrika (50 Laplaso 5-NN įverčio požymių) – 35.9(0.9)% tikslumas
- 3-NN koreliacijos atstumo metrika (50 Fišerio įverčio požymių) – 55.6(1.2)% tikslumas
- 3-NN Euklido atstumo metrika (visi požymiai) – 57.9(0.9)% tikslumas
- 3-NN Euklido atstumo metrika (50 χ^2 įverčio požymių) – 52.9(1.3)% tikslumas
- 3-NN Euklido atstumo metrika (50 ReliefF 5-NN įverčio požymių) – 54.2(1.3)% tikslumas
- 3-NN Euklido atstumo metrika (50 ReliefF 30-NN įverčio požymių) – 53.6(1.3)% tikslumas
- 3-NN Euklido atstumo metrika (50 Laplaso 30-NN įverčio požymių) – 46.5(0.9)% tikslumas
- 3-NN Euklido atstumo metrika (50 Laplaso 5-NN įverčio požymių) – 35.9(0.4)% tikslumas
- 3-NN Euklido atstumo metrika (50 Fišerio įverčio požymių) – 53.4(1.0)% tikslumas



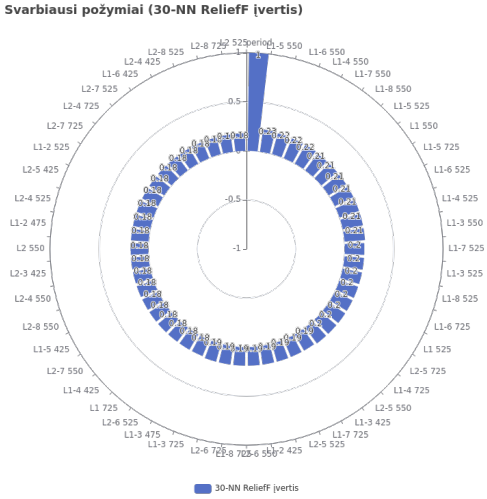
(a) 5-NN Laplaso įvertis, svarbiausi požymiai



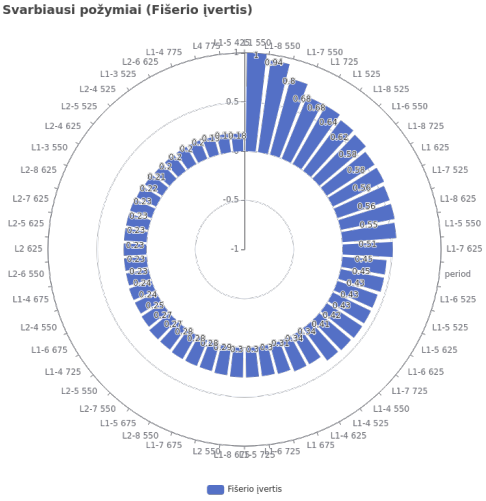
(b) 30-NN Laplaso įvertis, svarbiausi požymiai



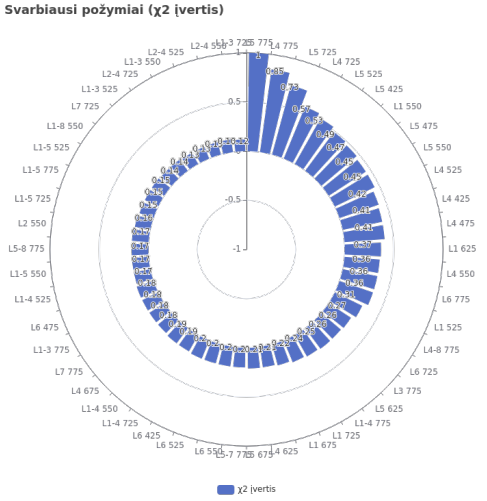
(c) 5-NN ReliefF įvertis, svarbiausi požymiai



(d) 30-NN ReliefF įvertis, svarbiausi požymiai



(e) Fišerio įvertis, svarbiausi požymiai



(f) χ^2 įvertis, svarbiausi požymiai

3.4 pav. Svarbiausių požymių paieškos, naudojant filtravimo metodus, rezultatai

Išvados

1. Tik LDA metodas pagerino klasifikavimo modelio gebėjimą tiksliau klasifikuoti požymių vektorius (C=1000.0, gamma=0.001 atveju). SVM, k-NN modeliai tiksliausiai klasifikavo vektorių rinkinį, turintį visus požymius, bet ne jų poaibį iki 50 požymių.

