VILNIAUS UNIVERSITETAS

LIJANA STABINGIENĖ

VAIZDŲ ANALIZĖ NAUDOJANT BAJESO DISKRIMINANTINES FUNKCIJAS

Daktaro disertacija
Fiziniai mokslai (P 000)
Informatika (09 P)

Vilnius, 2012 metai

Disertacija rengta 2011 – 2012 metais Vilniaus universiteto Matematikos ir informatikos institute.

Disertacija ginama eksternu.

Mokslinis konsultantas:

prof. dr. Kęstutis Dučinskas (Klaipėdos universitetas, fiziniai mokslai, informatika – 09 P).

Padėka

Dėkoju mylimam Jėzui už dovanotą suvokimą, kad pati neturiu nieko, ko nebūčiau iš Jo gavusi. Ačiū Jam už žmones, kurie konsultavo, egzaminavo, recenzavo ir vertino.

Reziumė

Vaizdų analizė šiomis dienomis yra labai svarbi dėl plataus pritaikymo daugelyje mokslo ir pramonės sričių. Ji apima sritis tokias, kaip segmentavimas, transformavimas, požymių išgavimas, objekto klasifikavimas (objekto atpažinimas) (*angl. pattern recognition*). Pastarasis vėlgi skaidomas į siauresnes dalis: neuroninių tinklų, sintaksės metodų, genetinių algoritmų, statistinių metodų ir kt. Taigi, statistiniais metodais paremtas objekto atpažinimas yra vadinamas statistiniu objekto atpažinimu (*angl. statistical pattern recognition*). Šis yra dviejų tipų: 1) klasifikavimas be mokymo ir 2) klasifikavimas su mokymu. Antrasis tipas, paremtas Bajeso diskriminantinėmis funkcijomis ir yra šio darbo objektas. Darbe pagrindinis dėmesys yra skiriamas būtent tokio tipo metodams.

Pagrindinė problema yra stacionaraus Gauso atsitiktinio lauko (GRF) stebinio klasifikavimas į vieną iš dviejų klasių, laikant, kad jis yra priklausomas nuo mokymo imties (TS) ir atsižvelgiant į jo ryšius su mokymo imtimi.

Tikslas yra pateikti klasifikavimo procedūrą, kuri GRF stebinius klasifikuotų optimaliai. Yra pasiūlyta nauja klasifikavimo su mokymu metodika, kuri duoda geresnius rezultatus, lyginant su įprastai naudojamomis Bajeso diskriminantinėmis funkcijomis.

Taip pat darbe pateiktos (išvestos) klaidų tikimybės Bajeso diskriminantinėms funkcijoms, kurios yra kaip šių funkcijų veikimo kriterijus. Be to, yra ištirta klaidų tikimybių priklausomybė nuo statistinių parametrų reikšmių.

Siūloma metodika tikrinama eksperimentų būdu, atstatant vaizdus, sugadintus erdvėje koreliuoto triukšmo. Tokia situacija pasitaiko natūraliai, pavyzdžiui, degant miškui dūmai uždengia nuotolinio stebėjimo vaizdą, gautą iš palydovo. Taip pat tokia situacija gana dažna esant debesuotumui. Esant tokiai situacijai erdvinės priklausomybės įvedimas į klasifikacijos problemą pasiteisina.

Disertaciją sudaro įvadas, trys skyriai, išvados, literatūros sąrašas ir autoriaus publikacijų disertacijos tema sąrašas.

Bendra disertacijos apimtis – 117 puslapių, numeruotų formulių 61, 30 paveikslų ir 13 lentelių. Literatūros sąrašą sudaro 53 šaltiniai.

Tyrimų rezultatai publikuoti 4 recenzuojamuose periodiniuose mokslo žurnaluose (viena iš jų ISI publikacija) ir 2 tarptautinės konferencijos recenzuojamuose leidiniuose. Rezultatai pristatyti tarptautinėje ir nacionalinėse konferencijose.

Raktiniai žodžiai: Statistinis objekto atpažinimas, vaizdo klasifikavimas, klasifikavimas su mokymu, vaizdo analizė, erdvinė koreliacija, Bajeso diskriminantinė funkcija, Gauso atsitiktinis laukas, Markovo atsitiktinis laukas, klaidos tikimybė, mokymo imtis, nuotolinio stebėjimo vaizdai.

Abstract

In our days image analysis is very important because of its usage in many different areas of science and industry. It is used in the areas such as segmentation, transformation, data mining and pattern recognition. Pattern recognition is also divided into several fields: neural networks, syntax methods, genetic algorithms, statistical methods and other. Pattern recognition methods based on statistical methods are called statistical pattern recognitions. The last one is of two types: 1) unsupervised classification and 2) supervised classification. The second type, based on Bayes discriminant functions is the object of this work. The biggest attention in this work is dedicated to the methods of such type.

The main problem is to classify stationary GRF observation into one off two classes, considering, that it is dependant on training sample ant taking in to account the relationship with training sample.

The objective of this work is to give the classification procedure, which optimally classifies GRF observations. The new supervised classification method, based on BDF methodology, is proposed and it gives better results comparing with other commonly used methods based on Bayes discriminant functions.

In this work error rates of Bayes discriminant functions are presented (derived), which are the criterion of these functions. Also, the dependences on statistical parameters are investigated for these error rates.

The proposed methods are investigated with experiments, restoring images, corrupted by spatially correlated noise. Such situation occurs naturally, for example, during the forest fire smoke covers the remotely sensed image, gathered from the satellite. Also such situation is often during cloudy days. During such situation the incorporation of the spatial dependences into the classification problem is useful.

Dissertation consists of introduction, three chapters, conclusions, the list of references and the list of author's publications associated with the topic of this dissertation.

The overall size of this dissertation is 117 pages, 61 numbered equations, 30 pictures and 13 tables. The list of references consists of 53 literally sources.

The results are published in 4 peer reviewed periodical science journals (one of them is ISI publication) and 2 international peer reviewed conference journals. Results are presented in international and national conferences.

Keywords: Statistical pattern recognition, image classification, supervised classification, image analysis, spatial correlation, Bayes discriminant function, Gaussian random fields, Markov random fields, error rate, training sample, remotely sensed images.

Žymėjimai

Simboliai

$\pi_l(y)$	apriorinės tikimybės
Δ_{0n}	sąlyginis Mahalanobio atstumas
Δ_0	marginalusis Mahalanobio atstumas
$\Phi(.)$	standartinio normaliojo skirstinio funkcija
$\{Z(s): s \in D\}$	atsitiktinis laukas
Ω_l	klasė, grupė
$r(s\text{-}u)$	erdvinės koreliacijos funkcija
L={1,2}	žymių aibė
ρ	klasterizavimo parametras
α	koreliacijos pločio parametras
h	atstumas tarp taškų
n_l	stebėjimo taškų skaičius su žyme l
X_y	plano matrica
σ^2	dispersija
δ_{ij}	Kronekerio delta
Σ_{0t}	kovariacijų matrica
$J(l,m)$	neneigiama divergencija tarp dviejų klasių
$\Delta(l)$	divergencijų vidurkis
S_8	aibė atitinkanti antros eilės kaimynystę s_0
ζ	erdvinis žymių planas

Santrumpos

AER – tikroji (faktinė) klaidos tikimybė.

BCR – Bajeso klasifikavimo taisyklė.

BDF – Bajeso diskriminantinė funkcija, kai klasifikuojamas stebinys yra priklausomas nuo mokymo imties.

BDFI – Bajeso diskriminantinė funkcija, kai klasifikuojamas stebinys yra nepriklausomas nuo mokymo imties.

CBER – sąlyginė Bajeso klaidos tikimybė.

CER – sąlyginė klaidos tikimybė.

DF – diskriminantinė funkcija.

DRF – diskretus atsitiktinis laukas.

EER – tiksli klaidos tikimybė.

EBER – vidutinė Bajeso klaidos tikimybė.

GLCM – pilkumo lygio pasikartojimų matrica.

GRF – Gauso atsitiktinis laukas.

USCL – klasifikavimas be mokymo.

KF – koreliacinė funkcija.

ME – klaidingo klasifikavimo tikimybė.

SCL – klasifikavimas su mokymu.

MRF – Markovo atsitiktinis laukas.

TS – mokymo imtis.

PBDF – įterpta Bajeso diskriminantinė funkcija, kai klasifikuojamas stebinys yra priklausomas nuo mokymo imties.

PBDFI – įterpta Bajeso diskriminantinė funkcija, kai klasifikuojamas stebinys yra nepriklausomas nuo mokymo imties.

RF – atsitiktinis laukas.

SLD – erdvinis žymių planas.

SSD – erdvinės imties planas.

STL – mokymo lokacijų aibė.

E_PER – tikėtina klaidos tikimybė.

Turinys

Žymėjimai viii

Turinys x

Įvadas 12
Tiriamoji problema 16
Darbo aktualumas 16
Tyrimų objektas 16
Darbo tikslas 17
Darbo uždaviniai 17
Tyrimų metodika 17
Darbo mokslinis naujumas ir jo reikšmė 17
Darbo rezultatų praktinė reikšmė 18
Ginamieji teiginiai 19
Darbo rezultatų aprobavimas 19
Disertacijos struktūra 20

Erdvinė informacija ir jos panaudojimas vaizdų analizėje 21
1.1. Erdviniai duomenys 21
1.2. Erdvinių imčių tipai 23
1.3. Erdvinių populiacijų modeliai 24
1.4. Erdvinių sąryšių struktūros populiacijoje 29
1.5. Erdvinės statistikos taikymas vaizdų analizėje 33
1.6. Statistinis objekto atpažinimas 39
1.7. Pirmojo skyriaus apibendrinimas ir disertacijos uždavinių formulavimas 43

Vaizdų, modeliuojamų GRF, klasifikavimo metodai 45
2.1. Bajeso sprendimo teorija 45
2.2. Diskriminantinės funkcijos 48
2.3. Bajeso diskriminantinės funkcijos 49
2.4. Priklausomybės nuo mokymo imties įvedimas 52
2.5. Empirinis klaidingo klasifikavimo vertinimas 54
2.6. Klasifikatoriaus veikimo vertinimas 55
2.7. Tiksli klaidos tikimybė Bajeso diskriminantinei funkcijai 57
2.8. Tikėtinos klaidos tikimybės aproksimacija 65
2.9. Vidutinė Bajeso klaidos tikimybė 71
2.10. Antrojo skyriaus išvados 83

Pasiūlytos metodikos taikymas ir eksperimentiniai rezultatai 85

3.1. Juodai balto vaizdo rekonstravimo pavyzdys, naudojant BDF ir PBDF 85
3.2. Klasifikavimas, paremtas pilkumo lygio pasikartojimų matricomis 93
3.3. Palydovinės nuotraukos vaizdo klasifikavimas 98
3.4. Klasifikavimas, realaus nuotolinio stebėjimo vaizdo, padengto debesimis 105
3.5. Skyriaus išvados 109

Bendrosios išvados 111

Literatūra ir šaltiniai 112

Autoriaus publikacijos disertacijos tema 116

Įvadas

Vaizdų analizė sparčiai plėtojama daugelyje mokslo bei pramonės sričių, tokių kaip medicina, astronomija, gynyba, robotų technika, mikroskopija, Žemės stebėjimas iš palydovų ir pan. Kiekviena sritis yra savita su specializuotomis sąvokomis bei algoritmais. Pavyzdžiui analizuojant pavojingus arba sunkiai pasiekiamus regionus informacija yra gaunama iš palydovo, kur tokie faktoriai kaip oro sąlygos, atmosferos užterštumas, skenerio iškraipymai ir pan., daro blogą įtaką duomenims (jie nebėra tokie tikslūs). Tokių duomenų analizei reikalingi gana sudėtingi tikimybiniai metodai, leidžiantys atsižvelgti į minėtus faktorius (Dučinskas and Šaltytė-Benth 2003). Taigi analizuojant tokius vaizdus, kaip nuotraukas iš palydovų, svarbus vaidmuo tenka erdvinės statistikos sričiai – statistiniam vaizdų klasifikavimui su mokymu, kuris ir yra šio *darbo objektas*.

Switzer pirmasis autorius, kuris pradėjo taikyti statistinių erdvinių duomenų požymius klasifikavime (Switzer 1980). Vėlesni autoriai aptaria tekstūros rotacijos įtaką vaizdo klasifikavimui, kai požymiai tenkina Gauso Markovo atsitiktinio lauko (GMRF) modelius (Deng and Clausi 2004). (Atkinson and Lewis 2000) apžvelgia geostatistinės informacijos panaudojimo galimybes nuotolinio stebėjimo vaizdų klasifikavime.

(Atkinson 2004) pateikė neparametrinį *k*–artimiausių kaimynų (k–NN) klasifikavimo su mokymu metodą skirtingų paviršiaus dangų taškų klasifikavimui. (Atkinson 2010) tyrė savo metodą (Atkinson 2004), įvedant į jį skirtingus atstumų svorius, bei pritaikė jį IKONOS palydovo nuotraukų klasifikavimui.

Žemės dangos klasifikavimas yra vienas iš svarbiausių nuotolinio stebėjimo (*remote sensing*) tikslų, kuris apjungia daugelį žmogaus ir fizinės aplinkos dalių (Foody 2002). Galimybė gauti informaciją apie Žemės dangą iš nuotolinio stebėjimo duomenų yra paremta daugeliu mokslinių taikymų (Townshend 1992).

Nuotolinio stebėjimo vaizdų klasifikavimas yra procesas, kurio metu iš vaizdų sukuriamas teminis žemėlapis (Tso and Mather 2001).

Nemažai klasifikavimo su mokymu schemų remiasi tradicinėmis statistikos technikomis, tokiomis kaip mažiausio vidurkių atstumo ir maksimalaus tikėtinumo (Tso and Mather 2001). Tipiniuose metoduose, kiekviena paveiksliuko spektrinio dažnio juosta yra atvaizduojama kaip viena ašis *n*–matėje požymių erdvėje, o visi paveiksliuko taškai gali būti atvaizduojami kaip taškai toje erdvėje. Skirtingos Žemės dangos taškai, remiantis duomenimis savybių erdvėje, gali būti atskiriami vieni nuo kitų, remiantis sprendimo priėmimo taisykle. Tokiu būdu, taškai gali būti suklasifikuoti į bendras Žemės paviršiaus klases.

Nuotolinio stebėjimo vaizdo klasifikavime laikoma, kad klasių žymės seka iš MRF (Nishii 2003), o požymių stebiniai yra sąlyginai nepriklausomi nuo klasių žymių (sąlyginė nepriklausomybė) (Cressie 1993). Tačiau esant duomenims arti vienas kito erdvėje yra labai tikėtina, kad jie tarpusavyje koreliuoja. Erdvinių priklausomybių įvedimas į klasifikacijos problemą yra itin aktualus, kai turimas vaizdas yra sugadintas erdvėje koreliuoto triukšmo. Paprastai tai atsitinka dėl gamtos reiškinių tokių, kaip rūkas, dūmai degant miškui, kurie ir pasižymi erdvine koreliacija (šią situaciją puikiai iliustruoja pav. 1.5). Darbe pasiūlyta klasifikavimo metodika gali būti taikoma ir vaizdams, sugadintiems debesimis. Tokių vaizdų labai dažnai pasitaiko Lietuvos teritorijos palydovinėse nuotraukose. Čia labiausiai tikėtina gauti „švarią“ nuotrauką gegužės mėnesį, nes daugiausia giedrų dienų būna gegužę (5,3 dienos, o tikimybė tesiekia 17 %) (Gudritienė 2007).

Darbe *sprendžiama problema*, kaip tiksliau suklasifikuoti erdvėje koreliuotą vaizdą, t.y., kaip optimaliai klasifikuoti Gauso atsitiktinio lauko (GRF) stebinius, priklausomus nuo mokymo imties. Darbe naujumas yra tas, jog klasifi-

kuojamas stebinys laikomas priklausomu nuo mokymo imties ir požymiai tenkina stacionaraus GRF modelį, o klasių žymės tenkina diskretaus lauko modelį.

Vienas iš pirmųjų autorių, panaikinusių nepriklausomumo prielaidą, jog klasifikuojami stebiniai yra nepriklausomi nuo mokymo imties, K. Dučinskas (Dučinskas 2009). Autoriaus darbuose sprendžiama problema yra klasifikavimas, kai stebėtą vaizdą tenka atskirti į keletą homogeniškų regionų pagal žymių taškus, pagrįstus tam tikra požymių informacija ir pagal informaciją apie erdvinės priklausomybės ryšius su mokymo imtimi. Įvedus stipresnę erdvinę koreliaciją tarp požymių stebinių yra gaunamos mažesnės erdvinio klasifikavimo paklaidos (Stabingienė and Dučinskas 2009), (Dučinskas and Stabingienė 2011), (Stabingienė and Dučinskas 2010). Šių metodų taikymas aprašytas darbe (Stabingienė *et al.* 2010).

Pikselio klasifikavimas į vieną iš klasių yra pagrindinė problema vaizdo struktūros analizėje (Mardia 1988). Autorius Mardia (Mardia 1984) tęsia šį tyrimą įtraukdamas erdvinius diskriminavimo metodus formuojant klasifikavimo žemėlapius. Taikymas erdvinio konteksto klasifikavimo metodų (klasifikavimo su mokymu metodų) geoerdvinių duomenų gavyboje yra aptartas autoriaus Shechar (Shechar *et al.* 2004). Reikia pažymėti, kad plačiai taikomi vaizdo užbaigimo metodai, paremti pavyzdžiu (Wu *et al.* 2010), yra glaudžiai susiję su erdvinės diskriminacijos metodais.

Erdvinės informacijos (vaizdo tekstūra, taškų forma, kryptingumas, atsikartojimas, artumas) įvedimas į nuotolinio stebėjimo vaizdų klasifikavimą yra nemažas potencialas (Haralick 1979). Buvę populiarūs metodai, kurie naudojo erdvinius filtrus, priglodinančius pradinę informaciją ar klasifikavimo rezultatus, buvo sukritikuoti autorių Barnsley ir Barr (Barnsley and Barr 1996). Anot šių autorių, priglodinimo filtrai ne tik sumažina erdvinę įvairovę, bet taip pat ignoruoja mažo dydžio heterogeniškumus, kurie pasitaiko urbanizuotose teritorijose.

Naudojant kontekstinius klasifikatorius, taškui priskirta klasė yra apibrėžiama kaip funkcija spektrinių savybių tiek pačio taško, tiek ir jo kaimynų, todėl ši kontekstinė informacija yra naudojama klasifikavimo tikslumo pagerini-

mui (Ketting and Landgrebe 1976). Šie klasifikatoriai naudojami perklasifikuoti klaidingai suklasifikuotus taškus arba perskirstyti teisingai suklasifikuotus taškus iš tam tikrų regionų, kai tie taškai nėra svarbūs vartotojui.

Anot autoriaus Dudani (Dudani 1976), pirmieji suformulavę k–NN sprendimo taisyklę buvo Fix ir Hodges, kurie naudojo šį metodą nežinomų stebėjimų priskyrimui tai klasei, kuri buvo dažniausia tarp k–NN (Fix and Hodges 1976). Vėliau autorius Dudani išplėtė šią taisyklę įvesdamas svorių schemą, kuri duoda daugiau svorio informacijai iš tų kaimynų, kurie yra arčiau nesuklasifikuoto stebėjimo, nei iš kaimynų, kurie yra toliau (Dudani 1976).

Daug kompiuterio resursų reikalaujantys metodai, kaip ir MCMC metodai, gali būti naudojami erdvinės klasifikacijos problemoms spręsti, tačiau atlikimas dažnai yra sunkus dėl skaičiavimo sudėtingumo. Tad, klaidų tikimybių išraiškų išvedimas yra labai svarbus klasifikavimo procedūrų veikimui nustatyti.

Pagrindinis *tikslas* yra ne tik optimaliai klasifikuoti GRF stebinius priklausomus nuo mokymo imties, bet ir pateikti (išvesti) analitines klaidų tikimybių išraiškas Bajeso diskriminantinėms funkcijoms bei ištirti minėtų klaidų tikimybių priklausomybes nuo tam tikrų statistinių parametrų reikšmių.

Tikslios klaidos tikimybės Bajeso klasifikavimo taisyklei, naudojant sąlyginės nepriklausomybės prielaidą, yra išvestos autoriaus Nishii darbe (Nishii and Eguchi 2006). Autoriaus Dučinsko darbe (Dučinskas 2009) aptariamos erdvinės klasifikacijos problemos Gauso požymių stebiniams, atsisakant sąlyginės nepriklausomybės prielaidos ir laikant klasių žymes fiksuotomis mokymo imtyje.

Tikslios klaidos tikimybės formulė ir pasiūlytos tikėtinos klaidos tikimybės, susijusios su PBDF, aproksimacija, pateikta autorių Dučinskas ir Stabingienė darbuose (Stabingiene and Dučinskas 2009), (Stabingiene and Dučinskas 2010). Šių autorių darbo apibendrinimas daugiamačiam požymio atvejui yra pateiktas darbe (Dučinskas and Stabingienė 2011). Čia aptariamas požymiams stacionaraus daugiamačio Gauso atsitiktinio lauko (GRF) modelis ir logistinio tipo diskretus atsitiktinio lauko modelis, paremtas 0-1 divergencija; vidutinė Bajeso klaidos tikimybė (EBER) yra išvesta dviejų klasių atveju.

Autoriai Raudys ir Pikelis (Raudys 1976), (Raudys and Pikelis 1980) ištyrė tikėtinos klaidos tikimybės priklausomybę nuo požymių erdvės dimensijos tam tikrų tiesinių klasifikatorių atveju, esant normaliam pasiskirstymui. Jie parodė, jog didinant požymių skaičių fiksuotam *n*, tikėtinos klaidos tikimybė pirmiausia mažėja, ir, po optimumo pasiekimo, auga vėl.

Straipsniuose (Stabingienė and Dučinskas 2009), (Dučinskas and Stabingienė 2011) yra ištirta minėtų klaidų tikimybių priklausomybė nuo statistinių parametrų reikšmių ir pagal gautus rezultatus teigiama, jog didesnė erdvinė koreliacija tarp požymių stebinių garantuoja mažesnes erdvinio klasifikavimo paklaidas.

Tiriamoji problema

Erdviniame vaizdų klasifikavime su mokymu problema yra lokacijos priskyrimas vienai iš klasių pagal mokymo imties požymių stebinius ir klasifikuojamo taško artumą su mokymo imtimi.

Darbo aktualumas

Vaizdų analizė yra svarbi medicinoje taikant kompiuterinei diagnostikai, dokumentų vaizdų analizė naudojama raidžių ar skaičių atpažinimui, svarbi srityse, tokiose kaip gynyboje, mikroskopijoje, Žemės stebėjime iš palydovų ir pan.

Darbe siūloma metodika ypač aktuali klasifikuojant vaizdus, gautus iš palydovų, kurie yra sugadinti triukšmo, atsirandančio dėl gamtos sąlygų, reiškinių, tokių kaip rūkas, debesuotumas, dūmai gaisro metu. Toks triukšmas gali būti modeliuojamas Gauso atsitiktiniais laukais.