2. Tyrimo metu buvo nustatyta daugelio požymių tarpusavio koreliacija, todėl klasifikavimo rezultatams vėliau pagerinti gali būti naudojami kintamojo infliacijos daugiklio (VIF,) laso (angl. *Lasso*), keteros (angl. *Ridge*) metodai [26].

3. Laplaso įverčio požymių išskyrimo metodas nustatė požymius prasčiau, nei atsitiktinio išskyrimo atveju, nes metodui yra būdinga išskirti tarpusavyje koreliuojančius požymius.

Literatūros sąrašas

1. WU, Xindong ir kt. *DOI 10.1007/s10115-007-0114-2 SURVEY PAPER Top 10 algorithms in data mining*. 2007.
2. HASTIE, Trevor; Robert TIBSHIRANI; Jerome FRIEDMAN. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York, NY, 2009.
3. JINRU, Xue; Baofeng SU. Significant Remote Sensing Vegetation Indices: A Review of Developments and Applications. *Journal of Sensors*. 2017, t. 2017, psl. 1–17. **urlfrom**DOI: 10.1155/2017/1353691.
4. SNOWBALL K., A.D. Robson. *Nutrient Deficiencies and Toxicities in Wheat: A Guide for Field Identification*. 1991.
5. PETKUS, Vytautas; Ernestas PETRAUSKAS. *Augalo augimo sąlygų diagnostikos būdas ir įrenginys*. 2012-08.
6. GOODFELLOW, Ian; Yoshua BENGIO; Aaron COURVILLE. *Deep Learning*. The MIT Press, 2016.
7. X. HUANG Y. Ye, H. Zhang. *Unsupervised Learning Algorithms*. Springer International Publishing, 2016.
8. ABU ALFEILAT, Haneen ir kt. Effects of Distance Measure Choice on K-Nearest Neighbor Classifier Performance: A Review. *Big Data*. 2019, t. 7. **urlfrom**DOI: 10.1089/big.2018.0175.
9. BEZDEK, James C.; Siew K. CHUAH; David LEEP. Generalized k-nearest neighbor rules. *Fuzzy Sets and Systems*. 1986, t. 18, nr. 3, psl. 237–256.
10. TOUSSAINT, Godfried. Open Problems in Geometric Methods for Instance-Based Learning. 2003, psl. 273–283.
11. JAKKULA, Vikramaditya. Tutorial on support vector machine (svm). *School of EECS, Washington State University*. 2006, t. 37, nr. 2.5, psl. 3.
12. SHAWE-TAYLOR, John; Shiliang SUN. A review of optimization methodologies in support vector machines. *Neurocomputing*. 2011, t. 74, nr. 17, psl. 3609–3618.
13. AKHTAR, A.; A. KHANUM; S. A. KHAN; A. SHAUKAT. Automated Plant Disease Analysis (APDA): Performance Comparison of Machine Learning Techniques. 2013, psl. 60–65. **urlfrom**DOI: 10.1109/FIT.2013.19.
14. CAMARGO, Anyela; Jeremy SMITH. An image-processing based algorithm to automatically identify plant disease visual symptoms. *Biosystems Engineering - BIOSYST ENG*. 2009, t. 102, psl. 9–21. **urlfrom**DOI: 10.1016/j.biosystemseng.2008.09.030.
15. AL-HIARY, Heba ir kt. Fast and Accurate Detection and Classification of Plant Diseases. *International Journal of Computer Applications*. 2011, t. 17. **urlfrom**DOI: 10.5120/2183-2754.
16. ALBASHISH, Dheeb; Malik BRAIK; Sulieman BANI-AHMAD. A framework for detection and classification of plant leaf and stem diseases. In: 2011, psl. 113–118. **urlfrom**DOI: 10.1109/ICSIP.2010.5697452.
17. GURU, Devanur; Putaplar Basavanthappa MALLIKARJUNA; Manjunath SHANTHARAMU. Segmentation and classification of tobacco seedling diseases. In: 2011. **urlfrom**DOI: 10.1145/1980422.1980454.