Tyrimų objektas

Darbo tyrimų objektas – vaizdų analizei skirtas statistinis klasifikavimo su mokymu metodas, paremtas Bajeso diskriminantinėmis funkcijomis.

Darbo tikslas

Pagrindinis tikslas yra optimaliai klasifikuoti Gauso atsitiktinio lauko stebinius priklausomus nuo mokymo imties. Pateikti analitines klaidų tikimybių išraiškas Bajeso diskriminantinėms funkcijoms.

Darbo uždaviniai

- Pasiūlytą klasifikavimo su mokymu metodiką pritaikyti vaizdų analizėje.
- Pasiūlytos metodikos veiksmingumą patikrinti eksperimentų būdu, palyginant su Bajeso diskriminantinėmis funkcijomis, ignoruojančiomis erdvinę priklausomybę tarp klasifikuojamo taško ir mokymo imties. Palyginimui panaudoti klasifikavimo be mokymo metodą, paremtą pilkumo lygio pasikartojimų matricomis.
- Ištirti artimiausių kaimynų skaičiaus įtaką vaizdo klasifikavimo kokybei.
- Skaitiškai panagrinėti Bajeso diskriminantinių funkcijų klaidų tikimybių priklausomybes nuo tam tikrų statistinių parametrų reikšmių.

Tyrimų metodika

Darbe taikomi tikimybių teorijos, klasikinės ir erdvinės statistikos metodai, palyginimas iš kelių planų bei skaitmeninio modeliavimo metodai.

Darbo mokslinis naujumas ir jo reikšmė

Rengiant disertaciją buvo gauti mokslui nauji rezultatai:

- Pasiūlytas originalus statistinio klasifikavimo su mokymu metodas, pagrįstas Bajeso diskriminantine funkcija (BDF) ir erdvinės statistikos elementais. Šis metodas pritaikomas vaizdų analizėje.

- Atvejui, kai požymis modeliuojamas atsitiktiniu Gauso lauku, o klasės žymė modeliujama diskrečiu Markovo lauku, išvestos originalios formulės Bajeso klaidos tikimybėms.
- Skaitmeniškai nagrinėjama išvestų klaidų tikimybių priklausomybė nuo tam tikrų statistinių parametrų reikšmių.
- Pasiūlytos tikėtinos klaidos tikimybės, susijusios su Bajeso įterpta diskriminantine funkcija, aproksimacija, gali būti naudojama kaip erdvinių imčių plano kriterijus.
- Pasiūlyto metodo veiksmingumas yra tikrinamas eksperimentų būdu, naudojant realų vaizdą ir vaizdus, sugadintus su erdvėje koreliuotu Gauso atsitiktinio lauko triukšmu su skirtingais erdvinės koreliacijos pločio parametrais.
- Eksperimentų būdu yra tiriama panaudojamo artimiausių kaimynų skaičiaus (markoviškumo eilės) įtaka klasifikuojamo vaizdo kokybei.
- Siūloma metodika yra nauja, todėl jokiuose statistiniuose paketuose, nėra įdiegta. Gana sudėtingas algoritmas yra realizuotas R sistemos aplinkoje.

Darbo rezultatų praktinė reikšmė

Naudojant praktikoje pasiūlytą metodiką yra pagerinami klasifikuojamo vaizdo rezultatai. Šis metodas ypač aktualus atveju, kada duomenyse esantis triukšmas pasižymi erdvine priklausomybe.

Pasiūlytos metodikos realizacija gali būti pavyzdžiu klasifikuojant vaizdus.

Išvestos klaidų tikimybių išraiškos gali būti naudojamos kaip Bajeso diskriminantinių funkcijų veikimo matmuo.

Pasiūlytos tikėtinos klaidos tikimybės, susijusios su Bajeso įterpta diskriminantine funkcija, aproksimacija, gali būti naudojama kaip erdvinių imčių plano kriterijus.

Ginamieji teiginiai

- Pasiūlytas originalus vaizdų analizėje taikomas statistinio klasifikavimo su mokymu metodas, pagrįstas Bajeso diskriminantine funkcija (BDF) ir erdvinės statistikos elementais.
- Atvejui, kai požymis modeliuojamas atsitiktiniu Gauso lauku, o klasės žymė modeliujama diskrečiu Markovo lauku, išvesta originali formulė Bajeso klaidos tikimybei.
- Pasiūlyta tikėtinos klaidos tikimybės, susijusios su Bajeso įterpta diskriminantine funkcija, aproksimacija, panaudojama kaip erdvinių imčių plano kriterijus. Laikant, kad požymis modeliuojamas atsitiktiniu Gauso lauku, o klasės žymė modeliujama diskrečiu Markovo lauku.
- Išvesta vidutinė Bajeso klaidos tikimybė (EBER) dviejų klasių atveju, t.y., požymiams aptariamas stacionaraus daugiamačio Gauso atsitiktinio lauko (GRF) modelis, o klasių žymėms - logistinio tipo diskretus atsitiktinio lauko modelis paremtas 0-1 divergencija.

Darbo rezultatų aprobavimas

Disertacijos tema yra atspausdinti 6 moksliniai straipsniai: keturi – periodiniuose recenzuojamuose mokslo žurnaluose (Stabingienė and Dučinskas 2010), (Stabingienė *et al*. 2010), (Dučinskas and Stabingienė 2011), (Stabingienė *et al*. 2011), kur vienas – mokslo žurnale, įtrauktame į ISI sąrašą (Dučinskas and Stabingienė 2011); du – recenzuojamoje tarptautinės konferencijos medžiagoje (Stabingienė and Dučinskas 2009), (Dučinskas *et al*. 2011).

Disertacijoje atliktų tyrimų rezultatai buvo paskelbti tarptautinėje ir nacionalinėse mokslinėse konferencijose:

Tarptautinėje taikomų stochastinių modelių ir duomenų analizės konferencijoje (ASMDA), 2009 m., Vilniuje;

Lietuvos matematikų draugijos konferencijoje, 2010 m., Šiauliuose;

Lietuvos matematikų draugijos konferencijoje, 2011 m., Vilniuje;

Lietuvos matematikų draugijos konferencijoje, 2012 m., Klaipėdoje.

Disertacijos struktūra

Disertaciją sudaro įvadas, trys skyriai, išvados, literatūros sąrašas ir autoriaus publikacijų disertacijos tema sąrašas.

Bendra disertacijos apimtis – 117 puslapių, numeruotų formulių 61, 30 paveikslų ir 13 lentelių. Literatūros sąrašą sudaro 53 šaltiniai.

1

Erdvinė informacija ir jos panaudojimas vaizdų analizėje

Skyriuje analizuojami erdvinių duomenų ir imčių tipai. Supažindinama su erdvinių populiacijų modeliais bei erdvinių sąryšių struktūromis populiacijoje. Plačiau pateiktos kaimynystės chemos. Aptariamos erdvinės statistikos taikymo galimybės vaizdų analizėje. Skyriaus pabaigoje pateikiamas apibendrinimas ir tikslinami disertacijos uždaviniai.

1.1. Erdviniai duomenys

Tokios sąvokos kaip duomenys ar informacija atrodo susijusios viena su kita, tačiau skirtumų yra. *Duomenys* – tai objektyviai egzistuojantys faktai, vaizdai arba garsai, kurie gali būti naudingi tam tikram uždaviniui spręsti. *Informacija* – tai duomenys, kurių forma ir turinys yra pateikti tinkamu naudoti sprendimų priėmimo procese būdu. Duomenys virsta informacija, kai jiems suteikiamas kontekstas ir jie susiejami su tam tikra problema ar sprendimu (Dzemyda *et al.* 2008).

Erdviniai duomenys – bet kokios žinios apie vietovę, formas, santykius tarp jų, geografines ypatybes. Tai apima nuotolinio stebėjimo duomenis taip pat kaip ir žemėlapių duomenis.

Erdvinė informacija – apibrėžia fizinę objektų sritį ir ryšį tarp objektų. Erdvinės informacijos pramonė (industrija) yra platesnės informacinių technologijų sektoriaus specializuota dalis, kuri turi ryšį su daugeliu kitų disciplinų tokių kaip planavimas, natūralių resursų valdymas, inžinerija ir sveikatos pa-

slaugos. Erdvinė informacija yra naudojama kuriant tokias programas kaip krūmynų ar miškų gaisrų valdymo sistemos, greitosios pagalbos siuntimo paslaugos ir pan.

Trumpiau tariant, *erdviniai duomenys* – tai rezultatas stebėjimų, atliktų erdvėje. Kai stebėjimas atliekamas *d*-matės Euklido erdvės taške $s \in R^d$, tada galimas stebinys Z(s) erdvės taške *s* yra atsitiktinis dydis. Tuomet matematinis erdvinių duomenų (erdvinės populiacijos) modelis yra atsitiktinis laukas (RF) $\{Z(s): s \in D\}$, kur $D \subset R^d$ yra erdvinių indeksų aibė, o atsitiktinio lauko realizacija žymima $\{z(s): s \in D\}$.

Atsitiktinis laukas $\{Z(s): s \in D\}$ - tai rinkinys atsitiktinių dydžių, apibrėžtų vienoje tikimybinėje erdvėje (Ω, F, P) ir įgyjančių reikšmes erdvėje *B*. Erdvė *B* yra dažnai vadinama būsenų erdve. Čia ji yra $B \in R^q$. Kai q=1, atsitiktinis laukas vadinamas *skaliariniu*, o kai q>1 – *vektoriniu*. Jei visi baigtiniamačiai *Z* skirstiniai yra Gauso skirstiniai, tai toks laukas vadinamas Gauso atsitiktiniu lauku (GRF) (*angl. Gaussian random field*), o jo stebiniai – Gauso duomenimis (Dučinskas and Šaltytė-Benth 2003).

Eedviniai duomenys yra trijų tipų: *geostatistiniai, gardelės* ir *taškiniai vaizdai*. Pirmieji yra susiję su situacija, kai sritis *D* yra fiksuotas erdvės R^d poaibis, o stebinys Z(s) yra atsitiktinio lauko stebinys taške $s \in D$. Geostatistinio tipo uždaviniuose erdvinis indeksas *s* gali tolydžiai kisti erdvės R^d poaibyje *D*. Tuo jie skiriasi nuo kitų tipų uždavinių, nagrinėjančių duomenis gardelėje ar taškiniuose vaizduose.

Taškinių vaizdų analizėje $D \in R^d$ yra taškinio proceso galimų realizacijų aibė. Čia yra svarbios įvykių vietos bei jų skirstinys ir išsidėstymo tipas (išsidėsčiusios klasterizuotai, atsitiktinai ar taisyklingai).

Gardelės duomenų indeksų aibė *D* yra fiksuotas suskaičiuojamas arba dažniausiai baigtinis taisyklingai ar netaisyklingai erdvėje R^d pasklidusių taškų rinkinys. Toks rinkinys paprastai vadinamas gardele, o jo taškai yra susiję su artimiausios eilės kaimynais, antros eilės kaimynais ir t.t. Netaisyklingos gar-

delės taškų išsidėstymui nėra būdinga kokia nors tvarka, be to, ne visada akivaizdžios ir taškų sąsajos (Dučinskas and Šaltytė-Benth 2003).

Svarbu paminėti, kad gardelės tipo duomenys dažnai gaunami atliekant nuotolinius Žemės paviršiaus stebėjimus (palydovinėse nuotraukose Žemės paviršius yra padalijamas į stačiakampius, vadinamus pikseliais).

Šiame disertaciniame darbe yra naudojami gardelės tipo duomenys ne tik vaizdo rekonstravimo pavyzdyje ar palydovinės nuotraukos vaizdo klasifikavime, bet ir klaidų tikimybių tyrime.

1.2. Erdvinių imčių tipai

Norint turėti žinių apie tam tikros erdvinės srities charakteristikas, reikalinga sudaryti imtį, nes paprastai nėra tokios galimybės, kaip atlikti matavimus visuose tiriamos srities taškuose. Sudarant imtį, kaip įprasta, yra svarbu, jog imtis būtų reprezentatyvi, įvertiniai kuo tikslesni ir imties sudarymo išlaidos kuo mažesnės.

Taigi, šiame etape yra svarbus dalykas optimalaus plano radimas imčiai. Norint sudaryti optimalų planą, reikia žinoti erdvinės srities variacijų prigimtį. Tam sužinoti yra reikalingi imties duomenys (naudojamos kontrolinės imtys arba ankstesnių tyrimų rezultatai).

Yra keli klasikiniai imčių sudarymo planai, pagal kuriuos galima parinkti dominančios srities taškus. Išskiriami trys atvejai: 1) *tikimybinės* imtys, 2) *netikimybinės* (determinuotos) imtys, 3) *mišrios* (hibridinės). Tarp tikimybinių pasitaiko: atsitiktinės, stratifikuotos atsitiktinės, klasterinės.

Tikimybinės nuo netikimybinių skiriasi tuo, kad tikimybinių imčių objektai parenkami naudojant atsitiktinį mechanizmą (kuriame nors etape), o netikimybinių – griežtai determinuotas ar subjektyvias taisykles.

Tikimybinės *atsitiktinės imtys* (*angl. random spatial sampling or a spatially random sample*) sudaromos parenkant taškus nepriklausomai vienas nuo kito ir taško parinkimo tikimybės yra vienodos. Trūkumas tas, jog taškai iš svarbių sričių gali nepatekti į imtį.

Tikimybinės *stratifikuotos imtys* (*angl. stratified spatial sampling*) gaunamos padalijant sritį į nepersidengiančias dalis. Kiekvienoje dalyje požymio reikšmių variacija turi būti kuo mažesnė. Iš kiekvienos stratos elementai į imtį imami paprastuoju atsitiktiniu ėmimu, o imties elementų skaičius turi būti proporcingas stratos dydžiui.

Tikimybinės *klasterinės imtys* gaunamos atsitiktinai parenkant erdviniu atžvilgiu artimų taškų grupes (Dučinskas and Šaltytė-Benth 2003).

Jei pradinį tašką parinksime atsitiktinai, o likusius tam tikru apibrėžtu būdu, turėsime *sisteminę atsitiktinę imtį* (bus tikimybinė). Jeigu pradinis taškas parenkamas neatsitiktinai, o likę tam tikru apibrėžtu būdu, turėsime *taisyklingą* (vadinamą gardele) arba *determinuotą* (bus netikimybinė). Jei sritis padalijama į kvadratines stratas, kurių centruose parenkamas taškas, tada imtis bus netikimybinė ir vadinsis *centruota sistemine imtimi*.

1.3. Erdvinių populiacijų modeliai

Erdvinių populiacijų (duomenų) modelių analizėje dažnai taikomas tradicinis adityvusis duomenų modelis, atskiriantis determinuotą dalį nuo atsitiktinės:

Duomenys = erdvinis trendas + erdvinė sklaida.

Pirmoji variacijos komponentė yra pirmos eilės arba vidurkio variacija, kuri aprašoma n-mačiu vektoriumi μ. Antroji variacijos komponentė yra antros eilės variacija apie vidurkį μ, aprašoma kovariacine funkcija C.

Stebinio Z(s) modelis yra

$$Z(s) = \mu(s) + \varepsilon(s), \qquad (1.1)$$

kur $E(Z(s)) \equiv \mu(s)$ yra vidurkio funkcija arba erdvinis trendas, o $\varepsilon(s)$ - atsitiktinė paklaida arba erdvinė sklaida. Atsitiktinės paklaidos matematinis modelis srityje D yra atsitiktinis laukas $\{\varepsilon(s) : s \in D\}$ su nuliniu vidurkiu ir kovariacine funkcija C.

Labai svarbi sąvoka yra stacionarumas. Matematikoje stacionariais procesais yra vadinami stochastiniai procesai, kurių pasiskirstymai nepasikeičia pastū-

mus erdvėje ar laike. Arba kitaip sakant, pastūmus visas stebėjimo matavimų vietas ta pačia kryptimi ir tuo pačiu atstumu, proceso reikšmių skirstinys ir tikimybinės charakteristikos lieka nepakitusios.

1.1 Apibrėžimas. Atsitiktinis laukas $\{Z(s): s \in D\}$ vadinamas griežtai stacionariu (griežtai homogenišku), jei bendras stebinių $(Z(s_1), Z(s_2), ..., Z(s_n))$ skirstinys yra toks pat kaip ir $(Z(s_1+h), Z(s_2+h), ..., Z(s_n+h))$ skirstinys visiems baigtiniams *n*.

Jei $\{Z(s): s \in D\}$ yra griežtai stacionarus atsitiktinis laukas, tai $E\{Z(s)\} = \mu(s) = \mu$.

1.2 Apibrėžimas. Atsitiktinis laukas $\{Z(s): s \in D\}$ vadinamas stacionariu, jei:

$$E\left\{|Z(s)|^2\right\} < \infty \text{ visiems } s \in D,$$

$$E(Z(s)) \equiv \mu(s) \text{ visiems } s \in D,$$

$$C(s_1, s_2) = C(s_1 - s_2) \text{ visiems } s_1, s_2 \in D.$$

Griežtai stacionarūs laukai yra ir stacionarūs, tačiau priešingas tvirtinimas bendrai nėra teisingas. Tik Gauso atsitiktinių laukų stacionarumas reiškia griežtą stacionarumą, nes pastarieji yra pilnai aprašomi vidurkiu ir kovariacine funkcija.

Stacionaraus lauko koreliacijos funkcija apibrėžiama lygybe

$$R(s) = \frac{C(s)}{C(0)},$$

kur $C(s_1, s_2) = C(s_1 - s_2) = E\{(Z(s_1) - \mu(s_1))(Z(s_2) - \mu(s_2))\}$, $s_1, s_2 \in D$ - kovariacinė funkcija.

Kovariacinė funkcija, priklausanti nuo vektoriaus $h=s_1-s_2$ ilgio ir nepriklausanti nuo jo krypties, vadinama *izotropine*.

1.1 Lema. Stacionaraus lauko kovariacinė funkcija $C(s)$ tenkina šias lygybes:

$$C(0) \geq 0,$$

$$C(-s) = C(s) \text{ visiem } s \in D,$$

$$C(s) \leq C(0) \text{ visiems } s \in D.$$

Kovariacinė funkcija vadinama panašumo matu ir jos reikšmė didžiausia, kai atstumas tarp taškų yra nulinis.

Anot autoriaus Lopes H. F. (2008) labiausiai naudojamos koreliacinės funkcijos (KF) yra šios:

eksponentinė

$$r_\alpha(h) = \exp\{-|h| / \alpha\} = e^{-\frac{|h|}{\alpha}}, \tag{1.2}$$

Gausinė

$$r_\alpha(h) = \exp\{-h^2 / \alpha^2\} = e^{-\frac{h^2}{\alpha^2}},$$

kur h – atstumas tarp dviejų taškų, o α – koreliacijos plotis (*angl. correlation range*). Koreliacijos plotis dar vadinamas ilgio parametru (*angl. length parameter*) arba pločio parametru (*angl. range parameter*).

Matérn

$$r_{\alpha,\nu}(h) = \frac{1}{2^{\nu-1}\Gamma(\nu)}\left(2h\alpha\sqrt{\nu}\right)^\nu B_\nu\left(2h\alpha\sqrt{\nu}\right),$$

kur h, α apibrėžti aukščiau, o dydis ν – glodumo parametras (*angl. smoothness parameter*). Kai $\nu = 1/2$ => eksponentinė koreliacinė funkcija, o kai $\nu \to \infty$ => Gausinė koreliacinė funkcija (Lopes 2008). Dydis $B_\nu(.)$ – ν eilės modifikuota Besselio funkcija (*angl. modified Bessel function*) (Diggle and Ribeiro 2006).

Šiame darbe, skaitiniuose pavyzdžiuose, yra naudojama eksponentinė koreliacinė funkcija.

Taip pat šiame darbe yra naudojami skaliarinio atsitiktinio lauko $\{Z(s): s \in D\}$ pastovaus vidurkio modelis bei regresijos vidurkio modelis, kurie yra aprašyti sekančiose pastraipose.

Tarkime požymis Z yra stebimas srities D taškuose $s_1, s_2, ..., s_n \in D$. Stebiniai šiuose taškuose sujungti į vektorių-stulpelį

$$\mathbf{Z}_n = (\mathrm{Z}(\mathbf{s}_1), \mathrm{Z}(\mathbf{s}_2), ..., \mathrm{Z}(\mathbf{s}_\mathrm{n}))^T .$$

Tuomet

$$\mu_n = (\mu(s_1), \mu(s_2), ..., \mu(s_\mathrm{n}))^T$$

ir

$$\varepsilon_n = (\varepsilon(s_1), \varepsilon(s_2), ..., \varepsilon(s_\mathrm{n}))^T$$

yra atitinkamai vidurkių ir paklaidų vektoriai modelyje (1.1).

Skaliarinio atsitiktinio lauko $\{Z(s): s \in D\}$ vidurkio modeliai yra dviejų tipų: pastovaus vidurkio ir nepastovaus vidurkio (trendo paviršiaus, regresinis).

Pastovaus vidurkio atveju vidurkio funkcija yra pastovi, t.y. $\mathrm{E}\{\mathbf{Z}(\mathbf{s})\} = \mu =$ *const.* Tuomet $\mathbf{Z}_\mathrm{n}$ vidurkių vektorius bus

$$\boldsymbol{\mu}_n = \mu \mathbf{1}_\mathbf{n},$$

kur $\mathbf{1}_\mathbf{n} = (1, 1, ..., 1)^T$ – vienetinis n-matis vektorius.

Trendo paviršiaus vidurkio modelis – tai vidurkio, priklausančio nuo stebinio taško koordinačių, modelis. Taškui $s_i = (x_i, y_i) \in D$, $i = 1, 2, \ldots, n$, šis modelis užrašomas tokiu būdu:

$$\mu(s_i) = \sum_{\substack{t+u \le k \\ 0 \le t,u \le k}} \lambda_{tu} x_i^t y_i^u ,$$

kur λ_{tu} yra siūlomo modelio parametras, o k apibrėžia trendo paviršiaus eilę. Kai k=0, turime pastovaus vidurkio atvejį; kai k=1, tiesinis (pirmos eilės) atvejis generuoja plokštumą, k=2, kvadratinis (antros eilės) ir t.t.

Turint *n* stebinių, vidurkių vektorių galima užrašyti taip:

$$\mu_n = A\lambda ,$$

kur *A* yra taškų s_1, s_2, …,s_n koordinačių atitinkamų sandaugų matrica

$$A = \begin{pmatrix} 1 & x_1 & y_1 & x_1^2 & y_1^2 & x_1 y_1 & \cdots & x_1^t y_1^u & \cdots \\ 1 & x_2 & y_2 & x_2^2 & y_2^2 & x_2 y_2 & \cdots & x_2^t y_2^u & \cdots \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ 1 & x_n & y_n & x_n^2 & y_n^2 & x_n y_n & \cdots & x_n^t y_n^u & \cdots \end{pmatrix},$$

o λ – trendo paviršiaus parametrų vektorius, t.y. $\lambda = \left(\lambda_{00}, \lambda_{10}, \lambda_{01}, \lambda_{20}, \lambda_{02}\lambda_{11}, \ldots, \lambda_{tu}\right)^T$.

Regresiniame modelyje vidurkis taške $s_i \in D,\ i = 1,2,...,n$, apibrėžiamas kaip funkcija nuo *q* aiškinamųjų kintamųjų (regresorių):

$$\mu\left(s_i\right) = x^T\left(s_i\right)\beta ,$$

kur $x^T\left(s_i\right) = \left(1, x_1\left(s_i\right), x_2\left(s_i\right), ..., x_q\left(s_i\right)\right)$ yra (q+1)–matis regresorių vektorius, o $\beta = \left(\beta_0, \beta_1, \beta_2, \ldots, \beta_q\right)^T$ yra (q+1)–matis regresijos parametrų vektorius. Tokiu būdu *n* stebinių vidurkių vektorius bus:

$$\mu_n = X\beta ,$$

kur *X* yra n×(q+1) eilės matrica:

$$X = \begin{pmatrix} 1 & x_1\left(s_1\right) & x_2\left(s_1\right) & \ldots & x_q\left(s_1\right) \\ 1 & x_1\left(s_2\right) & x_2\left(s_2\right) & \ldots & x_q\left(s_2\right) \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ 1 & x_1\left(s_n\right) & x_2\left(s_n\right) & \ldots & x_q\left(s_n\right) \end{pmatrix}.$$

Dydis $x_j\left(s_i\right)$ yra *j*-tojo aiškinamojo kintamojo stebinys taške s_i, i= 1,2,...,n, j= 1,2,...,q (Dučinskas and Šaltytė-Benth 2003).

Erdvinių duomenų variacijai aprašyti naudojamos kovariacinės funkcijos arba semivariogramos. Pastarosios ypatingą vietą užima geostatistikoje, kur taikomos krigingui. Praktikoje naudojant semivariogramų modelius, galima nustatyti erdvinių duomenų priklausomybę. Turimų duomenų pagrindu suda-

roma empirinė semivariograma, kuriai parenkamas ir priglodinamas vienas iš semivariogramos parametrinių modelių. Dažniausiai naudojami semivariogramų modeliai aprašomi keliais semivariogramos parametrais, kurių pagrindiniai yra grynuolis, slenkstis ir slenksčio atstumas. Semivariograma yra vadinama „nepanašumo“ matu, o kovariacinė funkcija – „panašumo“ matu. Dar ji vadinama pusvariograme ir žymima γ(h).

Tai funkcija, kuri aprašo mažėjančią koreliaciją tarp imties stebėjimų porų, kai jų atsiskyrimas (tarpusavio atstumas) didėja (Liu and Mason 2009). Semivariograma γ(h).yra funkcija, aprašanti kintamųjų arba procesų erdvinės priklausomybės laipsnį ir bendrai apibrėžiama šitaip:

$$\gamma(\mathbf{h}) = \frac{1}{2n}\sum_{i=1}^{n}\left[z\left(x_i\right) - z\left(x_i + h\right)\right]^2$$

kur n yra vertinamų stebėjimų porų skaičius, z – stebėjimo reikšmė, h – atstumas tarp dviejų stebėjimų, o x_i atitinka lyginamų taškų poziciją.

1.4. Erdvinių sąryšių struktūros populiacijoje

Taikant įvairius metodus svarbu žinoti, kaip taškai yra susieti tarpusavyje. Jiems susieti yra du metodai: vektorinis ir teseliacinis. *Vektoriniame* modelyje pagrindinis loginis vienetas – linija, o *teseliaciniame* – erdvės dalis. Vektoriniame modelyje sritys modeliuojamos kaip daugiakampiai, kurių briaunos apibrėžiamos linijomis (loginiais vienetais). Teseliacinis modelis yra sudarytas iš ląstelių, suskaidančių plokštumą, visumos. Teseliacijoje erdviniai ryšiai tarp loginių vienetų yra apibrėžti netiesiogiai.

Kai stebinių taškai, sritys pasklidusios netaisyklingai, gaunasi sistema netaisyklinga. Tad norint suvokti, kaip turi būti aprašoma erdvinė priklausomybė, reikia apibrėžti ryšius tarp atskirų sričių tokiose netaisyklingose sistemose. Tarkime, yra atlikti stebėjimai fiksuotuose taškuose $s_1,...,s_n$. Tiriamoje srityje sudaromas grafas, kurio tikslas yra nustatyti ryšius tarp taškų. Čia taškai reiškia nagrinėjamo grafo viršūnes. Taigi pradžioje reikia parinkti briaunų struktūrą, kuri apibrėžia kiekvieno taško kaimynus ir priskirti briaunoms skaitines reikš-

mes. Maksimalus skirtingų taškų porų, o tuo pačiu ir briaunų, skaičius bus n(n+1)/2 (gali būti ir mažesnis, nes ryšiai nebūtinai egzistuoja tarp visų porų). Sakoma, kad ryšys yra, jei tarp dviejų grafo viršūnių yra briauna (Dučinskas and Šaltytė-Benth 2003).

Yra nemažai kriterijų, kurie padeda parinkti tinkamą briaunų struktūrą. Vienas iš tokių kriterijų būtų tiesioginio ryšio kiekybinis, kuris remiasi artimiausių kaimynų skaičiumi *k* (*k*=1,2,3,…). Artimiausių kaimynų skaičius, apibrėžtas iš anksto, parodo, su keliais artimiausiais kaimynais ieškome ryšių.

Taigi ir šiame darbe, 2 skyrelyje, vaizdo atstatymas, klasifikavimo su mokymu metodu, remiasi artimiausių kaimynų skaičiumi. Todėl sekančiose pastraipose yra aptariamos kaimynystės sistemos.

Tam tikroje srityje *D* stebėjimo vietos yra susijusios viena su kita per kaimynystės sistemą, kur kaimynystės sistema sričiai *D* yra apibrėžiama kaip

$$\mathcal{N} = \left\{ \mathcal{N}_i \middle| \forall i \in D \right\},$$

kur $\mathcal{N}_i$ yra s_i kaimyninių stebėjimo taškų aibė (Winkler 2006). Kaimyninė priklausomybė pasižymi sekančiomis savybėmis:

1. Lokacija nėra kaimynė pati sau: $s_i \notin \mathcal{N}_i$.
2. Kaimynystės ryšys yra abipusis: $s_i \in \mathcal{N}_{i'} \Leftrightarrow s_{i'} \in \mathcal{N}_i$) (Li 2009).

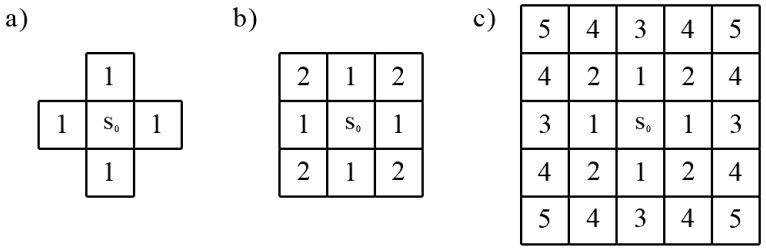
Reguliariai gardelei *D*, lokacijos s_i kaimynų aibė yra apibrėžiama, kaip aibė stebėjimų taškų nutolusių nuo lokacijos *si* per spindulį $\sqrt{r}$

$$\mathcal{N}_i = \left\{ s_{i'} \in D \middle| \left[dist\left(s_i, s_{i'}\right) \right]^2 \le r,\ s_{i'} \ne s_i \right\}$$

kur *dist*(*A, B*) yra Euklidinis atstumas tarp *A* ir *B*, o *r* yra sveikasis skaičius. Taip pat reikėtų atsižvelgti, jog lokacijos arčiau krašto turi mažiau kaimynų (Li 2009).

Pirmos eilės kaimynystės sistemoje, dar vadinamoje 4-ių kaimynų sistema, kiekviena lokacija (vidinė) turi po keturis kaimynus, kaip pavaizduota pav. 1 a) čia s_0 žymi lokaciją, o vienetai šios lokacijos kaimynus. Antros eilės kaimynystės schemoje, vadinamoje 8-ių kaimynų sistema, kiekvienai lokacijai (vidinei)

yra po 8 kaimynus pav. 1 b). Numeriai $n = 1,\ldots,5$ pavaizduoti pav. 1 c) atitinka tolimiausias kaimynines lokacijas *n*-os eilės kaimyninėse sistemose.

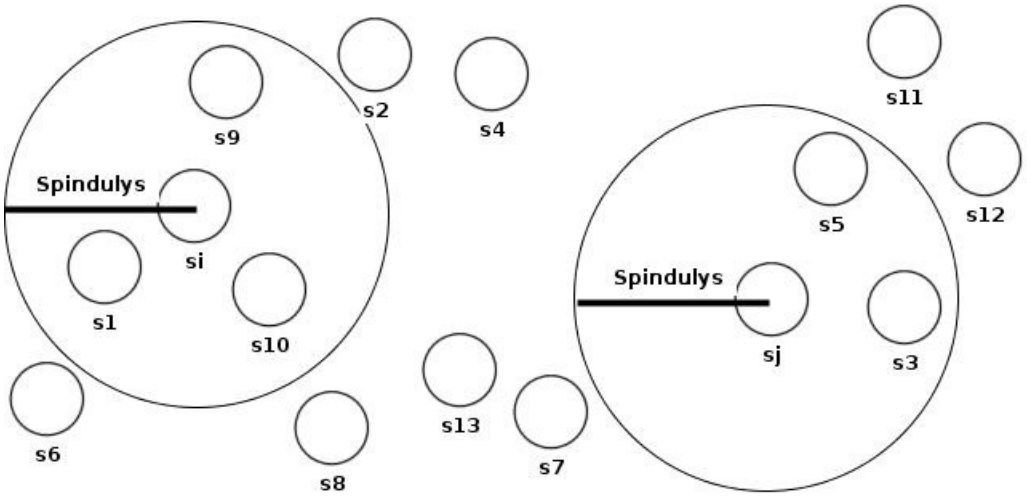


1.1 pav. a) 1-os eilės kaimynystės schema; b) 2-os eilės kaimynystės schema; c) 5-os eilės kaimynystės schema.

Kai yra apibrėžtas *D* stebinių vietų erdvėje eiliškumas, kaimynų aibė gali būti apibrėžta tiksliau. Pavyzdžiui, kai $D = \{s_1,\ldots,s_m\}$ yra sunumeruota stebinių vietų aibė ir jos elementai atitinka indeksuotus taškus 1D (vienmatis) paveiksliuko, vidinė lokacija $s_i \in \{s_2,\ldots,s_{m-1}\}$ turi du pirmos eilės artimiausius kaimynus $\mathcal{N}_i = \{s_{i-1}, s_{i+1}\}$, o kraštinės lokacijos turi tik po vieną $\mathcal{N}_1 = \{s_2\}$, $\mathcal{N}_m = \{s_{m-1}\}$. Analogiškai lokacijos reguliarioje kvadratinėje gardelėje $D = \{(s_i, s_j) | 1 \le i, j \le n\}$ atitinka $n \times n$ paveiksliuko taškus 2D (dvimatėje) erdvėje. Vidinė (ne paveiksliuko kraštuose esanti) lokacija (s_i, s_j) turi keturis artimiausius kaimynus $\mathcal{N}_{i,j} = \{(s_{i-1}, s_j), (s_{i+1}, s_j), (s_i, s_{j-1}), (s_i, s_{j+1})\}$ (Li 2009).

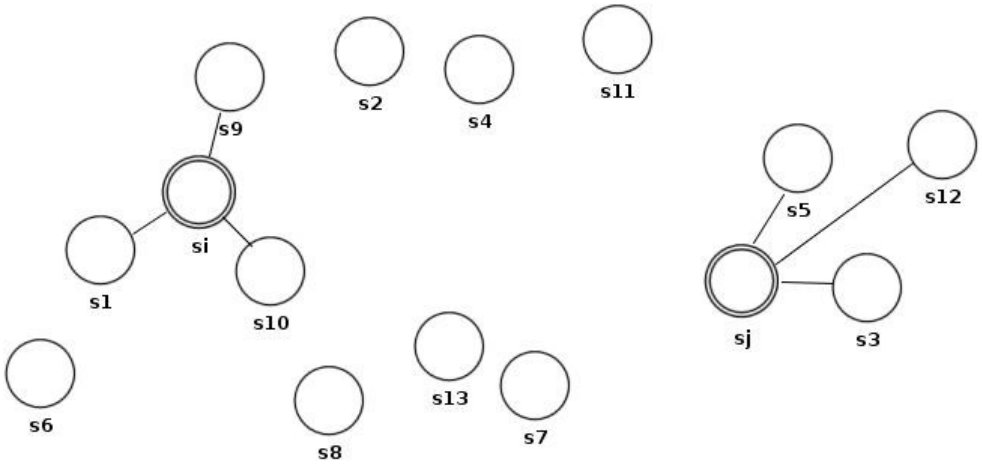
Esant nereguliariai sričiai *D*, lokacijos kaimynų aibė yra apibrėžiama analogiškai kaip ir reguliarios srities atveju, paimant visas lokacijas esančias spindulio $\sqrt{r}$ atstumu. Taip pat yra ir kitų kaimynystės sudarymo principų nereguliariai gardelei. Prie tam tikrų kaimynystės schemų, kaimynų aibės $\mathcal{N}_i$ nereguliariai *D* turi skirtingas formas ir dydžius. Nereguliarios lokacijos ir jų kaimynai pavaizduoti pav. 1.2. Kaimyninės sritys stebėjimo vietoms s_i ir s_j yra apvestos apskritimais. Šių dviejų kaimyninių aibių dydžiai yra $\#\mathcal{N}_i = 3$ ir $\#\mathcal{N}_j = 2$ (Li 2009).

Čia dydis r nėra įprastai naudojamo apskritimo spindulio žymėjimas. Tai yra dydis, nemažesnis už Euklidinio atstumo kvadratą tarp dviejų taškų. $\sqrt{r}$ atitinka Euklidinio atstumo tarp dviejų taškų ribą, iki kurios tie taškai laikomi kaimynais, pagal toje vietoje aprašomą kaimynystės schemą. Paveiksle 1.2 yra pateikta viena iš galimų kaimynystės schemų.



1.2 pav. Kaimynystės principas nereguliarios gardelės atveju.

Artimiausių kaimynų sistemos metodas remiasi tuo, jog iš visos D srities kiekvienai stebėjimo vietai s_i paimamas fiksuotas skaičius artimiausių lokacijų, kurios nuo lokacijos s_i nutolusios mažiausiu atstumu lyginant su kitomis lokacijomis. Pav. 1.3 pavaizduota lokacijų s_i ir s_j kaimyninės lokacijos sujungtos linijomis, šiame paveiksliuke naudojama trijų artimiausių kaimynų NN(3) schema. Šis principas taip pat gali būti taikomas ir reguliariai gardelei, kai konkrečiai lokacijai kaimynines lokacijas galima rinktis tik iš dalies visų reguliarios gardelės lokacijų, tai gali būti taikoma klasifikavime su mokymu, kada tame pačiame paveiksliuke yra keletas mokymo imties lokacijų (taškų) ir nežinomi taškai klasifikuojami tik pagal kaimyninius taškus iš mokymo taškų aibės. Toks principas taikomas šiame disertaciniame darbe atstatant sugadintą erdvėje koreliuoto triukšmo vaizdą.



1.3 pav. s_i ir s_j kaimynai remiantis NN3 kaimynystės schema.

1.5. Erdvinės statistikos taikymas vaizdų analizėje

Skaitmeninis vaizdas (*angl. digital image*) yra dvimatis skaičių masyvas. Kiekviena celė skaitmeninio vaizdo yra vadinama pikseliu ir skaičius nusakantis pikselių ryškumą yra vadinamas skaitmeniniu numeriu (*angl. digital number*) (Liu and Mason 2009). Paprastai vaizdai būna: 1) dvispalviai (*angl. binary*), kur pikseliui apibūdinti pakanka vieno bito, 2) pilkos skalės (*angl. gray-scale*), kur pikseliui apibūdinti naudojami aštuoni bitai, kurie leidžia atvaizduoti 256 pilkumo lygius, nuo 0 iki 255, 3) spalvoti (*angl. color*) vaizdai pateikiami žmogui matomo spektro diapazone, 4) daugiaspektriniai (*angl. multispectral*) vaizdai pateikiami ne tik žmogaus matomo spektro diapazone.

Tokių vaizdų analizė taikoma įvairiose srityse, tokiose kaip astronomija, gynyba, mikroskopija, Žemės stebėjimas iš palydovų, robotų technika ar medicina.

Kiekviena iš minėtų sričių yra individualios ir sudaro atskiras skaitmeninių vaizdų analizės sritis su daugybe specializuotų savo srities sąvokų ir algoritmų.

Analizuojant pavojingus arba sunkiai pasiekiamus regionus, informacija yra gaunama iš palydovo, įrašančio įrenginio, kuris neturi jokio fizinio kontakto su objektu. Tokių vaizdų analizei plačiai yra taikomi *erdvinės statistikos metodai*. Mat, čia gaunami erdviniai duomenys yra taisyklingos gardelės tipo, taip pat, palydovinėse nuotraukose Žemės paviršius yra padalijamas į mažus stačiakampius, vadinamus pikseliais (pikselyje esanti informacija paprastai sutelkta

centriniame to pikselio taške). Kiekvienas Žemės paviršiaus taškas skleidžia elektromagnetinę radiaciją. Elektromagnetinio spektro bangos ilgio intensyvumas priklauso nuo daugybės tokių faktorių, kaip paviršiaus tipas (dirbamas laukas, miškas, vandens telkinys, statybvietė ir kt.), paviršiaus sąlygos (sausas, drėgnas, ...), temperatūra, biologinis aktyvumas, radiacijos kampas ir kt. Palydovuose yra įmontuoti daugiaspektriniai skeneriai, kuriais matuojamas kelių skirtingų spektro juostų intensyvumas. Dėl tokių faktorių, kaip oro sąlygos, atmosferos užterštumas, skenerio iškraipymai ir pan., gauti spektro intensyvumo matavimų duomenys nėra visiškai tikslūs. Tokių duomenų analizei reikalingi gana sudėtingi tikimybiniai metodai, leidžiantys atsižvelgti į minėtus faktorius (Dučinskas Šaltytė-Benth 2003). Taigi analizuojant tokius vaizdus, kaip nuotraukas iš palydovų, svarbus vaidmuo tenka *erdvinės statistikos metodams*.

Erdvinės statistikos metodai gali būti taikomi bet kokio paveiksliuko atveju sprendžiant įvairiausias problemas, nes bet koks skaitmeninis paveiksliukas gali būti interpretuojamas kaip taškų stebinių matrica. Šią situaciją gerai iliustruoja paveikslėlis 1.4 RGB spalviniame režime.

RGB spalvinis režimas – spalvų perteikimo metodas, paremtas RGB spalvų modeliu (spalvinis modelis – tai spalvos parametrų, tokių kaip tonas, sodrumas, ryškis, aprašymo būdas).

Objektai, turintys kokią nors spalvą, gali ją sugerti arba atspindėti, todėl tokiems objektams aprašyti yra naudojami skirtingi spalvų modeliai.

RGB spalvinis modelis naudojamas prietaisuose, kurie spinduliuoja šviesą: televizoriuose, kompiuterių monitoriuose, šviestuvuose. Šis modelis remiasi pagrindinėmis spalvomis (spalvinėmis komponentėmis): raudona, žalia, mėlyna (*angl. Red*, *Green*, *Blue*).

Naudojant šį modelį yra priskiriamos reikšmės kiekvienai RGB komponentei nuo 0 iki 255. Šiais skaičiais yra matuojamas kiekvienos spalvos kiekis RGB modelyje (spalva turi turi 256 lygius). Kai visų spalvinių komponenčių (spalvinių kanalų) skaitinės reikšmės lygios, gaunama pilka spalva, kai visų komponenčių skaitinės reikšmės lygios 255 – balta spalva, kai visos skaitinės

reikšmės lygios 0 – juoda. Sumaišius visas tris spalvas skirtingomis proporcijomis, galima išgauti visą atspalvių įvairovę.

Maišant pagrindines spalvas, sukuriamos išvestinės spalvos. Maišant dvi pagrindines spalvas, gaunama šviesesnė spalva. Sumaišę raudoną ir žalią, gauname geltoną, žalią ir mėlyną – žydrą, mėlyną ir raudoną - purpurinę. Sumaišius visas tris pagrindines spalvas, gaunama balta spalva (vadinama adityvia).

Šiame darbe sugadinto vaizdo klasifikavimui naudojamas RGB spalvinio modelio paveikslėlis. Čia paveikslėlis yra iš dviejų klasių, kur pirmos klasės elementai – balti pikseliai, o antros klasės elementai – juodi pikseliai. Reikštų, kad baltos spalvos celė įgytų reikšmes: R-255, G-255, B-255. Juodos spalvos celė – R-0, G-0, B-0. Kadangi naudojamos tik dvi spalvos, tai pakanka RGB spalvinio modelio tik vieno kanalo. Baltos spalvos pikseliai turės reikšmes vienetus, o juodos spalvos pikseliai turės nulius.

Šiais laikais plačiai yra taikomas teksto atpažinimas, kada nuskaitmenintas tekstas perkeliamas į kompiuterį, ir paverčiamas tekstine informacija. Programos naudoja tam tikrus metodus teksto atpažinimui, tačiau ir šioje srityje yra naudojami erdvinės statistikos metodai. Viena iš problemų, su kuria susiduriama, yra tokia, jog tekstas ne visada būna ryškus ant realios tekstinės informacijos (raidžių paveiksliukų). Atsiradusius triukšmus galime interpretuoti kaip erdvėje koreliuotą triukšmą ε (modelyje (1.1)), o realią raidės taškų informaciją – kaip vidurkių modelį. Taigi, tokiems vaizdams galima taikyti erdvinio klasifikavimo metodus, ir, pašalinus triukšmą išgauti švarią vidurkių informaciją. Taip pertvarkytas paveiksliukas yra lengviau atpažįstamas teksto atpažinimo (OCR) programų (Gupta *et al*. 2005), (Gupta *et al*. 2009).

Tą patį principą galima taikyti ir palydovinės informacijos klasifikavimui, kai ant realios palydovinės nuotraukos atsiranda triukšmas, trukdantis tinkamai klasifikuoti informaciją. Tai gali nutikti dėl tam tikrų gamtoje vykstančių reiškinių, tokių kaip rūkas, debesys ar gaisro metu atsiradę dūmai. Tokią situaciją iliustruoja paveiksliukas (pav. 1.5).