18. MOHANTY, Sharada; David HUGHES; Marcel SALATHE. Using Deep Learning for Image-Based Plant Disease Detection. *Frontiers in Plant Science*. 2016, t. 7. **urlfrom**DOI: 10.3389/fpls.2016.01419.
19. TODA, Yosuke; Fumio OKURA. How Convolutional Neural Networks Diagnose Plant Disease. *Plant Phenomics*. 2019, t. 2019. **urlfrom**DOI: 10.1155/2019/9237136.
20. LI, Jundong ir kt. Feature selection: A data perspective. *ACM computing surveys (CSUR)*. 2017, t. 50, nr. 6, psl. 1–45.
21. KUMAR, Vipin. Sonajharia Minz,“ *Feature Selection: A literature Review”, Smart Computing Review*. 2014, t. 4, nr. 3, psl. 211–229.
22. BELKIN, Mikhail; Partha NIYOGI. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*. 2003, t. 15, nr. 6, psl. 1373–1396.
23. PHILLIPS, Jeff M.; Suresh VENKATASUBRAMANIAN. *A Gentle Introduction to the Kernel Distance*. 2011. **urlfrom** arXiv: 1103.1625 [cs.CG].
24. ROBNIK-ŠIKONJA, Marko; Igor KONONENKO. Theoretical and empirical analysis of ReliefF and RReliefF. *Machine learning*. 2003, t. 53, psl. 23–69.
25. PANCHENKO, Dmitry. *Chi-squared goodness-of-fit test*. 2011. Taip pat internete: https://ocw.mit.edu/courses/18-443-statistics-for-applications-fall-2006/896115b23d713bde212c45f62c086080_lecture11.pdf.
26. WIEL, Mark A. van de ir kt. *Think before you shrink: Alternatives to default shrinkage methods can improve prediction accuracy, calibration and coverage*. 2023. **urlfrom** arXiv: 2301.09890 [stat.ME].

Priedai

1.1. Priedas. Požymių išskyrimo metodų optimizavimo rezultatai

1.1 lentelė. Klasifikavimo tikslumas, k- optimaliausių požymių rinkinys, χ^2, 5-NN ReliefF, 5-NN Laplaso bei Fišerio įverčiai SVM (C=1000.0, gamma=0.001), 2343 vektorių rinkinys

		Požymių skaičius (k)										
		1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Metodas	**χ^2 įvertis**	0.15	0.165	0.166	0.166	0.166	0.28	0.291	0.306	0.328	0.356	0.404
	5-NN ReliefF įvertis	0.167	0.262	0.291	0.304	0.304	0.317	0.323	0.33	0.352	0.361	0.404
	30-NN ReliefF įvertis	0.167	0.276	0.288	0.295	0.304	0.312	0.321	0.328	0.354	0.36	0.405
	30-NN Laplaso įvertis	0.166	0.165	0.167	0.245	0.243	0.265	0.285	0.293	0.301	0.341	0.405
	5-NN Laplaso įvertis	0.166	0.165	0.17	0.172	0.173	0.267	0.285	0.291	0.303	0.325	0.404
	Fišerio įvertis	0.189	0.191	0.194	0.238	0.246	0.267	0.316	0.329	0.343	0.361	0.404
	Tiesinio diskriminanto analizė	0.25	0.334	0.371	0.417	0.427	0.447	0.452[1]	-	-	-	-
	Retų principinių komponentų analizė	0.166	0.166	0.267	0.271	0.288	0.355	0.371	0.387	0.396	0.396	0.402