Gaisro metu kylantys dūmai gali būti interpretuojami kaip erdvėje koreliuotas laukas, kurį galima modeliuoti Gauso atsitiktiniu lauku.

$$= R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1m} \\ r_{21} & r_{22} & \cdots & r_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nm} \end{pmatrix}, \; r_{i,j} \in [0,\ldots,255]$$

$$= G = \begin{pmatrix} g_{11} & g_{12} & \cdots & g_{1m} \\ g_{21} & g_{22} & \cdots & g_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ g_{n1} & g_{n2} & \cdots & g_{nm} \end{pmatrix}, g_{i,j} \in [0,\ldots,255]$$

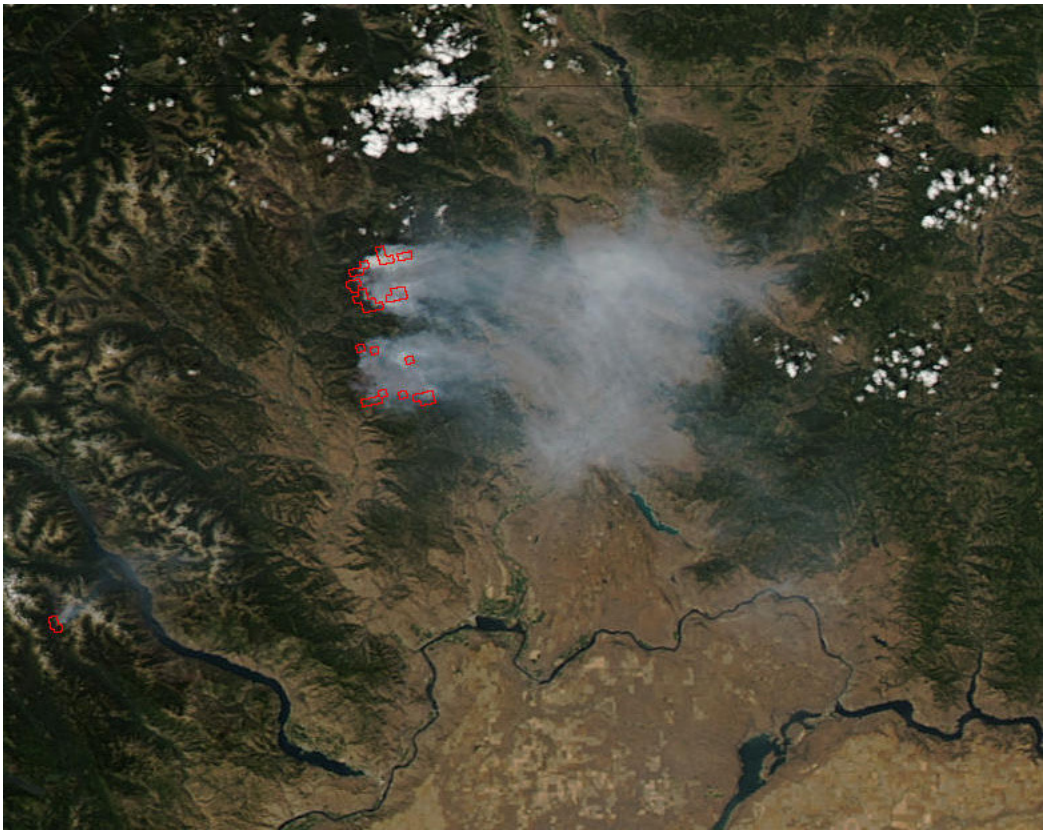
$$= B = \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1m} \\ b_{21} & b_{22} & \cdots & b_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nm} \end{pmatrix}, b_{i,j} \in [0,\ldots,255]$$

1.4 pav. Vaizdas RGB spalviniame režime

Tie taškai, kurie požymių erdvėje yra arčiau klasifikuojamo taško Z_0, jie turi didesnę tikimybę būti tos pačios klasės taškais, nei tie taškai, kurie yra toliau požymių erdvėje (Atkinson 2004), (Fix and Hodges 1951), (Liu 2009), (Dudani 1976). Pavyzdžiui, atsitikus tam tikrame objekte avarijai, išsiliejus teršalams ir pasklindus į aplinką, bus taip, kad, kuo arčiau to objekto, tuo didesnis teršalų susitelkimas ir išmatuotos reikšmės labiau panašios (koreliuos), o tolstant nuo objekto tam tikru atstumu reikšmės mažės ir susitelkimo (koreliuotumo) neliks. Prie tam tikro atstumo priklausomybė išliks. Išmatuotieji stebinio požymiai Z yra aprašomi modeliu (1.1), į kurį įeina požymio vidurkis $\mu(s)$ ir triukšmas $\varepsilon(s)$. Pastarasis dydis $\varepsilon(s)$ yra aprašomas kovariacine funkcija (Dučinskas and

Šaltyte-Benth 2003) ir jis yra paklaidų vektorius, o paklaidos dažnai būna pasiskirsčiusios normaliai, todėl jį ir modeliuojame kaip erdvėje koreliuotą Gauso atsitiktinį lauką.

Tiek debesys, tiek dūmai yra susitelkę (koreliuoja). Jei bus paveikslėlis sudarytas iš miško ir pievos teritorijos, tai taškai atitinkantys pievą turės aukštesnes reikšmes (didesnis intensyvumas), o miškas – žemesnes. Ir kai ant miško teritorijos užplauks debesėlis, jis miško reikšmes padarys aukštesnes. Tad tokį triukšmą modeliuojant nereikėtų ignoruoti erdvinės koreliacijos.



1.5 pav. Vaizdas iš NASA MODIS palydovo (Schmaltz 2006). Gaisro metu teritoriją dengia dūmai, pasižymintys erdvėje koreliuotu triukšmu

Taip pat labai svarbu tai, jog analizuojant vaizdus galima taikyti ir Markovo atsitiktinių laukų savybes bei modelius.

Vaizdų analizėje dažnai taikomi Markovo atsitiktiniai laukai (MRF) (*angl. Markov random field*) (Winkler 2006). Markovo atsitiktinių laukų teorija yra tikimybių teorijos šaka analizuojanti erdvines arba kontekstines fizikinių po-

žymių priklausomybes. Ši teorija yra naudojama vizualiniam žymėjimui tam, kad nustatyti sąveikaujančių žymių tikimybinius skirstinius.

Tegul $F = \{F_1,\ldots,F_m\}$ yra šeima atsitiktinių dydžių apibrėžtų aibėje *D*, kurioje kiekvienas atsitiktinis dydis F_i įgyja reikšmę iš $\mathcal{L}$. Taip apibrėžta šeima *F* yra vadinama atsitiktiniu lauku. Toliau bus naudojami žymėjimas $F_i = f_i$ pažymėti įvykį kai F_i priskiriama reikšmė f_i ir žymėjimas $(F_1 = f_1,\ldots,F_m = f_m)$ pažymėti jungtinį įvykį. Paprastumui jungtinis įvykis yra supaprastinamas iki $F = f$, kur $f = (f_1,\ldots,f_m)$ yra *F* konfigūracija atitinkanti lauko realizaciją. Diskrečiai žymių aibei $\mathcal{L}$, tikimybė, jog atsitiktinis dydis F_i įgys reikšmę f_i žymima kaip $P(F_i = f_i)$ yra sutrumpinama iki $P(f_i)$. Jungtinė tikimybė yra žymima $P(F = f) = P(F_1 = f_1,\ldots,F_m = f_m)$ ir sutrumpinama iki *P*(*f*). Esant tolydžiai $\mathcal{L}$, naudojamos tikimybinės tankio funkcijos $p(F_i = f_i)$ ir $p(F = f)$ (Li 2009).

1.3 Apibrėžimas. *F* yra vadinamas Markovo atsitiktiniu lauku srityje *D* su kaimynystės sistema $\mathcal{N}$ tada ir tik tada, kai tenkinamos sekančios sąlygos:

$P(f) > 0,\ \forall f \in \mathbb{F}$ (teigiamumo),

$P\left(f_i \middle| f_{D-\{s_i\}}\right) = P\left(f_i \middle| f_{\mathcal{N}_i}\right)$ (Markoviškumo),

kur $D - \{s_i\}$ yra aibių skirtumas, $f_{D-\{s_i\}}$ žymi aibę žymių lokacijose $D - \{s_i\}$, ir $f_{\mathcal{N}_i} = \{f_{i'} | i' \in \mathcal{N}_i\}$ yra *si* lokacijai kaimyninių lokacijų žymių aibė.

$\mathbb{F}$ – visų galimų konfigūracijų aibė (galimi klasių žymių išsidėstymai (kombinacijos) apie klasifikuojamą stebinį). Pavyzdžiui, jei klasifikuojant naudosime 4 artimiausių kaimynų schemą, tai gali būti konfigūracija tokia: 3 stebiniai vienos klasės, 1 – kitos. Gali būti vienos ir kitos klasės po 2 stebinius. Taip pat galima situacija, kad vienos klasės stebinių bus 4, o kitos – 0.

Teigiamumo sąlyga yra naudojama dėl tam tikrų techninių priežasčių ir dažnai praktikoje gali būti išpildoma. Griežtas teigiamumas čia reiškia tai, jog visos įmanomos žymių realizacijos (konfigūracijos) yra galimos. Kiekviena konfigūracija (erdvinis žymių išsidėstymas), apie klasifikuojamą tašką Z_0, pagal tam tikrą kaimyninę schemą, yra galima (daugiau ar mažiau tikėtina). Nėra

tokio stebinių žymių išsidėstymo (Y=y), kurio negalėtų būti, todėl tikimybė visada didesnė už nulį.

Markoviškumo sąlyga apibrėžia F charakteristikas. MRF modeliuose tik kaimyninės žymės tiesiogiai įtakoja vienos kitą. Jei mes pasirinktume didžiausią įmanomą kaimynystės sistemą, kurioje kiekvienos lokacijos kaimynai yra visos likusios srities D lokacijos, tada bet kuris F yra MRF su tokia kaimynystės sistema (Li 2009).

Vaizdų klasifikavime dažnai laikoma, kad požymių stebinių žymės tenkina MRF modelį. Šiame darbe, trečiame skyriuje, tiriant klaidų tikimybių priklausomybę nuo statistinių parametrų reikšmių, yra laikoma, kad požymius tenkina stacionaraus Gauso atsitiktinio lauko modelis, o klasių žymės seka iš MRF modelio.

Markovo atsitiktinių laukų pritaikomumas grindžiamas tuomi, kad klasifikuojamam stebiniui Z_0, įtakos nedaro mokymo imties taškai, tie, kurie nėra susieti su Z_0 pagal tam tikrą kaimynystės schemą. MRF modelyje tik kaimyninės žymės tiesiogiai įtakoja vienos kitą (Li 2009) ir kai yra klasifikuojama pagal BD funkcijas, koreliacijas tarp mokymo imties elementų skaičiuojame ne visiems, bet tiems taškams, kurie su Z_0 yra susieti tam tikra kaimynystės schema. Pavyzdžiui, jei klasifikuojamą požymį Z_0, susiesiu su keturių artimiausių kaimynų schema, tai klasifikuojant, bus naudojami tik keturi, arčiausiai Z_0 esantys stebiniai.

1.6. Statistinis objekto atpažinimas

Ankstesniame skyrelyje aptarta erdvinės statistikos svarba vaizdų analizėje. Šiame skyrelyje dėmesys yra skiriamas siauresnei vaizdų analizės sričiai - statistiniam objekto atpažinimui (vaizdo, šablono, struktūros, raštų ir kt. atpažinimui).

Struktūros, raštų, šablonų ir kt. objektų klasifikavimas (*angl. pattern classification*) dažniau vadinamas objektų atpažinimu (*angl. pattern recognition*),

yra pagrindine kliūtimi automatizavimo užduotyse (Alder 2001), pavyzdžiui kuriant robotus su jutikliais.

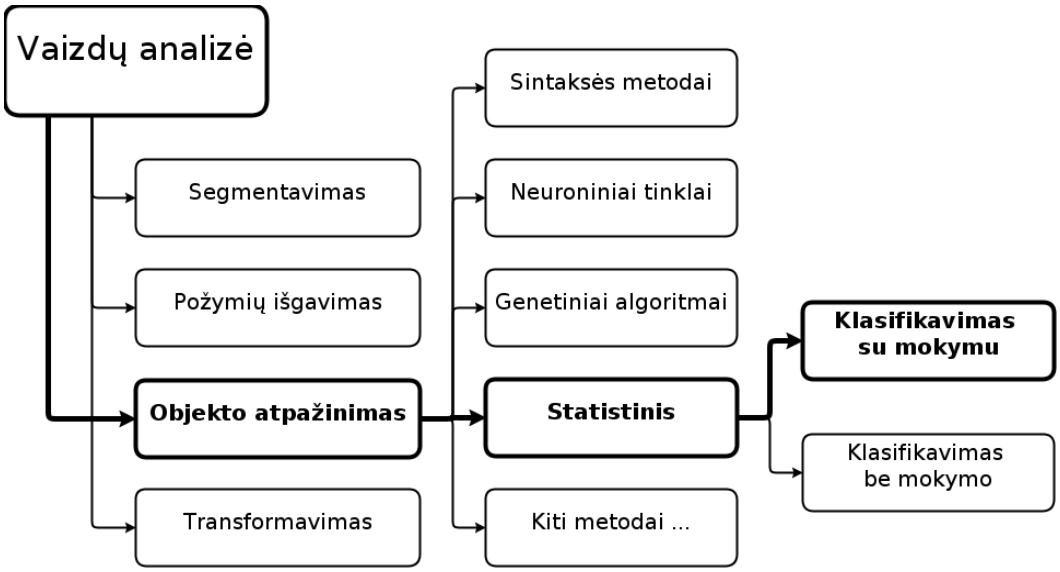
Žodis „*pattern*“ lietuvių kalboje verčiasi labai įvairiai, todėl nelengva padaryti tinkamą vertimą. Turbūt teisingiausia būtų jį sieti su kontekstu.

„*Pattern*“ – tai neaiškiai apibrėžtas objektas (objektyvioji realybė), kuriam galėtų būti suteiktas pavadinimas, pavyzdžiui, DNR seka, ranka rašytas žodis, žmogaus veidas, pirštų antspaudų vaizdas ir pan.

Taigi, pradžioje svarbu paminėti, kad vaizdų analizė (*angl. image analysis*) apima siauresnes sritis: 1) segmentavimas, 2) transformavimas, 3) požymių išgavimas (*angl. feature extraction*), 4) objekto klasifikavimas (*angl. pattern classification*) arba objekto atpažinimas (*angl. pattern recognition*).

Pastarasis (objekto atpažinimas) vėlgi skaidomas į siauresnes dalis: neuroninių tinklų, sintaksės metodų, genetinių algoritmų, statistinių metodų ir kt. Taigi, statistiniais metodais paremtas objekto atpažinimas (objekto klasifikavimas) yra vadinamas statistiniu objekto atpažinimu (*angl. statistical pattern recognition*).

Statistinis ir neuroninis objekto atpažinimas gali būti diferencijuojamas į atpažinimą su mokymu (*angl. supervised*) ir atpažinimą be mokymo (*angl. unsupervised*) (Bäse 2004).



1.6 pav. Vaizdų analizės metodai

Paveikslėlyje 1.6 pateikta vaizdų analizės metodų schema. Šioje schemoje storesnėmis rodyklėmis yra parodyta sritis, kuriai ir priklauso šiame darbe siūloma klasifikavimo metodika.

Lentelėje 1.1 yra pateikti objekto atpažinimo (klasifikavimo) taikymo pavyzdžiai.

1.1 Atpažinimo teorijos taikymo sritys

Taikymo sritis	*Taikymas*	*Pradinis objektas*	*Objekto klasės*
Dokumentų vaizdų analizė	Raidžių atpažinimas	Dokumento paveiksliukas	Raidės, žodžiai
Dokumentų klasifikavimas	Paieška internete	Tekstiniai dokumentai	Semantinės kategorijos
Daugialypės informacijos paieška	Paieška internete	Vaizdo filmai	Žanrai
Kalbos atpažinimas	Telefoninė pagalba	Kalbos garso įrašas	Sakomi žodžiai
Natūralios kalbos apdorojimas	Informacijos išgavimas	Sakiniai	Pokalbio dalys
Biometrinis atpažinimas	Asmens identifikavimas	Veidas, rainelė, pirštų antspaudai	Sistemos vartotojai
Medicina	Kompiuterinė diagnostika	Mikroskopu gauti vaizdai	Vėžinės/sveikos ląstelės
Kariuomenė	Automatinis taikinio atpažinimas	Optinis arba infraraudonas paveiksliukas	Taikinio tipai
Industrinė automatika	Vaisių rūšiavimas	Konvejerio paveiksliukai	Kokybės klasės
Nuotolinis stebėjimas	Javų derliaus prognozavimas	Daugiaspektrinis paveiksliukas	Žemės paviršiaus klasės
Bioinformatika	Sekos analizė	DNR seka	Žinomi genai
Duomenų gavyba	Reikšmingų objektų paieška	Taškai daugiamatėje erdvėje	Gerai atskirti klasteriai

(Aksoy 2011)

Kaip buvo minėta statistinis objekto atpažinimas yra dviejų tipų: klasifikavimas su mokymu (SCL) ir klasifikavimas be mokymo, kurie ir bus toliau aptariami.

SCL metodas yra paremtas statistika mokymo sričių (*angl. training areas*), atstovaujančių skirtingiems, subjektyviai pasirinktiems pagrįstiems objektams

(Liu and Mason 2009). Čia klasifikavimo funkcijos parametrai nustatomi pagal mokymo imties informaciją.

Mokymo imtis (TS) – informacija apie dalies tiriamų objektų požymių reikšmes ir priklausomybę vienai ar kitai grupei. Apibrėžiama tokiu būdu:

$$T' = \left(Z', Y'\right),$$

kur $Y = (Y(s_1),...,Y(s_n))'$ – žymių vektorius, $Z = (Z(s_1),...,Z(s_n))'$ – požymių vektorius (Čekanavičius and Murauskas 2008).

Mokymo imtis susideda iš aibės mokymo pavyzdžių. Klasifikavime su mokymu kiekvienas mokymo imties elementas yra pora, sudaryta iš stebinio, kurį dažniausiai sudaro stebinių vektorius, ir klasės, kuriai priklauso stebinys, žymės.

Klasifikavimui su mokymu labai dažnai naudojama diskriminantinė analizė ir diskriminantinės funkcijos.

Pirmasis iš autorių, kuris pradėjo taikyti tiesines diskriminantines funkcijas su erdvine priklausomybe buvo Paul Switzer (1980). Autorius savo darbe naudojo diskriminantines funkcijas su erdvine priklausomybe klasifikuojant skirtingas teritorijų klases pagal daugiaspektrinių palydovinių nuotraukų informaciją. Iki tol klasifikuojant daugiaspektrinių palydovinių nuotraukų informaciją būdavo naudojami tik klasifikuojamos lokacijos s_0 (taško) stebinių vektorius $Z(s_0)$ (Switzer 1980). Autorius papildomai įveda priklausomybę nuo klasifikuojamo taško ir jam kaimyninių taškų.

Klasifikavimo su mokymu algoritmai išanalizuoja mokymo informaciją ir suformuoja funkciją, kuri naudojama klasifikavimui. Klasifikavimo funkcija privalo prognozuoti bet kuriam naujam stebiniui, atitinkančiam funkcijos apibrėžimo sritį.

Klasifikacija su mokymu yra kontroliujama pagal vartotojų žinias, bet kita vertus yra ribota ir netgi gali būti neobjektyvi pagal jų subjektyvią nuomonę. Todėl klasifikacija gali būti nevykusi pagal netinkamą arba netikslią informaciją mokymo imtyje ir/arba neišsamias vartotojo žinias (Liu and Mason 2009).

Klasifikavimo be mokymo metodai yra naudojami tada, kai klasės nėra apibrėžtos iš anksto arba kada jos yra, bet duomenys naudojami tam, kad patvirtintų jog jie yra tinkamų klasių. Pastarojo tipo pavyzdžiai yra gana dažni biologijoje, kur rūšys dažnai apibrėžiamos fizinėmis savybėmis, ir biocheminių matavimų duomenų rinkiniai tampa prieinami.

Šie metodai paprastai yra skirti vizualizacijai (*angl. visualization*), tačiau kartais naudojami klasifikuoti (Ripley 1996).

Klasifikavimas be mokymo stipriai susijęs su pasiskirstymo įvertinimu statistikoje, tačiau klasifikavimas be mokymo taip pat remiasi daugeliu kitų technikų, kurių taikymo metu stengiamasi nustatyti pagrindinius duomenų požymius. Pagal nustatytus duomenų požymius atitinkami objektai priskiriami atitinkamai klasei. Vieni iš dažnai taikomų tokio tipo principų remiasi pasikartojimo matricomis (*angl. co-occurrence matrix*), pagal kurias nustatomi vaizdų požymiai.

1.7. Pirmojo skyriaus apibendrinimas ir disertacijos uždavinių formulavimas

- Taikant erdvinės statistikos metodus, būtina žinoti duomenų išsidėstymą erdvėje. Vienas esminių bruožų, išskiriančių erdvinę statistiką iš klasikinės, yra tas, jog ji naudojama modeliuoti ne tik erdviniam trendui, bet ir erdvinei koreliacijai.
- Erdvinių duomenų požymio Z modelį sudaro erdvinio proceso vidurkio ir klaidų dedamosios. Dažniausiai išskiriami pastovaus ir nepastovaus vidurkio modeliai.
- Duomenų variacija aprašoma kovariacine funkcija arba semivariograma. Pastarosios dažnai naudojamos praktiniuose skaičiavimuose, kur turimų duomenų pagrindu sudaroma empirinė semivariograma. Jai parenkamas ir priglodinamas vienas iš semivariogramos parametrinių modelių. Tokiu būdu galima įvertinti erdvinės koreliacijos pločio parametrą α.

- Erdvinės statistikos metodai gali būti taikomi bet kokio paveiksliuko atveju, sprendžiant įvairiausias problemas, nes bet koks skaitmeninis paveiksliukas gali būti interpretuojamas kaip taškų stebinių matrica.
- Vaizdus, kurie yra sugadinti tokių reiškinių, kaip dūmai, debesys, rūkas, galima modeliuoti Gauso atsitiktiniais laukais. Tokius triukšmus galima interpretuoti, kaip erdvėje koreliuotą triukšmą ε, modelyje (1.1), o realią vaizdo taškų informaciją – kaip vidurkių modelį.
- Esminis disertacijos uždavinys – pasiūlyti klasifikavimo su mokymu metodiką, paremtą BDF, naudojamą klasifikavimui vaizdų, sugadintų su erdvėje koreliuotu triukšmu.
- Ištirti artimiausių kaimynų skaičiaus įtaką vaizdo klasifikavimo kokybei bei skaitiškai panagrinėti BDF klaidų tikimybių priklausomybes nuo tam tikrų statistinių parametrų reikšmių.

2

Vaizdų, modeliuojamų GRF, klasifikavimo metodai

Skyrelyje apžvelgiama Bajeso sprendimo teorija, diskriminantinės funkcijos, klasifikavimo vertinimas. Pagrindinis dėmesys skiriamas klasifikavimo su mokymu metodams, paremtiems Bajeso diskriminantinėmis funkcijomis. Į klasifikavimo problemą yra įvedama erdvinė priklausomybė. Požymių stebiniai priklausomi ir tenkina GRF modelį, o klasių žymės – diskretaus lauko modelį. Taip pat pateiktos klaidų tikimybių išraiškos Bajeso diskriminantinėms funkcijoms.

Skyriaus tematika yra paskelbti keturi autorės straipsniai [1A], [3A], [5A], [6A].