		Požymių skaičius (k)										
		1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
	Radialinės bazinės funkcijos PCA (Nystromo transformacija)	0.166	0.18	0.194	0.196	0.207	0.228	0.229	0.23	0.23	0.23	0.23
	Atsitiktinis parametras	0.151 (0.025)	0.186 (0.045)	0.209 (0.035)	0.227 (0.026)	0.241 (0.031)	0.269 (0.025)	0.285 (0.022)	0.308 (0.023)	0.337 (0.006)	0.35 (0.006)	0.404 (0.0)

1.2 lentelė. Klasifikavimo tikslumas, k- optimaliausių požymių rinkinys, χ^2, 5-NN ReliefF, 5-NN Laplaso bei Fišerio įverčiai SVM (C=1000.0, gamma=0.001), 2928 vektorių rinkinys

		Požymių skaičius (k)										
		1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Metodas	χ^2 **įvertis**	0.15	0.167	0.167	0.167	0.168	0.285	0.299	0.31	0.337	0.363	0.409
	5-NN ReliefF įvertis	0.168	0.264	0.295	0.306	0.305	0.32	0.326	0.335	0.356	0.363	0.409
	30-NN ReliefF įvertis	0.167	0.276	0.291	0.299	0.306	0.315	0.322	0.332	0.358	0.363	0.409
	30-NN Laplaso įvertis	0.165	0.165	0.168	0.246	0.245	0.267	0.289	0.297	0.307	0.345	0.409
	5-NN Laplaso įvertis	0.165	0.166	0.17	0.173	0.174	0.271	0.289	0.295	0.308	0.331	0.409
	Fišerio įvertis	0.19	0.19	0.194	0.241	0.251	0.272	0.321	0.334	0.345	0.364	0.409

Metodas	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Tiesinio diskriminanto analizė	0.251	0.335	0.372	0.418	0.43	0.449	0.456[1]	-	-	-	-
Retų principinių komponentų analizė	0.169	0.167	0.27	0.274	0.296	0.359	0.376	0.389	0.4	0.4	0.407
Radialinės bazinės funkcijos PCA (Nystromo transformacija)	0.166	0.182	0.209	0.214	0.22	0.242	0.245	0.246	0.246	0.246	0.246
Atsitiktinis parametras	0.154 (0.026)	0.187 (0.047)	0.212 (0.036)	0.231 (0.026)	0.244 (0.031)	0.273 (0.025)	0.29 (0.022)	0.313 (0.022)	0.342 (0.006)	0.355 (0.006)	0.409 (0.0)

1.3 lentelė. Klasifikavimo tikslumas, k- optimaliausių požymių rinkinys, χ^2, 5-NN ReliefF, 5-NN Laplaso bei Fišerio įverčiai SVM (C=1000.0, gamma=0.001), 4686 vektorių rinkinys

Metodas	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
χ^2 įvertis	0.15	0.167	0.168	0.169	0.174	0.292	0.305	0.318	0.35	0.374	0.42
5-NN ReliefF įvertis	0.17	0.271	0.297	0.306	0.306	0.322	0.33	0.342	0.361	0.372	0.42

	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
30-NN ReliefF įvertis	0.17	0.281	0.294	0.302	0.307	0.319	0.328	0.338	0.364	0.371	0.42
30-NN Laplaso įvertis	0.166	0.165	0.169	0.251	0.25	0.272	0.298	0.307	0.316	0.352	0.42
5-NN Laplaso įvertis	0.166	0.166	0.171	0.176	0.178	0.279	0.298	0.304	0.319	0.343	0.42
Fišerio įvertis	0.192	0.193	0.198	0.246	0.259	0.273	0.327	0.34	0.352	0.371	0.42
Tiesinio diskriminanto analizė	0.252	0.336	0.375	0.419	0.433	0.455	0.465[1]	-	-	-	-
Retų principinių komponentų analizė	0.177	0.185	0.277	0.283	0.303	0.366	0.38	0.396	0.409	0.408	0.417
Radialinės bazinės funkcijos PCA (Nystromo transformacija)	0.166	0.181	0.225	0.234	0.237	0.274	0.277	0.278	0.278	0.278	0.278
Atsitiktinis parametras	0.156 (0.028)	0.189 (0.049)	0.216 (0.037)	0.237 (0.027)	0.249 (0.031)	0.28 (0.026)	0.299 (0.023)	0.322 (0.023)	0.353 (0.006)	0.367 (0.005)	0.42 (0.0)