2.1. Bajeso sprendimo teorija

Bajeso sprendimo teorija atstovauja fundamentinį statistinį metodą objekto klasifikavimo problemai spręsti. Ši technika yra grindžiama prielaida, kad sprendimo problema yra suformuluota tikimybinėse sąlygose ir kad visos atitinkamos tikimybių reikšmės yra pateiktos.

Paprasta šio metodo apžvalga gali būti pateikta sutelkiant dėmesį į dviejų klasių atvejį Ω_1, Ω_2. Daroma prielaida, kad apriorinės tikimybės $P(\Omega_1)$ ir $P(\Omega_2)$ bus žinomos, nes jos gali būti lengvai nustatomos iš turimų duomenų aibės. Taip pat yra žinomos tikimybinio tankio funkcijos $p(x_i|\Omega_i)$, $i=1,2$. Funkcija $p(x_i|\Omega_i)$ taip pat vadinama tikėtinumo funkcijos vardu.

Prisimindami Bajeso taisyklę, turime

$$P\left(\Omega_i|\mathbf{x}\right)=\frac{p\left(\mathbf{x}|\Omega_i\right)P\left(\Omega_i\right)}{p\left(\mathbf{x}\right)},$$

kur $p\left(\mathbf{x}\right)$ yra $\mathbf{x}$ tikimybinė tankio funkcija, o $\mathbf{x}$ - *n*-dimensinis požymio vektorius (*angl. feature vector*). Tad $p\left(\mathbf{x}\right)$ turės pavidalą

$$p\left(\mathbf{x}\right)=\sum_{i=1}^{2}p\left(\mathbf{x}|\Omega_i\right)P\left(\Omega_i\right).$$

Bajeso klasifikavimo taisyklė dabar gali būti parodyta dviejų klasių atveju Ω_1, Ω_2.

Jeigu, $P\left(\Omega_1|\mathbf{x}\right) > P\left(\Omega_2|\mathbf{x}\right)\Rightarrow \mathbf{x}\in\Omega_1$.

Jeigu, $P\left(\Omega_1|\mathbf{x}\right) < P\left(\Omega_2|\mathbf{x}\right)\Rightarrow \mathbf{x}\in\Omega_2$.

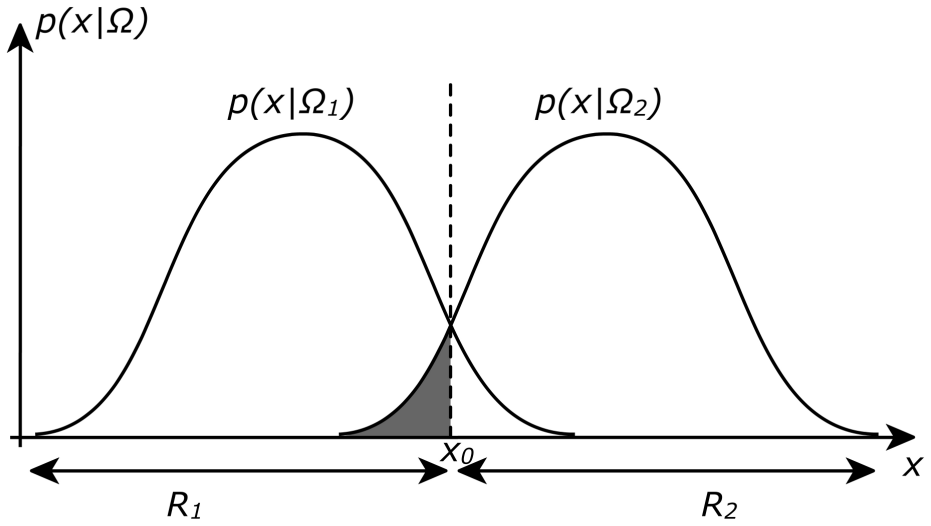
Remiantis aukščiau pateiktomis išraiškomis, galima daryti išvadą, kad požymio vektorius gali būti priskirtas vienai arba kitai klasei. Analogiškai galima užrašyti

$$p\left(\mathbf{x}|\Omega_1\right)P\left(\Omega_1\right) > (<)\ p\left(\mathbf{x}|\Omega_2\right)P\left(\Omega_2\right)\Rightarrow \mathbf{x}\in\Omega_1\ \left(\mathbf{x}\in\Omega_2\right).$$

Tai atitinka sąlyginių tikimybinių tankio funkcijų nustatytą maksimumą, įvertintą pagal $\mathbf{x}$. Paveiksle 2.1 yra pateikta vienodai tikėtinos klasės ir sąlyginės tikimybinės tankio funkcijos $p\left(x|\Omega_i\right)$, $i=1,2$, kaip *x* funkcijos. Punktyrinė linija ant x_0 atitinka ribą, dalijančią vienmatę požymių erdvę į du regionus R_1, R_2. Remiantis Bajeso klasifikavimo taisykle, visos reikšmės $x\in R_1$ yra priskiriamos klasei Ω_1, o visos reikšmės $x\in R_2$ yra priskiriamos klasei Ω_2 (Bäse 2004).

Sprendimo klaidos tikimybė yra apskaičiuojama pagal formulę

$$P_e=\int_{-\infty}^{x_0}p\left(x|\Omega_2\right)dx+\int_{x_0}^{+\infty}p\left(x|\Omega_1\right)dx.$$



2.1 pav. Dvi vienodai tikėtinos klasės atitinkančios regionus R_1, R_2.

Bajeso klasifikavimo taisyklė pasiekia minimalią klaidos tikimybę. Žinoma, kad klasifikavimo klaida minimali, jei požymių aibė pasiskirsčiusi į du regionus R_1 ir R_2 taip, kad

$$R_1 : P\left(\Omega_1 \middle| \times\right) > P\left(\Omega_2 \middle| \times\right)$$

$$R_2 : P\left(\Omega_2 \middle| \times\right) > P\left(\Omega_1 \middle| \times\right) \ .$$

Apibendrinimas M klasių Ω_1, Ω_2, ..., Ω_M, yra paprastas. Požymių vektorius $\times$ yra priskiriamas klasei Ω_i, jei

$$P\left(\Omega_i \middle| \times\right) > P\left(\Omega_j \middle| \times\right) \quad \forall j \neq i \, . \tag{2.1}$$

Kiekvieną kartą priskiriant objektą į klasę yra rizika suklysti, o daugelio klasių atveju klaidingas klasifikavimas gali turėti rimtesnių pasekmių. Kiekybiniu būdu galima tai įvertinti, apskaičiuojant pagal formulę, vadinamą nuostolių funkcija. Tegul $L(i, j)$ „nuostolis“ priskiriant objektą į klasę i, kai iš tiesų priklauso klasei j.

Iš to, kas pasakyta, matoma, kad skirtinga klasifikavimo galimybė yra pasiekta apibrėžiant nuostolių funkciją $L(i, j)$ su i, j=1,2,…,M. $L(i, j)$=0, jei požymio vektorius $\times$ teisingai priskirtas į klasę, ir didesnė už nulį $L(i, j)$>0, jei $\times$ yra priskirtas į klasę neteisingai.

Sąlyginis nuostolių narys $R_i(\times)$ yra apibrėžiamas

$$R_i(\times) = \sum_{j=1}^{M} L(i,j)\, P\left(\Omega_j \middle| \times\right)$$

arba ekvivalenčiai,

$$R_i(\times) = \sum_{j=1}^{M} L(i,j)\, p\left(\times \middle| \Omega_j\right) P\left(\Omega_j\right).$$

Remiantis ankstesniais apibrėžimais, gauname nežymiai pakeistą Bajeso klasifikavimo taisyklę (BCR): požymių vektorius × priskiriamas klasei Ω_i, kuriai $R_i(\times)$ minimalus (Bäse 2004).

2.2. Diskriminantinės funkcijos

Apibrėžiant objekto klasifikatorių galimybių yra įvairių. Vienas metodas, kuris gali būti laikomas kaip klasifikatorių kanoninė forma yra vadinamas diskriminantinėmis funkcijomis (DF). *M* klasių atveju jos yra naudojamos požymių erdvės padalijimui. Daugelyje situacijų yra paprasčiau spręsti su lygiaverčia tikimybine funkcija nei dirbti tiesiogiai su tikimybėmis, pavyzdžiui $g_i(\times) = f\left(P\left(\Omega_i \middle| \times\right)\right)$, kur $f(\cdot)$ yra monotoniškai didėjanti funkcija. $g_i(\times)$ yra vadinama diskriminantine funkcija. Remiantis ankstesniais sąryšiais, formulei (2.1) gauname sekančią ekvivalenčią išraišką arba, paprasčiau, sprendimo taisyklę: × priklauso klasei Ω_i, jei

$$g_i(\times) > g_j(\times) \quad \forall j \neq i\,. \tag{2.2}$$

Darant prielaidą, kad R_1, R_2 yra kaimyninių regionų, tada jie gali būti atskirti hyperplokštuma daugiamatėje erdvėje (*angl. multidimensional space*). Formulė, apibrėžianti atskyrimo plokštumą yra:

$$g_{ij}(\times) = g_i(\times) - g_j(\times) = 0, \;\; i,j = 1,2,\ldots,M \;\; i \neq j\,. \tag{2.3}$$

Diskriminantinės funkcijos yra labai naudingos tada, kai sprendžiama su Gauso tikimybine tankio funkcija, kuri apibrėžiama šitaip:

$$p(\mathbf{x}|\Omega_i) = \frac{1}{(2\pi)^{n/2}|\Sigma_i|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i)\right), \quad i = 1,\ldots,M$$

kur μ_i yra vidurkio reikšmė ir Σ_i kovariacijų matrica klasėje Ω_i. Kovariacijų matrica yra apskaičiuojama pagal formulę

$$\Sigma_i = E\left[(\mathbf{x}-\mu_i)(\mathbf{x}-\mu_i)^T\right]$$

Toliau renkamės monotoninę logaritminę diskriminantinę funkciją $\ln(\cdot)$

$$g_i(\mathbf{x}) = \ln\left(p(\mathbf{x}|\Omega_i)P(\Omega_i)\right) = \ln p(\mathbf{x}|\Omega_i) + \ln P(\Omega_i)$$

tokiu būdu normaliojo tankio funkcijai yra gaunama

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x}-\mu_i)^T \Sigma_i^{-1}(\mathbf{x}-\mu_i) + \ln P(\Omega_i) + c_i$$

kur $c_i = -(n/2)\ln 2\pi - (1/2)\ln|\Sigma_i|$ yra konstanta (Bäse 2004).

Taigi, metodo esmė yra tokia: kiekvienai grupei sudaroma klasifikavimo funkcija, ir objektas priskiriamas tai grupei, kurios klasifikavimo funkcija, objektą atitinkančiam stebiniui, įgyja didžiausią reikšmę (Čekanavičius and Murauskas 2008).

2.3. Bajeso diskriminantinės funkcijos

Svarbiausias darbo uždavinys yra klasifikuoti Gauso atsitiktinio lauko stebinius $\{Z(s): s \in \mathcal{D} \subset \mathrm{R}^2\}$, kur stebinio $Z(s)$ modelis klasėje Ω_l yra apibrėžtas formule (1.1).

Darant prielaidas, kad klasės visiškai apibrėžtos, taip pat žinant populiacijos apriorines tikimybes π_1, π_2, ($\pi_1+\pi_2=1$), Bajeso diskriminantinė funkcija (BDF), minimizuojanti klaidingo klasifikavimo tikimybę, yra suformuota pagal sąlyginio tankio santykio logaritmą (Fukunaga, 1990).

Kai Z_0 yra *nepriklausomas* nuo mokymo imties, tada BDFI pagal McLachlan (2004) yra apibrėžiama:

$$W_k(Z_0) = \left(Z_0 - \frac{1}{2}(\mu_1 + \mu_2) \right)(\mu_1 - \mu_2) / \sigma^2 + \gamma(k) \qquad (2.4)$$

kur $\gamma(k) = \ln\left(\pi_1(k)/\pi_2(k)\right)$, Z_0 – klasifikuojamo stebinio požymis. μ_1 – pirmos klasės požymio reikšmių vidurkis, μ_2 – antros. $\pi_1(k)$, $\pi_2(k)$ – apriorinės klasių žymių tikimybės.

Tai atvejis *su žinomais parametrais*.

Tačiau praktiškai klasifikuojant objektus labai dažnai populiacijos parametrai yra nežinomi. Tokiu atveju yra naudojami jų įvertiniai, o pačios funkcijos yra vadinamos įterptosiomis Bajeso diskriminantinėmis funkcijomis (PBDFI). Taigi, įterpta Bajeso diskriminantinė funkcija, kai Z_0 *nepriklauso* nuo mokymo imties $T=t$ yra

$$W_k\left(Z_0; \hat{\mu}; \hat{\sigma}^2\right) = \left(Z_0 - \frac{1}{2}\left(\hat{\mu}_1 + \hat{\mu}_2\right) \right)\left(\hat{\mu}_1 - \hat{\mu}_2\right)/\hat{\sigma}^2 + \gamma(k), \qquad (2.5)$$

$$\hat{\mu} = (X_y' R^{-1} X_y)^{-1} X_y' R^{-1} Z = \begin{pmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \end{pmatrix}$$

$$\hat{\sigma}^2 = (Z - X_y \hat{\mu}) R^{-1} (Z - X_y \hat{\mu}) / (n-2)$$

kur $\gamma(k) = \ln\left(\pi_1(k)/\pi_2(k)\right)$, $\pi_1(k)$, $\pi_2(k)$ – apriorinės tikimybės (k nurodo skirtumą tarp skirtingų klasių stebinių skaičiaus), R – koreliacijų matrica tarp mokymo imties stebinių. X_y – plano matrica, turinti n eilučių ir stulpelių tiek, kiek klasių. Dydis n – mokymo lokacijų, susietų tam tikra kaimynystės schema su Z_0, skaičius.

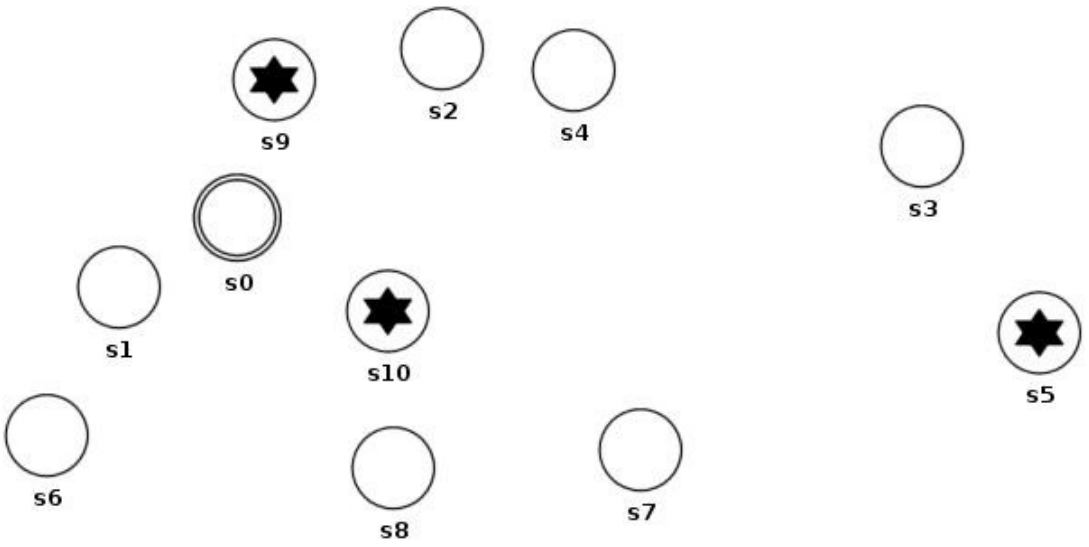
Tai atvejis *su dalinai žinomais parametrais*.

Koreliacijų matricos pavidalas:

$$R = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1n} \\ r_{21} & r_{22} & \cdots & r_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ r_{n1} & r_{n2} & \cdots & r_{nn} \end{pmatrix},$$

kur r_{ij} – koreliacinė funkcija tarp taškų s_i ir s_j, $i, j=1,...,n$. Šio darbo realizacijos dalyje naudojama eksponentinė KF, apibrėžta formule (1.2).

Dydis X_y – plano matrica, kurios elementai imami iš aibės {0,1}. Šios matricos eilučių skaičius atitinka mokymo imties elementų skaičių, o stulpelių skaičius – klasių skaičių. Plano matricoje elementas įgyja 1-o reikšmę kiekvienoje eilutėje ties tuo stulpeliu, kurią klasę jis ir atitinka. Plano matricos sudarymas pateiktas sekančiu pavyzdžiu (pav. 2.2).



2.2 pav. Taškų išsidėstymas, kur s_0 – klasifikuojamas taškas, balti skrituliukai – pirmos klasės taškai, o skrituliukai su žvaigždute – antros klasės taškai.

Pagal pav. 2.2 pateiktą schemą, kur s_1, s_2, s_3, s_4, s_6, s_7, s_8, yra taškai priklausantys pirmai klasei, o s_5, s_9, s_{10} – antrai klasei, sudaroma sekanti plano matrica:

$$X_y = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix}$$

$Y = (Y(s_1),...,Y(s_n))'$ – žymių vektorius, $Z = (Z(s_1),...,Z(s_n))'$ – požymių vektorius. $T' = (Z', Y')$ – mokymo imtis.

Vektoriaus Z_n modelis duotam $Y_n = y_n$ yra

$$Z = X_y \mu + E_n, \tag{2.6}$$

kur X_y – plano matrica, $\mu' = (\mu_1, \mu_2)$ ir E yra n-vektorius atsitiktinių paklaidų, turinčių Gauso pasiskirstymą $N_n(0, \sigma^2 R)$.

2.4. Priklausomybės nuo mokymo imties įvedimas

Skyrelyje 2.3 aprašytos Bajeso diskriminantinės funkcijos, taikomos tada, kai klasifikuojamas taškas yra laikomas nepriklausomu nuo mokymo imties. Tačiau dažnai taikomais atvejais mokymo imtis ir klasifikuojamas taškas yra tame pačiame erdviniame koreliuotame lauke, todėl į aptartas diskriminantines funkcijas naudinga įvesti klasifikuojamo taško erdvinę priklausomybę su mokymo imtimi (Dučinskas 2009), (Stabingienė *et al*. 2010).

Tarp klasifikuojamo stebinio požymio Z_0 ir požymių vektoriaus Z_n komponentų, erdvinių koreliacijų vektorių pažymėsime dydžiu r_0. Kadangi Z_0 yra koreliuotas su mokymo imtimi, todėl naudojame sąlyginį Z_0 Gauso pasiskirstymą duotai mokymo imčiai $T = t\,(Z = z, Y = y)$ su vidurkiu

$$\mu_{lt}^0 = E\left(Z_0 \middle| T = t; Y(s_0) = l\right) = \mu_l + \alpha_0' \left(z_0 - X_y \mu\right),\ l = 1,2$$

ir dispersija

$$\sigma_{0t}^2 = V\left(Z_0 \middle| T = t; Y(s_0) = l\right) = \sigma^2 R_{0n},$$

kur $\alpha_0' = r_0' R^{-1}$, $R_{0n} = 1 - r_0' R^{-1} r_0$. Čia r_0 - koreliacijų vektorius, apibrėžtas aukščiau $r_0' = (r_{01} \quad r_{02} \quad \cdots \quad r_{0n})$, kur r_{0i} yra koreliacinės funkcijos reikšmės.

Prielaida 2.1 Klasių žymės Y ir $Y(s_0)$ seka iš Markovo atsitiktinio lauko (MRF) modelio, kuriame lokacijos iš srities D, yra susietos viena su kita per kaimynystės sistemą.

Prielaida 2.2 Sąlyginis pasiskirstymas $Y(s_0)$ žinomam $\kappa=k$ yra nusakytas tik pagal žymes kaimynystėje N_0, t.y.

$$\pi_1(k) = P(Y(s_0) = 1 \mid \kappa = k) = 1/(1+\exp(-\lambda(k))$$

$$\pi_2(k) = 1 - \pi_1(k), \quad k = 0,..,K.$$

$$\lambda(k) = k\rho / K,$$

kur ρ – klasterizacijos parametras, o $\pi_1(k)$, $\pi_2(k)$ – apriorinės tikimybės, apibrėžtos prielaidoje 2.2.

Laikant Z_0 koreliuotą su mokymo imtimi, formulėje 2.4 vidurkio ir dispersijos dydžius keičiame sąlyginiais vidurkiais ir dispersija, apibrėžtais žemiau.

BDF, klasifikuojanti Z_0 *priklausomą* nuo mokymo imties $T=t$ ($\kappa=k$), yra

$$W_{tk}(Z_0) = \left(Z_0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0\right)\right)\left(\mu_{1t}^0 - \mu_{2t}^0\right)\Big/\sigma_{0t}^2 + \gamma(k) \qquad (2.7)$$

kur $\gamma(k) = \ln\left(\pi_1(k)/\pi_2(k)\right)$;

$$\mu_{1t}^0 = E(Z_0 \mid T = t; Y(s_0) = 1) = \mu_1 + \alpha_0'\left(z_0 - X_y\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right)$$

$$\mu_{2t}^0 = E(Z_0 \mid T = t; Y(s_0) = 2) = \mu_2 + \alpha_0'\left(z_0 - X_y\begin{pmatrix}\mu_1\\ \mu_2\end{pmatrix}\right)$$

$$\sigma_{0t}^2 = V(Z_0 \mid T = t; Y(s_0) = l) = \sigma^2 R_{0n}, \quad l = 1,2.$$

PBDF, klasifikuojanti Z_0 *priklausomą* nuo mokymo imties $T=t$ ($\kappa=k$), (Dučinskas 2009) yra:

$$W_{tk}\left(Z_0; \hat{\mu}; \hat{\sigma}^2\right) = \left(Z_0 - \frac{1}{2}\left(\hat{\mu}_{1t}^0 + \hat{\mu}_{2t}^0\right)\right)\left(\hat{\mu}_{1t}^0 - \hat{\mu}_{2t}^0\right)\Big/\hat{\sigma}_{0t}^2 + \gamma(k) \qquad (2.8)$$

$$\hat{\mu}_{1t}^0 = E(Z_0 \mid T = t; Y(s_0) = 1) = \hat{\mu}_1 + \alpha_0'\left(z_n - X_y\begin{pmatrix}\hat{\mu}_1\\ \hat{\mu}_2\end{pmatrix}\right)$$

$$\hat{\mu}_{2t}^0 = E(Z_0 \mid T = t; Y(s_0) = 2) = \hat{\mu}_2 + \alpha_0'\left(z_n - X_y\begin{pmatrix}\hat{\mu}_1\\ \hat{\mu}_2\end{pmatrix}\right)$$

$$\hat{\sigma}_{0t}^2 = V(Z_0 \mid T = t; Y(s_0) = l) = \hat{\sigma}^2 R_{0n}.$$

Tai atvejis, kai neturėdami tikrų parametrų reikšmių formulėje 2.4, įterpiame įvertinius.

2.5. Empirinis klaidingo klasifikavimo vertinimas

Klasifikuojant stebinius labai svarbu įvertinti kaip tiksliai tai atliekama. Vaizdų analizėje dažnai taikomos tokios charakteristikos, kaip bendras klasifikavimo tikslumas bei klaidingo klasifikavimo tikimybių įverčiai. Naudojant Bajeso diskriminantines funkcijas naudinga įvertinti Bajeso klaidos tikimybes, nes prie skirtingų situacijų suklystama nevienodai (tam turi įtakos mokymo imties dydis bei statistinių parametrų reikšmės).

Klasifikavimo kokybę taip pat nusako ir klaidingo klasifikavimo tikimybių įverčiai, kurie parodo, kokia yra tikimybė suklysti klasifikavimo metu kiekvienai iš klasių (Čekanavičius and Murauskas 2008).

Dviejų klasių atveju klaidingo klasifikavimo tikimybių p'_{12} ir p'_{21} įverčiai apskaičiuojami pagal formulę:

$$\hat{p}'_{12} = \frac{n'_{12}}{n_2} \tag{2.9}$$

čia n_2 - antros klasės stebinių skaičius, n'_{12} - antros klasės klaidingai suklasifikuotų stebinių skaičius.

$$\hat{p}'_{21} = \frac{n'_{21}}{n_1} \tag{2.10}$$

čia n_1 – pirmos klasės stebinių skaičius, n'_{21} – pirmos klasės klaidingai suklasifikuotų stebinių skaičius.

p'_{12} įvertina tikimybę stebinį priskirti pirmai klasei nors iš tikrųjų jis priklauso antrai, o p'_{21} įvertina tikimybę stebinį priskirti antrai klasei kai jis priklauso pirmajai. Kuo klaidingo klasifikavimo tikimybė mažesnė abiems klasėms, tuo klasifikavimas atliktas tiksliau.

Dažnesnis klaidų tikimybių lygio įvertis (Härdle and Simar 2003) yra pateiktas sekančia išraiška

$$\frac{n'_{12} + n'_{21}}{n_2 + n_1} \tag{2.11}$$

Kita klasifikavimo kokybės įvertinimo charakteristika yra bendras klasifikavimo tikslumas (*angl. overall accuracy of classification*). Ši charakteristika parodo kokia dalis yra suklasifikuota teisingai. Bendro klasifikavimo tikslumo įvertinys:

$$\hat{P}_t = \frac{n_t}{N} \tag{2.12}$$

čia n_t – teisingai suklasifikuotų stebinių skaičius, o N – bendras stebinių skaičius.

Tokie empiriniai metodai veda į pernelyg optimistiškus vertinimus, tačiau tai yra gruboki kokybės matai diskriminantinei taisyklei.

2.6. Klasifikatoriaus veikimo vertinimas

Atpažinimo ciklas prasideda duomenų rinkimu, po to pradinių duomenų apdorojimo analize, o tada klasifikavimo taisyklės (klasifikatoriaus) parinkimu. Klasifikatoriaus veikimas yra svarbus aspektas objekto (šablono) atpažinimo cikle. Tad natūraliai kyla klausimai, tokie kaip: kaip gerai klasifikatorius veikia palyginus su konkuruojančiais? Kaip pagerinti klasifikatorių veikimą? Tad tokie klausimai atsakymus randa klasifikatorių veikimo vertinime (*angl. performance assessment*). Klasifikatoriaus veikimo vertinimas dažnai yra atskira dalis, nors turėtų būti klasifikatoriaus konstravimo (*angl. classifier design*) dalis, nes konstruojant klasifikatorių svarbu numatyti, kaip jis „elgsis“ prie skirtingų statistinių parametrų reikšmių, prie skirtingų klasių kaimynų skaičiaus ar prie skirtingų erdvinių imčių planų (*angl. spatial sampling design*).

Klasifikatoriaus konstravimo ir jo veikimo vertinimo dalys yra atskiros galbūt todėl, kad kriterijus, naudojamas klasifikatoriaus konstravimui dažnai skiriasi nuo kriterijų, naudojamų jiems vertinti. Pavyzdžiui, sudarant diskriminantinę taisyklę, galime rinktis parametrus pagal taisyklę, optimizuojančią ma-

tavimo kvadratines paklaidas, o taisyklės veikimą vertiname naudojant skirtingas atlikimo priemones, tokias kaip klaidų tikimybes (*angl. error rate*).

Klasifikavimo taisyklės veikimą nusako tokie aspektai: 1) taisyklės *atskiriamumas* (*angl. discriminability*) (kaip ji gerai klasifikuoja nematytus duomenis), 2) taisyklės *patikimumas* (*angl. reliability*) – tai yra matas (priemonė), kaip gerai ji įvertina (nustato) klasės priklausymo (*angl. class membership*) aposteriorines tikimybes.

Nemažai yra priemonių klasifikavimo taisyklės atskiriamumui vertinti, tačiau dažnesnės – klaidingo klasifikavimo tikimybės (*angl. misclassification rate*) arba tiesiog klaidų tikimybės. Klaidų tikimybės dažnai vertinamos pagal turimus duomenis, nes bendrai, nelengvas uždavinys yra gauti klaidų tikimybių analitinę išraišką (Webb 2002).

Šiame disertaciniame darbe yra pateiktos išvestos klaidų tikimybių analitinės išraiškos Bajeso diskriminantinėms funkcijoms. Jos reikalingos tam, kad įvertinti šių funkcijų veikimą. Be to, svarbu paminėti, kad daugelis autorių tyrė BDF veikimą, tačiau netyrė šių tikimybių priklausomybių nuo statistinių parametrų reikšmių, kas ir yra pateikta šio darbo skaitiniuose pavyzdžiuose.

Prieš pateikiant keletą populeresnių klaidų tikimybių įvertinimų tipų, pažymėsime reikalingus dydžius. Mokymo duomenys (*angl. training data*) $T' = \left(Z', Y'\right)$. $Z = (Z(s_1),...,Z(s_n))'$ – požymių vektorius. $Y = (Y(s_1),...,Y(s_n))'$ – žymių vektorius, $Y(s_i)_l = 1$, *jei* $Z\left(s_i\right) \in \Omega_l$ ir $Y(s_i)_l = 0$, *jei* $Z\left(s_i\right) \notin \Omega_l$. Pažymėsime $\Omega\left(Y(s_i)\right)$ besąlyginę klasės žymę. Sprendimo taisyklę sudarytą naudojant mokymo duomenis pažymėsime $\eta\left(\mathbf{z}_0; T\right)$ (η yra klasė į kurią yra priskirtas $\mathbf{z}_0$ požymis pagal sudarytą klasifikatorių naudojant mokymo duomenis *T*=*t*). Nuostolių funkciją pažymėsime $Q\left(\Omega\left(Y(s)\right), \eta\left(\mathbf{z}_0; T\right)\right)$, t.y.

$$Q\left(\Omega\left(Y(s)\right), \eta\left(\mathbf{z}_0; T\right)\right) = \begin{cases} 0, & jei \;\; \Omega\left(Y(s)\right) = \eta\left(\mathbf{z}_0; T\right) \;\; (klasif.\, teisng.) \\ 1 & (priešingu\ atveju) \end{cases}$$

Tariamoji klaidos tikimybė (*angl. apparent error rate*) e_A, kur klaidų tikimybei įvertinti naudojama numatyta aibė (*angl. design set*)

$$e_A = \frac{1}{n}\sum_{i=1}^{n} Q\left(\Omega\left(Y(s_i)\right), \eta\left(Z\left(s_i\right);T\right)\right).$$

Įvertis gali būti griežtai optimistiškai šališkas ypač sudėtingiems klasifikatoriams ir mažai duomenų aibei. Didinant mokymo imčių skaičių sumažinamas šališkumas.

Faktinė klaidos tikimybė (*angl. true error rate*) arba tikroji (*angl. or actual error rate or conditional error rate*) e_T, – tai klasifikatoriaus, klaidingai klasifikuojančio atsitiktinai parinktą objektą, tikėtina tikimybė. Tai klaidos tikimybė su be galo didele bandymo aibe sudaryta iš to paties skirstinio kaip mokymo duomenų.

Vidutinė klaidos tikimybė (*angl. expected error rate*) e_E, t.y. faktinės klaidos tikimybės vidutinė reikšmė per tam tikro dydžio mokymo aibes, $e_E=E[e_T]$.

Bajeso klaidos tikimybė arba optimali klaidos tikimybė e_B – tai faktinės klaidos tikimybės teorinis minimumas, reikšmė faktinės klaidos tikimybės jei klasifikatorius sudarytas iš teisingų grupės priklausymo aposteriorinių tikimybių (Webb 2002).

2.7. Tiksli klaidos tikimybė Bajeso diskriminantinei funkcijai

Šiame skyrelyje yra išvesta tikslios klaidos tikimybės formulė izotropinės (priklausančios nuo atstumo, bet ne nuo krypties) eksponentinės erdvinės koreliacijos atveju. Tai atlikta panaikinant nepriklausomumo prielaidą, t.y., paprastai erdviniame klasifikavime dažnai laikoma, jog požymių stebiniai priskiriant klasių žymes yra nepriklausomai pasiskirstę (dažnai praktiškai pasitaiko, jog erdvėje taškai yra šalia vienas kito, todėl tikėtina, kad jie koreliuoja). Taigi, ši nepriklausomumo prielaida yra panaikinta, remiantis stacionariu Gauso lauko modeliu požymių stebiniams ir dar, laikome, jog klasių žymių vektorius seka iš Markovo atsitiktinio lauko modelio (vaizdų analizėje įprasta laikyti, jog klasių žymes tenkina MRF modelis). Išvestos tikslios klaidos tikimybės priklausomybė nuo statistinių parametrų reikšmių yra tiriama statistiškai Markovo atsitiktiniams laukams su pirmos eilės kaimynų sistema.

Pagrindinis šio darbo tikslas yra klasifikuoti Gauso atsitiktinio lauko stebinius $\{Z(s): s \in D \subset R^2\}$.

Stebinio Z(s) modelis klasėje Ω_l yra

$$Z(s) = x'(s)\beta_l + \varepsilon(s), \tag{2.13}$$

kur $x(s)$ yra $q \times 1$ dydžio neatsitiktinių regresorių vektorius ir β_l yra $q \times 1$ dydžio parametrų vektorius, *l*=1,2. Klaidos narys yra nulinio vidurkio stacionarus Gauso atsitiktinis laukas $\{\varepsilon(s): s \in D\}$ su kovariacine funkcija, apibrėžta sekančiu modeliu visiems $s, u \in D$

$$\operatorname{cov}\{\varepsilon(s), \varepsilon(u)\} = \sigma^2 r(s-u),$$

kur *r*(*s*-*u*) yra erdvinės koreliacijos funkcija ir σ^2 yra dispersija kaip mastelio parametras.

Tegul $L = \{1,2\}$ yra žymių aibė (*angl. label set*). Klasės, susietos su *Z*(*s*) žymė yra žymima $Y(s)$, $Y(s) \in L, s \in D$. Tegul $S_n = \{s_i \in D;\ i = 1,...,n\}$ yra mokymo sričių aibė. Tada, aibė $T = \{(Z(s_i), Y(s_i)) \in R \times L;\ i = 1,...n\}$ sudaro mokymo aibę (*angl. training set*). Problema – klasifikavimas požymio stebinio $Z_0 = Z(s_0)$ į vieną iš dviejų klasių (įvertinimas $Y(s_0)$) su pateikta mokymo imtimi *T*.

Marginalųjį Mahalanobio atstumą (*angl. marginal Mahalanobis distance*) žymime šitaip: $\Delta_0 = \left| x_0^{'}(\beta_1 - \beta_2) \right| / \sigma$. Šis dydis parodo atstumą tarp dviejų tikimybinių skirstinių. Ši sąvoka pirmą kartą paminėta autoriaus Mahalanobio darbe (Mahalanobis 1936).

Neprarandant bendrumo, tarkime, kad n_1 lokacijų iš S_n turi žymę lygią 1 ir likusios lokacijos $n_2 = n\text{-}n_1$ turi žymę lygią 2 (pirma klasė – 1, o antra klasė - 2). Tegul *n* yra fiksuotas ir n_2, n_1 yra atsitiktiniai kintamieji. Aibė žymių ir požymio reikšmių apibrėžiamos atitinkamai, $Y_n = (Y(s_1),...,Y(s_n))'$ ir $Z_n = (Z(s_1),...,Z(s_n))'$.

Taigi, vektoriaus Z_n modelis konkrečiai aibei Y_n=y_n yra

$$Z_n = X\beta + E_n \tag{2.14}$$

kur X yra $n\times 2q$ dydžio plano matrica, $\beta' = (\beta_1', \beta_2')$ ir E yra n-vektorius atsitiktinių paklaidų, turinčių daugiamatį Gauso skirstinį $N_n(0, \sigma^2 R)$. Plano matrica X formulėje (2.14) yra apibrėžiama

$$X = X_1 \oplus X_2,$$

kur simbolis $\oplus$ reiškia tiesioginę matricų sumą ir X_l, l=1,2 yra $n_l\times q$ dimensijos regresorių matrica mokymo imčiai (TS) T=t.

Pažymime r_0 vektoriumi erdvinių koreliacijų tarp Z_0 ir Z_n; R – erdvinių koreliacijų matricą tarp komponentų Z_n. Kadangi Z_0 yra koreliuotas su mokymo imtimi, naudojame sąlyginį Z_0 Gauso skirstinį duotai T=t (Z_n=z_n) su vidurkiais μ_{lt}^0 ir dispersija σ_{0t}^2, apibrėžtais sekančiomis formulėmis

$$\mu_{lt}^0 = E\left(Z_0 \middle| T = t; Y(s_0) = l\right) = x_0'\beta_l + \alpha\left(z_n - X\beta\right),\ l = 1,2,$$

ir

$$\sigma_{0t}^2 = V\left(Z_0 \middle| T = t; Y(s_0) = l\right) = \sigma^2 R_{0n},$$

kur $x_0 = x(s_0)$, $\alpha = r_0' R^{-1}$, $R_{0n} = 1 - r_0' R^{-1} r_0$.

Pagal prielaidą, kad klasės yra visiškai nusakytos ir yra žinomos populiacijų apriorinės tikimybės π_1 ir π_2 $(\pi_1 + \pi_2 = 1)$, Bajeso diskriminantinė funkcija (BDF), minimizuojanti klaidingo klasifikavimo tikimybę (ME) yra suformuota pagal sąlyginių tankių santykio logaritmą (Fukunaga 1990).

Prielaida 2.3. Žymių vektorius Y ir $Y(s_0)$ laikoma, kad seka iš MRF su tam tikra kaimynų sistema. Tegul N_0 kaimynystė s_0 sudaryta iš $2K$ kaimynų.

Be to, tariame, kad kaimynų skaičius su žyme i yra atsitiktinis ir žymimas m_i, i=1,2. Apibrėžiame naują atsitiktinį kintamąjį, nusakantį skirtumus tarp klasių: $\kappa = |m_1 - m_2|/2$. Dydis π_k pateikia tikimybes reikšmėms κ, t.y., $\kappa = k$, k=0,1,...,K.

Taip pat galioja aukščiau pateikta 2.2 prielaida, t.y.: sąlyginis skirstinys $Y(s_0)$ duotam $\kappa = k$ yra apibrėžiamas pagal žymes kaimynystėje N_0 tokiu būdu:

$$\pi_1(k) = P(Y(s_0) = 1 | \kappa = k) = 1/(1+\exp(-\lambda(k)),$$

$$\pi_2(k) = 1 - \pi_1(k),\ k = 0,\ldots,K,$$

kur $\lambda(k) = k\rho / K$, o ρ – neneigiama konstanta vadinama klasterizavimo parametru (Nishii and Eguchi 2006). Neneigiamas parametras ρ pateikia klasių žymių erdvinės priklausomybės laipsnį.

Tada BDF duotam $T=t$ ir $\kappa = k$ yra

$$W_{tk}(Z_0) = \left(Z_0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0\right)\right)' \left(\mu_{1t}^0 - \mu_{2t}^0\right) \Big/ \sigma_{0t}^2 + \gamma(k) \qquad (2.15)$$

kur $\gamma(k) = \ln\left(\pi_1(k)/\pi_2(k)\right) = \lambda(k)$.

2.1 Apibrėžimas Bajeso diskriminantinei funkcijai (BDF) $W_{tk}(Z_0)$ sąlyginė klaidos tikimybė $P_k(t)$ yra apibrėžiama

$$P_k(t) = \sum_{l=1}^{2} \pi_l(k) P_{lk}(t), \qquad (2.16)$$

kur

$$P_{lk}(t) = P\left((-1)^l W_{tk}(Z_0)\right) > 0 \left| Y(s_0) = l\right),\ l = 1,2. \qquad (2.17)$$

$P_{lk}(t)$ yra tikimybė stebinio reikšmę Z_0 priskirti ne tai klasei (kai diskriminantinės funkcijos $W_{tk}(Z_0)$ reikšmė yra didesnė už 0, priskiriame stebinį Z_0 pirmai klasei, kai $W_{tk}(Z_0) < 0$ – antrai).

Sąlyginis Mahalanobio atstumas pateikiamas tokiu būdu

$$\Delta_{0n} = \left|\left(\mu_{1t}^0 - \mu_{2t}^0\right)\right| \Big/ \sigma_{0t} = \Delta_0 / \sqrt{R_{0n}},$$

kur $\Delta_0 = \left|x_0^{'}(\beta_1 - \beta_2)\right| \Big/ \sigma$, $R_{0n} = 1 - r_0' R^{-1} r_0$.

Akivaizdu, kad Δ_{0n} priklauso nuo T tik per S_n.

Lema 2.1. Tarkime, kad požymių vektorius Z_n seka iš modelio (2.14) ir klasės žymių vektorius – iš MRF, detalizuoto prielaidose 2.3 ir 2.2, tokiu atveju sąlyginė klaidos tikimybė funkcijai $W_{tk}\left(Z_0\right)$ (BDF) yra $P_k(t)$

$$P_{lk}(t)=\sum_{l=1}^{2}\pi_l(k)\Phi\left(-\Delta_{0n}/2+\left(-1\right)^l\lambda(k)/\Delta_{0n}\right). \tag{2.18}$$

Įrodymas. Naudojant normaliojo skirstinio savybes iš 2.17 turime

$$P_l(t)=\Phi\left(-\Delta_{0n}/2+\left(-1\right)^l\lambda(k)/\Delta_{0n}\right). \tag{2.19}$$

Tada įterpdami šią formulę į (2.16), užbaigiame lemos 2.1 įrodymą.

Iš 2.17 gaunamas 2.19.

►$T=t$

$$W_{tk}\left(t\right)|_{T=t}\sim N\left(\begin{array}{l}\left(\mu_{lt}^0-\frac{1}{2}\left(\mu_{1t}^0+\mu_{2t}^0\right)\right)'\left(\mu_{1t}^0-\mu_{2t}^0\right)\Big/\sigma_{0t}^2+\gamma\left(k\right);\\ \sigma_{0t}^2\left(\left(\mu_{1t}^0-\mu_{2t}^0\right)/\sigma_{0t}^2\right)^2\end{array}\right)$$

$$Z_0\,|_{\Omega_l}\sim N\left(\mu_{lt}^0;\,\sigma_{0t}^2\right)$$

$$E\left(Z_0\right)=\mu_{lt}^0$$

$$D\left(Z_0\right)=\sigma_{0t}^2$$

$$E\left(W_{tk}\right)=\left(\mu_{lt}^0-\frac{1}{2}\left(\mu_{1t}^0+\mu_{2t}^0\right)\right)'\left(\mu_{1t}^0-\mu_{2t}^0\right)\Big/\sigma_{0t}^2+\gamma\left(k\right)$$

$$D\left(W_{tk}\right)=\sigma_{0t}^2\left(\left(\mu_{1t}^0-\mu_{2t}^0\right)/\sigma_{0t}^2\right)^2=\frac{\left(\mu_{1t}^0-\mu_{2t}^0\right)^2}{\sigma_{0t}^2}$$

Mahalanobio atstumas: $\Delta_{0t}=\left(\mu_{1t}^0-\mu_{2t}^0\right)/\sigma_{0t}$

Normaliojo skirstinio savybės

$$P\left(X<x\right)=\Phi\left(x\right)$$

$$P\left(X>x\right)=1-\Phi\left(x\right)$$

$$\Phi(x) \sim N(0,1)$$

I klasė: l=2, nes skaičiuojame klaidą

$$P(W_{tk} > 0) = P\left(\frac{W_{tk} - E(W_{tk})}{\sqrt{D(W_{tk})}} > \frac{0 - E(W_{tk})}{\sqrt{D(W_{tk})}} \right) = 1 - \Phi\left(-\frac{E(W_{tk})}{\sqrt{D(W_{tk})}} \right) =$$

$$= \Phi\left(\frac{E(W_{tk})}{\sqrt{D(W_{tk})}} \right) = \Phi\left(\frac{\left(\mu_{2t}^0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0 \right) \right)\left(\mu_{1t}^0 - \mu_{2t}^0 \right) \Big/ \sigma_{0t}^2 + \gamma(k)}{\sqrt{\left(\mu_{1t}^0 - \mu_{2t}^0 \right)^2 \Big/ \sigma_{0t}^2}} \right) =$$

$$= \Phi\left(\frac{\left(\mu_{2t}^0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0 \right) \right)\left(\mu_{1t}^0 - \mu_{2t}^0 \right)\sigma_{0t}}{\sigma_{0t}^2\left(\mu_{1t}^0 - \mu_{2t}^0 \right)} + \frac{\gamma(k)\sigma_{0t}}{\left(\mu_{1t}^0 - \mu_{2t}^0 \right)} \right) =$$

$$= \Phi\left(\frac{\left(\mu_{2t}^0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0 \right) \right)}{\sigma_{0t}} + \frac{\lambda(k)}{\Delta_{0t}} \right) = \Phi\left(\frac{\left(\frac{1}{2}\left(-\mu_{1t}^0 - \mu_{2t}^0 + 2\mu_{2t}^0 \right) \right)}{\sigma_{0t}} + \frac{\lambda(k)}{\Delta_{0t}} \right) =$$

$$= \Phi\left(\frac{\left(\frac{1}{2}\left(\mu_{2t}^0 - \mu_{1t}^0 \right) \right)}{\sigma_{0t}} + \frac{\lambda(k)}{\Delta_{0t}} \right) = \Phi\left(\frac{\left(-\frac{1}{2}\left(-\mu_{2t}^0 + \mu_{1t}^0 \right) \right)}{\sigma_{0t}} + \frac{\lambda(k)}{\Delta_{0t}} \right) =$$

$$= \Phi\left(\frac{-\Delta_{0t}}{2} + \frac{\lambda(k)}{\Delta_{0t}} \right)$$

II klasė: *l*=1, nes skaičiuojame klaidą

$$P(W_{tk} < 0) = P\left(\frac{W_{tk} - E(W_{tk})}{\sqrt{D(W_{tk})}} < \frac{0 - E(W_{tk})}{\sqrt{D(W_{tk})}} \right) = \Phi\left(-\frac{E(W_{tk})}{\sqrt{D(W_{tk})}} \right) =$$

$$= \Phi\left(-\frac{\left(\mu_{1t}^0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0 \right) \right)\left(\mu_{1t}^0 - \mu_{2t}^0 \right) \Big/ \sigma_{0t}^2 + \lambda(k)}{\sqrt{\left(\mu_{1t}^0 - \mu_{2t}^0 \right)^2 \Big/ \sigma_{0t}^2}} \right) =$$

$$= \Phi\left(- \frac{\left(\mu_{1t}^0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0 \right) \right)\left(\mu_{1t}^0 - \mu_{2t}^0 \right)\sigma_{0t}}{\sigma_{0t}^2 \left(\mu_{1t}^0 - \mu_{2t}^0 \right)} - \frac{\lambda(k)\sigma_{0t}}{\left(\mu_{1t}^0 - \mu_{2t}^0 \right)} \right) =$$

$$= \Phi\left(- \frac{\left(-\frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0 - 2\mu_{1t}^0 \right) \right)}{\sigma_{0t}} - \frac{\lambda(k)}{\Delta_{0t}} \right) =$$

$$= \Phi\left(- \frac{\left(\frac{1}{2}\left(\mu_{1t}^0 - \mu_{2t}^0 \right) \right)}{\sigma_{0t}} - \frac{\lambda(k)}{\Delta_{0t}} \right) =$$

$$= \Phi\left(- \frac{\Delta_{0t}}{2} - \frac{\lambda(k)}{\Delta_{0t}} \right)$$

◄

2.2 Apibrėžimas Tiksli klaidos tikimybė aptartai klasifikavimo procedūrai $W_{tk}(Z_0)$ yra apibrėžta kaip $P_0(S_n) = E_T\left(P_k(T)\right)$, kur E_T reiškia vidurkį atsižvelgiant į T skirstinį.