1.4 lentelė. Klasifikavimo tikslumas, k- optimaliausių požymių rinkinys, χ^2, 5-NN ReliefF, 5-NN Laplaso bei Fišerio įverčiai SVM (C=10000.0, gamma=0.01), 2343 vektorių rinkinys

		Požymių skaičius (k)										
		1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Metodas	χ^2 **įvertis**	0.15	0.166	0.176	0.192	0.193	0.31	0.341	0.366	0.395	0.432	0.464
	5-NN ReliefF įvertis	0.167	0.272	0.308	0.316	0.317	0.356	0.384	0.396	0.417	0.425	0.464
	30-NN ReliefF įvertis	0.167	0.285	0.3	0.31	0.316	0.349	0.374	0.399	0.418	0.423	0.468
	30-NN Laplaso įvertis	0.165	0.165	0.166	0.25	0.256	0.292	0.33	0.339	0.364	0.395	0.468
	5-NN Laplaso įvertis	0.165	0.165	0.167	0.182	0.188	0.312	0.325	0.336	0.363	0.375	0.464
	Fišerio įvertis	0.195	0.214	0.222	0.263	0.284	0.301	0.354	0.372	0.398	0.429	0.464
	Tiesinio diskriminanto analizė	0.25	0.333	0.375	0.417	0.429	0.438	0.438[1]	-	-	-	-
	Retų principinių komponentų analizė	0.189	0.249	0.307	0.32	0.329	0.395	0.41	0.427	0.45	0.449	0.465

	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Radialinės bazinės funkcijos PCA (Nystromo transformacija)	0.165	0.162	0.262	0.267	0.279	0.34	0.385	0.397	0.409	0.415	0.417
Atsitiktinis parametras	0.159 (0.028)	0.193 (0.051)	0.221 (0.042)	0.249 (0.031)	0.266 (0.032)	0.317 (0.019)	0.346 (0.021)	0.378 (0.018)	0.41 (0.004)	0.422 (0.004)	0.464 (0.0)

1.5 lentelė. Klasifikavimo tikslumas, k- optimaliausių požymių rinkinys, χ^2, 5-NN ReliefF, 5-NN Laplaso bei Fišerio įverčiai SVM (C=10000.0, gamma=0.01), 2928 vektorių rinkinys

		Požymių skaičius (k)										
		1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Metodas	χ^2 **įvertis**	0.15	0.166	0.177	0.192	0.193	0.311	0.344	0.372	0.404	0.438	0.481
	5-NN ReliefF įvertis	0.168	0.275	0.309	0.315	0.316	0.358	0.385	0.4	0.423	0.427	0.481
	30-NN ReliefF įvertis	0.167	0.286	0.299	0.312	0.316	0.354	0.379	0.402	0.423	0.427	0.481
	30-NN Laplaso įvertis	0.166	0.166	0.168	0.25	0.259	0.295	0.332	0.343	0.368	0.399	0.481
	5-NN Laplaso įvertis	0.166	0.165	0.172	0.184	0.19	0.314	0.325	0.34	0.368	0.378	0.481
	Fišerio įvertis	0.203	0.216	0.223	0.264	0.283	0.302	0.355	0.374	0.403	0.434	0.481

	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Tiesinio diskriminanto analizė	0.25	0.333	0.38	0.421	0.433	0.452	0.456[1]	-	-	-	-
Retų principinių komponentų analizė	0.191	0.248	0.31	0.325	0.333	0.398	0.419	0.437	0.459	0.46	0.479
Radialinės bazinės funkcijos PCA (Nystromo transformacija)	0.164	0.162	0.262	0.271	0.285	0.342	0.387	0.397	0.413	0.42	0.424
Atsitiktinis parametras	0.16 (0.029)	0.193 (0.051)	0.222 (0.042)	0.25 (0.031)	0.268 (0.032)	0.321 (0.02)	0.35 (0.02)	0.383 (0.018)	0.415 (0.005)	0.43 (0.004)	0.481 (0.0)