Lema 2.2 Pagal lemos 2.1 prielaidas duotam $\{\pi_k\}$ tiksli klaidos tikimybė EER (*angl. exact error rate*) funkcijai $W_{tk}(Z_0)$ yra

$$\begin{aligned} P_0(S_n) &= \pi_0 \Phi\left(-\Delta_{0n}/2\right) + \\ &+ 2\sum_{k=1}^{K}\sum_{l=1}^{2} \pi_k \pi_l(k) \Phi\left(-\Delta_{0n}/2 + (-1)^l \lambda(k)/\Delta_{0n}\right) \end{aligned} \tag{2.20}$$

Įrodymas. Formulė seka vidurkinant išraišką $P_k(t)$, pateiktą (2.18), skirtingoms atsitiktinėms k reikšmėms su tikimybėmis π_k, k=0,1,…,K ir pastebint, kad $\lambda(0)=0$. π_k nusako kiek tikėtina, kad skirtumas tarp skirtingų klasių taškų yra k.

Pavyzdyje skaitiškai analizuojama tikslios klaidos tikimybės $P_0(S_n)$ priklausomybė nuo tam tikrų statistinių parametrų reikšmių. Tarkime D yra dvimatė gardelė ir $S_8 = \{(0,1),\ldots,(-1,-1)\}$, $S_0 = (0,0)$. Pažymime, kad S_8 yra ant-

ros eilės kaimynų aibė su S_0. Laikysime, kad MRF yra susietas su pirmos eilės kaimynystės sistema t.y., *K*=2 ir $N_0 = \{(0,1),(1,0),(0,-1),(-1,0)\}$.

Aptariame modelio (2.14) atvejį su pastoviais vidurkiais ir izotropine eksponentine erdvinės koreliacijos funkcija $r(h) = \exp\{-|h|/\alpha\}$.

Tikslios klaidos tikimybės (formulė(2.20)) reikšmės, skirtingoms klasterizavimo parametro ρ reikšmėms ir koreliacijos pločio parametrui α, esant fiksuotam Δ_0 ir $\{\pi_k\}$, pateiktos lentelėje 2.1.

Lentelė 2.1. Reikšmės $P_0(S_n)$ su $\Delta_0 = 0,2$ ir $\pi_0 = 0,6$, $\pi_1 = 0,3$, $\pi_2 = 0,1$

ρ \ α	0.5	1	1.5	2	2.5	3
0	0.45877	0.45110	0.44271	0.43503	0.42805	0.42166
0.2	0.45562	0.44843	0.44043	0.43302	0.42624	0.42000
0.4	0.44698	0.44095	0.43394	0.42725	0.42100	0.41518
0.6	0.43462	0.42989	0.42410	0.41834	0.41284	0.40761
0.8	0.42014	0.41655	0.41191	0.40712	0.40240	0.39783
1	0.40457	0.40189	0.39824	0.39432	0.39036	0.38643
1.2	0.38851	0.38656	0.38372	0.38055	0.37725	0.37392
1.4	0.37235	0.37096	0.36879	0.36625	0.36353	0.36072
1.6	0.35633	0.35537	0.35374	0.35173	0.34950	0.34715
1.8	0.34060	0.33997	0.33878	0.33721	0.33540	0.33345
2	0.32529	0.32490	0.32405	0.32285	0.32140	0.31979

Lentelėje 2.1 yra matyti, kad tiksli klaidos tikimybė silpnai atskirtoms klasėms (Δ_0=0.2) yra monotoniškai mažėjanti keičiantis parametrui α su fiksuota ρ reikšme. Taip pat galima teigti, jog stebiniai su stipresne erdvine priklausomybe gali būti tiksliau klasifikuojami.

Stipriau atskirtoms klasėms (Δ_0=1) vyrauja panašios tendencijos (lentelėje 2.2)

Taigi, skaitinės analizės rezultatai pateikia stiprius argumentus teigti, kad didesnis klasių žymių klasterizavimas ir stipresnė erdvinė koreliacija tarp stebinių požymių užtikrina mažesnes erdvinės klasifikacijos paklaidas.

Lentelė 2.2. Reikšmės $P_0(S_n)$ su $\Delta_0 = 1$ ir $\pi_0 = 0,6$, $\pi_1 = 0,3$, $\pi_2 = 0,1$

ρ \ α	0.5	1	1.5	2	2.5	3
0	0.30237	0.26946	0.23562	0.20671	0.18229	0.16151
0.2	0.30180	0.26901	0.23526	0.20642	0.18205	0.16130
0.4	0.30011	0.26766	0.23419	0.20554	0.18131	0.16068
0.6	0.29735	0.26545	0.23243	0.20410	0.18011	0.15965
0.8	0.29358	0.26243	0.23002	0.20212	0.17845	0.15824
1	0.28891	0.25866	0.22700	0.19964	0.17638	0.15648
1.2	0.28343	0.25422	0.22343	0.19671	0.17391	0.15438
1.4	0.27726	0.24919	0.21938	0.19337	0.17111	0.15199
1.6	0.27052	0.24367	0.21491	0.18967	0.16800	0.14933
1.8	0.26334	0.23775	0.21009	0.18568	0.16462	0.14645
2	0.25581	0.23150	0.20499	0.18143	0.16104	0.14338

2.8. Tikėtinos klaidos tikimybės aproksimacija

Šiame skyrelyje pateikta tikslios klaidos tikimybės formulė, kuri išvesta Bajeso diskriminantinei funkcijai, o dalinai žinomų parametrų atveju (vidurkiai ir dispersija nežinomi) yra pateikta aproksimacija tikėtinos klaidos tikimybės, susijusios su įterpta BDF. Minėtų klaidos tikimybių priklausomybė nuo statistinių parametrų reikšmių ir aptartų modelių yra ištirta skaitiškai mokymo lokacijų aibei, sudarant antros eilės kaimynystę klasifikuojamo stebinio lokacijoms.

Pagrindinis darbo tikslas klasifikuoti požymių stebinius, modeliuojamus stacionariu Gauso atsitiktiniu lauku $\left\{Z(s): s \in D \subset \mathrm{R}^2\right\}$.

Stebinio $Z(s)$ marginalinis modelis klasėje Ω_l yra

$$Z(s) = \mu_l + \varepsilon(s),$$

kur μ_l pastovus vidurkis, o klaidos narys nulinio vidurkio stacionarus Gauso atsitiktinis laukas $\left\{\varepsilon(s): s \in D\right\}$ su kovariacine funkcija.

Tegul $L = \{1,2\}$ žymių aibė. Lokacijos žymė $s \in D$ susieta su $Z(s)$ yra atsitiktinis kintamasis. $Y(s)$ įgyja reikšmes iš L. Tegul $S_n = \left\{s_i \in D; i = 1,...,n\right\}$

mokymo lokacijos. Aibė $Y=(Y(s_1),...,Y(s_n))'$ – žymių vektorius ir $Z=(Z(s_1),...,Z(s_n))'$ – požymių vektorius.

Vektorius $T'=\left(Z',Y'\right)$ vadinamas mokymo imtimi.

Tarkime, kad atvejis $\{T=t\}$ yra ekvivalentus atvejui $\{Z=z\}\cap\{Y=y\}$, kur *t, z, y* yra realizacijos atitinkamų atsitiktinių vektorių.

R – erdvinių koreliacijų matrica tarp komponentų *Z*. Tarkime, kad S_n yra fiksuotas, bet žymės yra pasiskirsčiusios atsitiktinai.

Kai *Y*=*y*, S_n yra padalintas į du poaibius, t.y. $S_n=S_y^{(1)}\cup S_y^{(2)}$, kur $S^{(l)}$ yra poaibis S_n, kur n_l yra lokacijos su žymėmis lygiomis *l*, *l*=1,2 (n_1+n_2 = *n*).

Tada modelis vektoriaus Z duotam $Y=y$ yra

$$Z=X_y\mu+E_n \tag{2.21}$$

kur X_y yra $n\times 2$ plano matrica, $\mu'=\left(\mu_1,\mu_2\right)$ ir E_n yra *n*-vektorius atsitiktinių paklaidų, turinčių daugiamatį Gauso skirstinį $N_n\left(0,\sigma^2R\right)$.

Čia problema yra stebinio $Z_0=Z(s_0)$, $s_0\in D$, $s_0\notin S_n$ klasifikavimas (įvertinimas $Y(s_0)$) su pateikta mokymo imtimi *T*.

Dydis r_0 – vektorius erdvinių koreliacijų tarp Z_0 ir Z. Naudojamas sąlyginis Z_0 Gauso skirstinys duotam $T=t$ su vidurkiais

$$\mu_{lt}^0=E\left(Z_0\middle|T=t;Y(s_0)=l\right)=\mu_l+\alpha_0'\left(Z-X_y\mu\right),\ l=1,2$$

ir dispersija

$$\sigma_{0t}^2=V\left(Z_0\middle|T=t;Y(s_0)=l\right)=\sigma^2R_{0n},$$

kur $\alpha_0'=r_0'R^{-1}$, $R_{0n}=1-r_0'R^{-1}r_0$.

Pažymime žymių vektoriaus *Y* pasiskirstymą $\left\{\pi\left(y\right)=P\left(Y=y\right)\right\}$.

Prielaida 2.4. *Sąlyginis skirstinys* $Y(s_0)$ *duotam T=t priklauso tik nuo* $Y=y$, (žymių realizacijos) *t.y.*

$$\pi_l(y)=P(Y(s_0)=l\mid T=t),l=1,2.$$

Pagal prielaidą, kad klasės yra visiškai tiksliai nusakytos, BDF (Fukunaga, 1990), minimizuojanti klaidingo klasifikavimo tikimybę, yra suformuota pagal sąlyginių tankių santykio logaritmą, aprašytą aukščiau.

Taigi, BDF, skirta klasifikuoti stebinį Z_0, kai duota mokymo imtis $T = t$, yra

$$W_t(Z_0) = \left(Z_0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0\right) \right)' \left(\mu_{1t}^0 - \mu_{2t}^0\right) / \sigma_{0t}^2 + \gamma(y) \qquad (2.22)$$

kur $\gamma(y) = \ln\left(\pi_1(y)/\pi_2(y)\right)$.

Ši situacija vadinama pilnai žinomų parametrų atveju.

Sąlyginis Mahalanobio atstumas duotam $T = t$ yra

$$\Delta_{0n} = \left| \left(\mu_{1t}^0 - \mu_{2t}^0\right) \right| / \sigma_{0t} = \Delta_0 / \sqrt{R_{0n}},$$

kur $\Delta_0 = \left|\mu_1 - \mu_2\right| / \sigma$ yra marginalinis Mahalanobio atstumas. Akivaizdu, kad Δ_{0n} priklauso nuo S_n, bet nepriklauso nuo *t*.

Sąlyginė Bajeso klaidos tikimybė (duotam $T = t$) klasifikuojamo Z_0 pagal BDF $W_t(Z_0)$ yra

$$P_0(t) = \sum_{l=1}^{2} \pi_l(y) \Phi\left(-\Delta_{0n}/2 + (-1)^l \gamma(y)/\Delta_{0n}\right), \qquad (2.23)$$

kur $\Phi(\cdot)$ standartinė normaliojo skirstinio funkcija.

Tiksli Bajeso klaidos tikimybė funkcijai $W_t(Z_0)$ yra

$$E_T(P_0(T)) = \sum_{ky1l=1}^{2} \pi(y)\pi_l(y)\Phi\left(-\Delta_{0n}/2 + (-1)^l \gamma(y)/\Delta_{0n}\right), \qquad (2.24)$$

kur E_T reiškia vidurkį pagal T skirstinį ir *k* skirtumus tarp klasių.

Tarkime, kad $\{\mu_l\}$ ir σ^2 yra nežinomi parametrai ir turi būti įvertinti iš mokymo imties *T*.

Tegul $\hat{\mu}$ ir $\hat{\sigma}^2$ įvertiniai μ ir σ^2, paremti mokymo imtimi $T = t$ Pažymėkime trijų parametrų vektorių dydžiu $\Psi' = \left(\mu, \sigma^2\right)$ ir vektorių trijų įvertinių $\hat{\Psi}' = \left(\hat{\mu}, \hat{\sigma}^2\right)$.

Įterpta BDF (PBDF) yra gauta pakeičiant parametrus funkcijoje BDF jų įvertiniais, paremtais mokymo imtimi $T = t$. Aukščiau apibrėžtai klasifikavimo problemai PBDF yra

$$W_t\left(Z_0; \hat{\Psi}\right) = \left(Z_0 - \frac{1}{2}\left(\hat{\mu}_{1t}^0 + \hat{\mu}_{2t}^0\right)\right)\left(\hat{\mu}_{1t}^0 - \hat{\mu}_{2t}^0\right) / \hat{\sigma}_{0t}^2 + \gamma(y), \tag{2.25}$$

kur $\hat{\mu}_{lt}^0 = E\left(Z_0 \middle| T = t; Y(s_0) = l\right) = \mu_l + \alpha_0'\left(z_n - X_y\hat{\mu}\right)$, $l = 1, 2$

ir

$$\hat{\sigma}_{0t}^2 = V\left(Z_0 \middle| T = t; Y(s_0) = l\right) = \hat{\sigma}^2 R_{0n}.$$

Aptartu atveju, tikroji (faktinė) klaidos tikimybė (Dučinskas, 2009) funkcijai $W_t\left(Z_0; \hat{\Psi}\right)$ yra apibrėžta

$$P_t\left(\hat{\Psi}\right) = \sum_{l=1}^{2} \pi_l(y)\Phi\left(\hat{Q}_l(t)\right), \tag{2.26}$$

ir

$$\begin{aligned} \hat{Q}_l(t) = (-1)^l \Bigg(& \mu_{lt}^0 - \frac{1}{2}\left(\hat{\mu}_{1t}^0 + \hat{\mu}_{2t}^0\right)\Bigg)\operatorname{sgn}\left(\hat{\mu}_{1t}^0 - \hat{\mu}_{2t}^0\right) \\ & + \gamma(y)\hat{\sigma}_{0t}^2 / \left|\hat{\mu}_{1t}^0 - \hat{\mu}_{2t}^0\right|\Big) / \sigma_{0t} \end{aligned} \tag{2.27}$$

2.3 Apibrėžimas Tikrosios klaidos tikimybės vidurkis $\left(E_T\left\{P_T\left(\hat{\Psi}\right)\right\}\right)$ atsižvelgiant į bendrą skirstinį T, $E_T\left\{P\left(\hat{\Psi}\right)\right\}$, yra vad. tikėtina klaidos tikimybe.

Tikėtina klaidos tikimybė (E_PER) naudinga nurodant funkcijos PBDF veikimo orientyrą, nes tai yra faktiškai suformuota iš mokymo imties. Vadinasi, E_PER Z_0 klasifikavimo problemai pagal PBDF yra

$$E_{T,}\left(P_T\left(\hat{\Psi}\right)\right) = E_T\left\{\sum_{l=1}^{2} \pi_l(Y)\Phi\left(\hat{Q}_l(T)\right)\right\}. \tag{2.28}$$

Lema 2.3 Tarkime, kad klasifikuojamas stebinys Z_0 pagal PBDF apibrėžtą formule (2.21) ir asimptotinė aproksimacija E_PER apibrėžta formule (2.28) bei remiantis teoremos (Dučinskas 2009) prielaidomis, turime:

$$AEP_0 = \sum_y \sum_{l=1}^{2} \pi(y)\pi_l(y)\Phi\left(-\Delta_{0n}/2 + (-1)^l \gamma(y)/\Delta_{0n}\right) + \sum_y \sum_l \pi_l(y)\pi(y)\phi(Q_1(y))(C(y) + 2\gamma^2(y)/(n-2))/\Delta_{0n} \quad (2.29)$$

kur $\varphi(\cdot)$ yra standartinio normaliojo tankio funkcija ir

$$C(y) = \Lambda' R_\mu \Lambda \Delta_{0n}^2 / \rho_0,\ \Lambda = X_y^{'}\alpha_0 - H/2 + \gamma(y)G/\Delta^2{}_{0n}.$$

Aibė $H = (1,1)', G = (1,-1)'$.

Įrodymas. Lemos įrodymas yra paremtas išraiškos $P_T(\hat{\Psi})$ (2.26), (2.27) Teiloro eilutės išplėtimu apie $\mu = \hat{\mu}$ ir $\hat{\sigma}^2 = \sigma^2$ taškus.

Tokiu atveju atsižvelgdami į Teiloro pagrindinio nario vidurkį (expectation) baigiame lemos įrodymą. Detalesnė informacija teoremos įrodymo yra darbe (Dučinskas 2009).

Žemiau esančiose lentelėse skaitiškai yra analizuojama tikslios klaidos tikimybės priklausomybė nuo tam tikrų statistinių modelio parametrų reikšmių. Tarkime *D* yra dvimatė taisyklinga gardelė su vienetiniu proporcingu didinimu, $S_0 = (0,0)$ ir S_8 – antros eilės kaimynų aibė su S_0.

Aptariamas modelio (2.21) atvejis su pastoviais vidurkiais ir izotropine eksponentine erdvinės koreliacijos funkcija $r(h) = \exp\{-|h|/\alpha\}$, kur α pločio parametras.

Aibė $Y_i = Y(s_i)$, $y_i = y(s_i)$, *i*=1,…,*n*. Tarkime, kad sąlyginis skirstinys $Y(s_0)$ duotam $Y = y$ yra

$$\pi_1(y) = P(Y(s_0) = 1 | Y = y) = 1/(1 + \exp(\rho(1 - 2n_1/n)),$$

ir apriorinis klasių žymių skirstinys yra $\pi(y) = \pi_{n_1}/C_8^{n_1}$, kur

$$n_1 = \#\left(i : y_i = 1,\, i = 1,...,n\right),\ \text{ir}\ \pi_{n_1} = P\left(\sum_{i=1}^{n} 1\{Y_i = 1\} = n_1\right),\ n_1 = 0,...,n\,.$$

Lentelė 2.3. Reikšmės AEP_0 su Δ_0=0.2, π_4=0.5, π_3= π_5=0.15, π_2= π_6=0.1

ρ \ α	0.5	1	1.5	2	2.5	3
0.2	0.45828	0.45023	0.44223	0.43495	0.42832	0.42224
0.4	0.45462	0.44710	0.43959	0.43274	0.42651	0.42078
0.6	0.44917	0.44231	0.43548	0.42926	0.42364	0.41844
0.8	0.44256	0.43633	0.43023	0.42479	0.41992	0.41542
1	0.43530	0.42959	0.42420	0.41960	0.41561	0.41194
1.2	0.42771	0.42242	0.41769	0.41396	0.41094	0.40820
1.4	0.42000	0.41505	0.41093	0.40811	0.40614	0.40442
1.6	0.41229	0.40761	0.40408	0.40222	0.40138	0.40078
1.8	0.40466	0.40021	0.39727	0.39641	0.39680	0.39743
2	0.39718	0.39291	0.39056	.390786	0.39251	0.39447

Lentelė 2.4. Reikšmės AEP_0 su Δ_0=3, π_4=0.5, π_3= π_5=0.15, π_2= π_6=0.1

ρ \ α	0.5	1	1.5	2	2.5	3
0.2	0.07570	0.03865	0.02005	0.01027	0.00521	0.00263
0.4	0.07687	0.03933	0.02046	0.01051	0.00534	0.00269
0.6	0.07831	0.04018	0.02097	0.01079	0.00549	0.00277
0.8	0.08006	0.04124	0.02158	0.01113	0.00567	0.00287
1	0.08214	0.04252	0.02232	0.01153	0.00589	0.00298
1.2	0.08459	0.04403	0.02319	0.01200	0.00613	0.00311
1.4	0.08742	0.04579	0.02419	0.01254	0.00641	0.00325
1.6	0.09066	0.04782	0.02533	0.01315	0.00673	0.00341
1.8	0.09434	0.05012	0.02662	0.01384	0.00709	0.00360
2	0.09846	0.05271	0.02807	0.01460	0.00748	0.00380

Skaitinės analizės rezultatai leidžia teigti, kad didelis klasterizavimas (susitelkimas) klasių žymių ir stipresnė erdvinė koreliacija tarp požymių stebinių garantuoja mažesnes erdvinės klasifikacijos paklaidas.

2.9. Vidutinė Bajeso klaidos tikimybė

Skyrelyje aptariama statistinio klasifikavimo problema naudojant daugiamatį stacionarų Gauso atsitiktinį lauką modeliavimui sąlyginio tankio duotoms požymių stebinių klasių žymėms. Klasės yra apibrėžtos daugiamačiu regresijos vidurkių modeliu ir bendra faktorizuota kovariacijų funkcija. Yra išvesta vidutinė Bajeso klaidos tikimybė (EBER) dviejų klasių atvejui, kai klasių žymės modeliuojamos atsitiktiniu lauku (RF), paremtu 0-1 divergencija. Laikoma, kad klasifikuojamas stebinys yra priklausomas nuo mokymo imties. Mokymo imties dydžio efektas ir statistinių parametrų reikšmių įtaka klaidos tikimybei (EBER) yra skaitiškai ištirta tokiu atveju, kai duomenų erdvinė struktūra yra dvimatės taisyklingos gardelės poaibis su vienetiniais atstumais.

Toliau trumpai yra apibrėžiami požymių bei žymių modelių aprašymai.

Tarkime, kad erdviniai duomenys susideda iš stebėtų požymio kintamojo reikšmių, kurios yra modeliuojamos *p*-mačiu atsitiktiniu lauku

$$\{Z(s): s \in D \subset \mathrm{R}^2\}.$$

Laikoma, kad kiekviena lokacija srityje D priklauso vienai iš klasių Ω_1, Ω_2. Klasės žymė ar tiesiog žymė lokacijai $s \in D$ yra žymima $Y(s)$, ir traktuojama kaip atsitiktinis kintamasis žymių aibėje $L = \{1,2\}$.