1.6 lentelė. Klasifikavimo tikslumas, k- optimaliausių požymių rinkinys, χ^2, 5-NN ReliefF, 5-NN Laplaso bei Fišerio įverčiai SVM (C=10000.0, gamma=0.01), 4686 vektorių rinkinys

	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
χ^2 įvertis	0.15	0.167	0.178	0.195	0.196	0.313	0.347	0.381	0.415	0.458	0.509
5-NN ReliefF įvertis	0.17	0.276	0.311	0.316	0.317	0.361	0.389	0.406	0.431	0.443	0.509

	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
30-NN ReliefF įvertis	0.17	0.289	0.301	0.315	0.316	0.357	0.385	0.41	0.434	0.439	0.508
30-NN Laplaso įvertis	0.166	0.167	0.168	0.254	0.264	0.301	0.341	0.351	0.372	0.403	0.508
5-NN Laplaso įvertis	0.166	0.166	0.174	0.187	0.192	0.316	0.331	0.347	0.374	0.382	0.509
Fišerio įvertis	0.217	0.217	0.224	0.262	0.284	0.305	0.357	0.381	0.408	0.448	0.509
Tiesinio diskriminanto analizė	0.251	0.337	0.382	0.426	0.441	0.476	0.482[1]	-	-	-	-
Retų principinių komponentų analizė	0.192	0.254	0.313	0.327	0.336	0.406	0.429	0.452	0.48	0.481	0.506
Radialinės bazinės funkcijos PCA (Nystromo transformacija)	0.17	0.165	0.264	0.27	0.292	0.347	0.393	0.405	0.42	0.429	0.433
Atsitiktinis parametras	0.16 (0.029)	0.192 (0.052)	0.223 (0.043)	0.251 (0.032)	0.271 (0.032)	0.326 (0.02)	0.357 (0.02)	0.39 (0.018)	0.426 (0.005)	0.443 (0.004)	0.509 (0.0)

1.7 lentelė. Klasifikavimo tikslumas, k- optimaliausių požymių rinkinys, χ^2, 5-NN ReliefF, 5-NN Laplaso bei Fišerio įverčiai SVM (C=100000.0, gamma=0.1), 2343 vektorių rinkinys

		Požymių skaičius (k)										
		1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Metodas	**χ^2 įvertis**	0.15	0.173	0.183	0.198	0.199	0.314	0.35	0.39	0.416	0.439	0.445
	5-NN ReliefF įvertis	0.167	0.282	0.325	0.331	0.331	0.391	0.412	0.423	0.428	0.43	0.445
	30-NN ReliefF įvertis	0.167	0.291	0.305	0.326	0.331	0.384	0.412	0.424	0.432	0.426	0.445
	30-NN Laplaso įvertis	0.164	0.164	0.168	0.261	0.275	0.309	0.361	0.366	0.38	0.405	0.445
	5-NN Laplaso įvertis	0.163	0.164	0.187	0.196	0.2	0.325	0.348	0.361	0.38	0.392	0.445
	Fišerio įvertis	0.223	0.224	0.227	0.275	0.303	0.326	0.379	0.401	0.42	0.447	0.445
	Tiesinio diskriminanto analizė	0.236	0.289	0.311	0.359	0.376	0.412	0.435[1]	-	-	-	-
	Retų principinių komponentų analizė	0.185	0.274	0.314	0.324	0.343	0.38	0.393	0.412	0.429	0.431	0.444

	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Radialinės bazinės funkcijos PCA (Nystromo transformacija)	0.187	0.257	0.304	0.335	0.344	0.389	0.405	0.417	0.431	0.434	0.445
Atsitiktinis parametras	0.17 (0.028)	0.211 (0.049)	0.252 (0.034)	0.284 (0.022)	0.303 (0.025)	0.359 (0.02)	0.381 (0.019)	0.406 (0.019)	0.431 (0.006)	0.439 (0.006)	0.445 (0.0)

1.8 lentelė. Klasifikavimo tikslumas, k- optimaliausių požymių rinkinys, χ^2, 5-NN ReliefF, 5-NN Laplaso bei Fišerio įverčiai SVM (C=100000.0, gamma=0.1), 2928 vektorių rinkinys

		Požymių skaičius (k)										
		1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Metodas	χ^2 **įvertis**	0.15	0.174	0.185	0.199	0.201	0.317	0.356	0.395	0.428	0.458	0.468
	5-NN ReliefF įvertis	0.168	0.282	0.324	0.333	0.333	0.393	0.418	0.43	0.444	0.448	0.468
	30-NN ReliefF įvertis	0.167	0.288	0.304	0.328	0.334	0.387	0.416	0.434	0.448	0.444	0.465
	30-NN Laplaso įvertis	0.163	0.164	0.168	0.264	0.278	0.309	0.364	0.369	0.388	0.415	0.465
	5-NN Laplaso įvertis	0.164	0.165	0.186	0.198	0.201	0.329	0.355	0.367	0.39	0.401	0.468
	Fišerio įvertis	0.224	0.225	0.229	0.276	0.302	0.328	0.381	0.405	0.433	0.466	0.468

Metodas	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
Tiesinio diskriminanto analizė	0.232	0.287	0.301	0.352	0.38	0.426	0.449[1]	-	-	-	-
Retų principinių komponentų analizė	0.171	0.268	0.304	0.325	0.347	0.391	0.407	0.429	0.454	0.456	0.471
Radialinės bazinės funkcijos PCA (Nystromo transformacija)	0.182	0.256	0.306	0.338	0.347	0.396	0.413	0.43	0.444	0.453	0.467
Atsitiktinis parametras	0.17 (0.028)	0.211 (0.049)	0.253 (0.034)	0.285 (0.022)	0.304 (0.025)	0.363 (0.02)	0.388 (0.019)	0.415 (0.02)	0.444 (0.008)	0.455 (0.006)	0.468 (0.0)

1.9 lentelė. Klasifikavimo tikslumas, k- optimaliausių požymių rinkinys, χ^2, 5-NN ReliefF, 5-NN Laplaso bei Fišerio įverčiai SVM (C=100000.0, gamma=0.1), 4686 vektorių rinkinys

Metodas	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
χ^2 įvertis	0.15	0.177	0.185	0.201	0.204	0.323	0.371	0.417	0.454	0.489	0.509
5-NN ReliefF įvertis	0.17	0.283	0.326	0.333	0.334	0.395	0.43	0.443	0.474	0.483	0.509

	Požymių skaičius (k)										
	1	**2**	**3**	**4**	**5**	**9**	**14**	**21**	**35**	**50**	**289**
30-NN ReliefF įvertis	0.17	0.291	0.305	0.329	0.335	0.389	0.425	0.454	0.477	0.474	0.514
30-NN Laplaso įvertis	0.165	0.165	0.169	0.265	0.281	0.319	0.375	0.38	0.401	0.436	0.514
5-NN Laplaso įvertis	0.165	0.165	0.186	0.196	0.201	0.333	0.359	0.376	0.405	0.417	0.509
Fišerio įvertis	0.223	0.225	0.227	0.276	0.306	0.334	0.387	0.415	0.452	0.5	0.509
Tiesinio diskriminanto analizė	0.229	0.272	0.265	0.319	0.374	0.446	0.468[1]	-	-	-	-
Retų principinių komponentų analizė	0.167	0.243	0.289	0.311	0.348	0.404	0.432	0.462	0.494	0.494	0.511
Radialinės bazinės funkcijos PCA (Nystromo transformacija)	0.179	0.254	0.311	0.344	0.354	0.41	0.431	0.453	0.479	0.49	0.508
Atsitiktinis parametras	0.172 (0.028)	0.213 (0.05)	0.255 (0.034)	0.289 (0.022)	0.308 (0.025)	0.371 (0.02)	0.401 (0.018)	0.434 (0.019)	0.472 (0.009)	0.49 (0.008)	0.509 (0.0)

[1]Naudojama 12 komponentų, kadangi maksimalus komponentų skaičius negali būti didesnis nei vektorių klasių skaičius $N - 1$