Požymio stebinio $Z(s)$ klasėje Ω_l (t.y. $Y(s) = l$) modelis yra

$$Z(s) = B_l' x(s) + \varepsilon(s), \tag{2.30}$$

kur $x(s)$ yra $q \times 1$ dydžio neatsitiktinių regresorių vektorius ir B_l yra $q \times p$ dydžio parametrų matrica, *l*=1,2. Reikalaujama, kad $B_1 \neq B_2$. Klaidos narys išraiškoje (2.30) yra *p*-matis nulinio vidurkio stacionarus GRF $\{\varepsilon(s): s \in D\}$ su kovariacine funkcija apibrėžta sekančiu modeliu visiems $s, u \in D$

$$\mathrm{cov}\{\varepsilon(s), \varepsilon(u)\} = r(s-u)\Sigma, \tag{2.31}$$

kur $r(s-u)$ erdvinės koreliacijos funkcija ir Σ yra požymio kovariacijų matrica.

Aibė $Y = (Y(s_1),...,Y(s_n))'$ yra mokymo žymių vektorius, o $Z = (Z(s_1),...,Z(s_n))'$ mokymo požymių matrica. Todėl mokymo imtį sudaro matrica $T = (Z, Y)$ dydžio $n \times (p+1)$.

Pažymime $S_n = \{s_i \in D;\ i = 1,...,n\}$ lokacijų aibę, kurioje mokymo imtis yra imama T ir ją pavadiname mokymo lokacijų aibe. Taip pat vadinama mokymo imties erdvine struktūra (Sheckar *et al.* 2002).

Tarkime, kad atsitiktinių mokymo kintamųjų $Y = y$ ir $Z = z$ realizacijos atitinka mokymo imties $T = t$ realizaciją. Vadinasi mokymo požymių matricos Z skirstinys duotam $Y = y$ yra matric-matis (*angl. matrix-variate*) normalus skirstinys, t.y.

$$Z \,|\, Y = y \sim N_{n \times p}\left(X_y B,\ R \otimes \Sigma\right), \tag{2.32}$$

kur X_y yra $n \times 2q$ plano matrica, $B' = (B_1', B_2')$ yra $p \times 2q$ dydžio vidurkio parametrų matrica.

Čia R reiškia erdvinių koreliacijų matricą požymių stebiniams iš S_n su elementais, kurie yra lygūs erdvinės koreliacijos r atitinkamoms reikšmėms formulėje (2.31).

Dydžio $n \times 2q$ plano matrica X_y yra suformuota tokiu būdu: į pirmuosius q stulpelius įeina požymių stebinių regresoriai iš Ω_1, ir į sekančius q stulpelius įeina požymių stebinių regresoriai iš Ω_2.

Toliau darbe aptariama požymio stebinio $Z_0 = Z(s_0),\ s_0 \in D$ klasifikavimo problema su nežinoma klasės žyme duotos mokymo imties atveju. Žymė lokacijai s_0 yra žymima Y_0.

Pažymime tarp Z_0 ir Z erdvinių koreliacijų vektorių dydžiu r_0 ir aibę Z^+

$$Z^+ = \begin{pmatrix} Z \\ Z_0 \end{pmatrix},\ R^+ = \begin{pmatrix} R & r_0 \\ r_0' & 1 \end{pmatrix},\ x_0 = x(s_0),\ \alpha_0 = R^{-1} r_0 .$$

Tai išplaukia iš (2.30) – (2.32), kai $l = 1, 2$

$$Z^+ \,|\, Y = y,\ \ Y_0 = l \sim N_{(n+1)\times p}(X_y^l B, R^+ \otimes \Sigma), \tag{2.33}$$

kur

$$X_y^l = \begin{pmatrix} X_y \\ x_0^l \end{pmatrix}, \quad x_0^l = (\delta_{1l} \mathrm{I}_q \otimes \delta_{2l} \mathrm{I}_q) x_0,$$

kur δ_{ij} Kronekerio delta ir I_q vienetinė matrica q eilės. Sąlyginis skirstinys Z_0 duotam $T = t$ yra Gauso, t.y.

$$Z_0 \big| T = t, \;\; Y_0 = l \sim N_p(\mu_{lt}^0, \Sigma_{ot}). \tag{2.34}$$

Sąlyginiai vidurkiai μ_{lt}^0 yra

$$\mu_{lt}^0 = E(Z_0 \big| T = t; Y_0 = l) = B_l' x_0 + (z - X_y B)' \alpha_0, \;\; l = 1, 2. \tag{2.35}$$

Sąlyginė kovariacijų matrica Σ_{0t} apibrėžta sekančiai

$$\Sigma_{0t} = Var\left(Z_0 \,|\, T = t; Y_0 = l\right) = R_{on} \Sigma, \tag{2.36}$$

su $R_{0n} = 1 - r_0' \alpha_0$.

Lokacijos $s = s_0$ požymio stebinio Marginalinis Mahalanobio atstumo kvadratas tarp populiacijų yra

$$\Delta_0^2 = (\mu_1^0 - \mu_2^0)' \Sigma^{-1} (\mu_1^0 - \mu_2^0) \tag{2.37}$$

kur $\mu_l^0 = B_l' x_0, \quad l = 1, 2.$

Mahalanobio atstumo kvadratas tarp Z_0 sąlyginių skirstinių duotam $T = t$ yra apibrėžiamas

$$\Delta_{0n}^2 = \left(\mu_{1t}^0 - \mu_{2t}^0\right)' \Sigma_{0t}^{-1} \left(\mu_{1t}^0 - \mu_{2t}^0\right). \tag{2.38}$$

Panaudojus (2.35), (2.36) išraiškas formulėse (2.37), (2.38) gauname

$$\Delta_{0n}^2 = \Delta_0^2 / R_{0n}.$$

Akivaizdu, kad Δ_{0n} priklauso nuo mokymo imties tik per S_n.

Laikome, kad mokymo lokacijų aibė (STL) (*angl. set of training locations*) S_n yra fiksuota, bet žymės jose yra pasiskirsčiusios atsitiktinai.

Taigi S_n yra padalinama į tarpusavyje nesusikertančių poaibių sąjungą, t.y $S_n = S^{(1)} \cup S^{(2)}$, kur $S^{(l)}$ yra atsitiktinis S_n poaibis, kuriame yra N_l lokacijų skaičius su žymėmis lygiomis $l, l = 1,2.$ Kadangi $N_1 + N_2 = n$, yra pakankama naudoti tiktai N_1 skirstinį.

Pažymime diskretaus atsitiktinio kintamojo N_1 skirstinį

$$\left\{\pi_j = P(N_1 = j), \quad j = 0,1,...,n\right\} \tag{2.39}$$

Šios tikimybės kartais yra vadinamos apriorinėmis žymių tikimybėmis.

Vadiname $\xi(y) = \left\{S^{(1)}, S^{(2)}\right\}$ erdviniu žymių planu (SLD) atitinkančiu mokymo žymių vektoriaus realizaciją $Y = y.$ Akivaizdu, kad $\xi(Y)$ atitinka Y. Tarkim, jei $Y = y$, tada $N_i = n_i, i = 1,2$, kur $n_1 + n_2 = n$.

Tegul $J(l,m)$ neneigiama divergencija tarp dviejų klasių Ω_l ir Ω_m, visiems $m,l = 1, 2$, tenkinantiems $J(l,l) = 0$.

Apibrėžkime Y_0 sąlyginį skirstinį duotam $Y = y$

$$\pi_l(y) = P(Y_0 = l \mid Y = y), \; l = 1,2.$$

Čia išplečiamas metodas, naudotas autorių Nishii ir Eguchi (Nishii and Eguchi 2006), žymių skirstinio, paremto $S_n \cup s_0$ modeliavimui. Laikoma, jog žymės Y_0 skirstinys sąlyginai pagal $Y = y$ yra apibrėžiamas žymėmis iš S_n ir divergencija.

Aptariame atvejį kai žymės Y_0 sąlyginis tankis tiesiogiai nepriklauso nuo S_n ir s_0.

Divergencijų vidurkis tarp lokacijos s_0 (su žyme l) ir žymių lokacijose S_n, apibrėžiamas tokiu būdu

$$\Delta(l) = \sum_{m=1}^{2} n_m J(l,m) / n \tag{2.40}$$

Prielaida 2.5 Y_0 sąlyginis skirstinys duotam $Y = y$ (su $N_l = n_l$) yra apibrėžiamas sekančia formule

$$\pi_l(y) = \exp\{-\rho\Delta(l)\} / \sum_{k=1}^{2} \exp\{-\rho\Delta(k)\}, \; l = 1,2, \qquad (2.41)$$

kur ρ neneigiama konstanta, vadinama klasterizacijos parametru. Ji parodo atsitiktinio lauko erdvinės priklausomybės laipsnį. Jei $\rho = 0,$ tada klasės yra erdvėje nepriklausomos.

Remiantis prielaida, jog klasės yra visiškai apibrėžtos, BDF (Fukunaga, 1990) minimizuojanti klaidingo klasifikavimo tikimybę, yra suformuota iš sąlyginių tankių (apibrėžtų anksčiau) santykio logaritmo. Ši funkcija Z_0 klasifikavimui pagal $T = t$ yra

$$W_t(Z_0) = \left(Z_0 - \frac{1}{2}\left(\mu_{1t}^0 + \mu_{2t}^0\right) \right)' \Sigma_{0t}^{-1} \left(\mu_{1t}^0 - \mu_{2t}^0\right) + \gamma(y), \qquad (2.42)$$

kur $\gamma(y) = \ln\left(\pi_1(y)/\pi_2(y)\right)$.

Apibrėžimas 2.4 Sąlyginė Bajeso klaidos tikimybė (CBER) yra apibrėžta kaip sąlyginė tikimybė nuo $T = t$, jog atsitiktinis stebinys Z_0 yra klaidingai suklasifikuojamas pagal BDF $W_t(Z_0)$. Ją pažymime $P_0(t)$.

Lema 2.4 Tarkim, kad Z_0 sąlyginis skirstinys, apibrėžtas formulėse (2.34)-(2.36) ir Y_0 sąlyginis skirstinys tenkina prielaidą 2.5. Tada sąlyginė Bajeso klaidos tikimybė, klasifikuojant Z_0 pagal funkciją $W_t(Z_0)$ (BDF), yra

$$P_0(t) = \sum_{l=1}^{2} \pi_l(y) \Phi\left(Q_l(t)\right) \qquad (2.43)$$

kur

$$Q_l(t) = -\Delta_{0n}/2 + (-1)^l \gamma(y)/\Delta_{0n} .$$

Čia $\Phi(\cdot)$ standartinio normaliojo skirstinio funkcija.

Lemos 2.4 įrodymas tiesiogiai seka iš apibrėžimo 2.4 ir iš daugiamačio Gauso skirstinio savybių.

Apibrėžimas 2.5 *Vidutinė Bajeso klaidos tikimybė* (*EBER*) *funkcijai* $W_t(Z_0)$ *yra apibrėžiama kaip* $P_{0n} = E_T(P_0(T))$, *kur* E_T *reiškia vidurkį atsižvelgiant į T skirstinį.*

Dėl lengvesnės interpretacijos, neprarandant bendrumo, naudojamas tam tikras kvazi atstumas.

Prielaida 2.6 *Kvazi atstumas J*(*l,m*) *yra 0-1 atstumas, apibrėžtas* $J(l,m) = 1 - \delta_{lm}$, *kur* δ_{lm} *yra Kronekerio delta.*

Pažymime, jog erdvinis modelis su siūlomu divergencijos tipu yra dažnai naudojamas vaizdų segmentavime (Besag 1986).

Remiantis prielaida 2.6, iš formulės (2.41), turime

$$\begin{aligned} \pi_1(y) &= 1/(1+\exp\{-\rho(n_1-n_2)/n\}), \\ \pi_2(y) &= 1/(1+\exp\{\rho(n_1-n_2)/n\}). \end{aligned} \tag{2.44}$$

Aukščiau minėtos tikimybės fiksuotam n priklauso nuo y tik per n_1. Taigi galima pristatyti naują žymėjimą

$$\pi_l^*(n_1) = \pi_l(y), \quad \gamma^*(n_1) = \gamma(y). \tag{2.45}$$

Kadangi atsitiktinis kintamasis N_1 yra Y funkcija, tai iš (2.39) ir (2.45) gauname, jog jungtinis skirstinys (*angl. joint distribution*) $\{P(Y=y, Y_0=l)\} = \{P(Y_0=l|Y=y)P(Y=y)\}$ turi sekančią formą

$$\left\{P\left(N_1=n_1, Y_0=l\right) = \pi_l^*(n_1)\cdot\pi_{n_1},\ n_1=0,...,n;\ l=1,2\right\}. \tag{2.46}$$

Lema 2.5 Prie lemos 2.4 sąlygų, prielaidos 2.6 ir duotoms apriorinėms tikimybėms $\{\pi_j\}$, funkcijos $W_t(Z_0)$ tikimybė EBER yra

$$P_{on} = \sum_{j=0}^{n}\sum_{l=1}^{2} \frac{\pi_j \Phi\left(-\Delta_{0n}/2 + (-1)^l \rho(2j/n-1)/\Delta_{0n}\right)}{\left(1+\exp\{(-1)^l\rho(2j/n-1)\}\right)} \tag{2.47}$$

Įrodymas. Naudojant prielaidą 2.6, formulę (2.46) ir (2.45) į išraišką (2.43) mes gauname sekančią CBER formulę

$$P_0(t) = \sum_{l=1}^{2} \pi_l^*(n_1)\Phi\left(-\Delta_{0n}/2 + (-1)^l \gamma^*(n_1)/\Delta_{0n}\right) \qquad (2.48)$$

Iš formulės (2.46) matome, jog CBER priklauso nuo $T = t$ tik per $N_1 = n_1$. Taigi galime pakeisti $P_0(T)$ vidurkinimą, atsižvelgiant į T skirstinį jį suvidurkinant pagal N_1 skirstinį, apibrėžtą (2.39) formule. Lemos 2.5 įrodymas yra gaunamas panaudojant formules nuo (2.44) iki (2.46) išraiškoje (2.48).

►Tikimybes $\pi_1(y)$ ir $\pi_2(y)$ gausime iš prielaidos 2.6 ir formulės 2.41:

Divergencija:

$$J(l,m) = 1 - \delta_{lm}$$

Divergencijų vidurkis:

$$\Delta(l) = \sum_{m=1}^{2} n_m J(l,m)/n$$

Tikimybės:

$$\pi_l(y) = \frac{\exp\{-\rho\Delta(l)\}}{\sum_{k=1}^{2}\exp\{-\rho\Delta(k)\}}$$

Kai *l*=1

$$\pi_1(y) = \frac{\exp\{-\rho\Delta(1)\}}{\sum_{k=1}^{2}\exp\{-\rho\Delta(k)\}} = \frac{\exp\left\{-\rho\left(\sum_{m=1}^{2} n_m J(1,m)/n\right)\right\}}{\sum_{k=1}^{2}\exp\left\{-\rho\left(\sum_{m=1}^{2} n_m J(k,m)/n\right)\right\}} =$$

$$= \frac{\exp\left\{-\rho\left(n_1 J(1,1)/n + n_2 J(1,2)/n\right)\right\}}{\exp\left\{-\rho\left(\sum_{m=1}^{2} n_m J(1,m)/n\right)\right\} + \exp\left\{-\rho\left(\sum_{m=1}^{2} n_m J(2,m)/n\right)\right\}} =$$

$$= \frac{\exp\{-\rho(n_2/n)\}}{\exp\left\{-\rho\left(\frac{n_1 J(1,1)}{n} + \frac{n_2 J(1,2)}{n}\right)\right\} + \exp\left\{-\rho\left(\frac{n_1 J(2,1)}{n} + \frac{n_2 J(2,2)}{n}\right)\right\}} =$$

$$= \frac{\exp\{-\rho(n_2/n)\}}{\exp\left\{-\rho\left(\frac{n_2}{n}\right)\right\}+\exp\left\{-\rho\left(\frac{n_1}{n}\right)\right\}} = \frac{1}{1+\exp\left\{-\rho\left(\frac{n_1}{n}\right)+\rho\left(\frac{n_2}{n}\right)\right\}} =$$

$$= \frac{1}{1+\exp\left\{-\rho\left(\frac{(n_1-n_2)}{n}\right)\right\}} = 1/\left(1+\exp\left\{\frac{-\rho(n_1-n_2)}{n}\right\}\right)$$

Kai *l*=2

$$\pi_2(y) = \frac{\exp\{-\rho\Delta(2)\}}{\sum_{k=1}^{2}\exp\{-\rho\Delta(k)\}} = \frac{\exp\left\{-\rho\left(\sum_{m=1}^{2} n_m J(2,m)/n\right)\right\}}{\sum_{k=1}^{2}\exp\left\{-\rho\left(\sum_{m=1}^{2} n_m J(k,m)/n\right)\right\}} =$$

$$= \frac{\exp\{-\rho(n_1 J(2,1)/n + n_2 J(2,2)/n)\}}{\exp\left\{-\rho\left(\sum_{m=1}^{2}\frac{n_m J(1,m)}{n}\right)\right\}+\exp\left\{-\rho\left(\sum_{m=1}^{2}\frac{n_m J(2,m)}{n}\right)\right\}} =$$

$$= \frac{\exp\{-\rho(n_1/n)\}}{\exp\left\{-\rho\left(\frac{n_1 J(1,1)}{n}+\frac{n_2 J(1,2)}{n}\right)\right\}+\exp\left\{-\rho\left(\frac{n_1 J(2,1)}{n}+\frac{n_2 J(2,2)}{n}\right)\right\}} =$$

$$= \frac{\exp\{-\rho(n_1/n)\}}{\exp\left\{-\rho\left(\frac{n_2}{n}\right)\right\}+\exp\left\{-\rho\left(\frac{n_1}{n}\right)\right\}} = \frac{1}{1+\exp\left\{-\rho\left(\frac{n_2}{n}\right)+\rho\left(\frac{n_1}{n}\right)\right\}} =$$

$$= \frac{1}{1+\exp\left\{\rho\left(\frac{n_1-n_2}{n}\right)\right\}} = 1\Big/\left(1+\exp\left\{\frac{\rho(n_1-n_2)}{n}\right\}\right)$$

Toliau naudojame gautas tikimybių išraiškas EBER išvedimui:

$$\pi_1(y) = 1/\left(1+\exp\{-\rho(n_1-n_2)/n\}\right)$$

$$\pi_2(y) = 1/\left(1+\exp\{\rho(n_1-n_2)/n\}\right).$$

$$P_0(t)=\sum_{l=1}^{2}\pi_l(y)\Phi\left(-\Delta_{0n}/2+(-1)^l\ln\left(\frac{\pi_1(y)}{\pi_2(y)}\right)/\Delta_{0n}\right)=$$

$$=\sum_{l=1}^{2}\frac{\Phi\left(-\Delta_{0n}/2+(-1)^l\ln\left(\frac{1+\exp\{\rho(n_1-n_2)/n\}}{1+\exp\{-\rho(n_1-n_2)/n\}}\right)/\Delta_{0n}\right)}{\left(1+\exp\{(-1)^l\rho(n_1-n_2)/n\}\right)}$$

$$=\left[\begin{array}{l}TVARKOM\quad \ln,\ t.y.\ \ln\left(\frac{1+\exp\{\rho(n_1-n_2)/n\}}{1+\exp\{-\rho(n_1-n_2)/n\}}\right)=\\ =\left[\rho(n_1-n_2)/n\triangleq x\right]=\\ =\ln\left(\frac{1+e^x}{1+\frac{1}{e^x}}\right)=\ln\left(\frac{e^x+e^{2x}}{e^x+1}\right)=\ln\left(\frac{e^x(1+e^x)}{(e^x+1)}\right)=x=\\ =-\rho(n_1-n_2)/n\end{array}\right]=$$

$$=\sum_{l=1}^{2}\frac{\Phi\left(-\Delta_{0n}/2+(-1)^l\rho\left((n_1-n_2)/n\right)/\Delta_{0n}\right)}{\left(1+\exp\{(-1)^l\rho(n_1-n_2)/n\}\right)}=$$

$$=\sum_{j=0}^{n}\sum_{l=1}^{2}\pi_j\frac{\Phi\left((-\Delta_{0n}/2)+\left((-1)^l\rho\left(\frac{2j}{n}-1\right)\right)/\Delta_{0n}\right)}{\left(1+\exp\left\{(-1)^l\rho\left(\frac{2j}{n}-1\right)\right\}\right)}$$

EBER gaunasi tokia:

$$EBER=\sum_{j=0}^{n}\sum_{l=1}^{2}\frac{\pi_j\Phi\left(\left(\frac{-\Delta_{0n}}{2}\right)+\left((-1)^l\rho\left(\frac{2j}{n}-1\right)\right)/\Delta_{0n}\right)}{\left(1+\exp\left\{(-1)^l\rho\left(\frac{2j}{n}-1\right)\right\}\right)}$$

Pastaba:

$\left(\frac{2j}{n}-1\right)$ gavome iš $(n_1-n_2)/n$, kai n_1 įgijinėja reikšmes nuo 0 iki n. ◄

Išvesta uždaros formos išraiška efektyviai gali būti naudojama kaip BDF veikimo matmuo ir kaip erdvinės imties plano optimalumo kriterijus.

Išvestos formulės EBER priklausomybė nagrinėjama nuo kai kurių statistinių parametrų reikšmių. Vertinamas mokymo imties dydžio efektas naudojant EBER. Skaitinei iliustracijai aptariamas (2.30) – (2.33) modelio vienmatis atvejis su pastoviais vidurkiais ir izotropine eksponentine kovariacine funkcija *C*(h)

$$C(h) = \sigma^2 \exp\{-|h|/\alpha\},$$

kur σ^2 dispersija, o α pločio parametras.

Tarkim, kad *D* yra 2-matė taisyklinga gardelė su vienetiniais tarpais ir $S_0 = (0,0)$. Be žymaus bendrumo praradimo imamas atvejis, kai $n = 2K$, kur K yra fiksuotas natūrinis skaičius. Tada, EBER P_{on} yra tokia

$$P_{on} = \sum_{j=0}^{2K}\sum_{l=1}^{2} \pi_j \Phi\left(-\Delta_{0n}/2 + (-1)^l \rho(j/K-1)/\Delta_{0n}\right) / \left(1+\exp\{(-1)^l \rho(j/K-1)\}\right) \quad (2.49)$$

Tiriama, kaip mokymo imties dydis įtakoja EBER reikšmes. Imties dydžio daroma įtaka klaidai EBER yra tiriama naudojant pirmos eilės $(K=2)$, antros eilės $(K=4)$ ir trečios eilės kaimynystes $(K=6)$ taškui s_0. Geresniam interpretavimui yra paimamas $\pi_0 = 1$. Šiuo atveju tikimybė lygi 1, kad iš aibės S_n lokacijų skaičius su žyme 1 yra lygus lokacijų skaičiui su žyme 2. Tada EBER yra išreiškiama formule

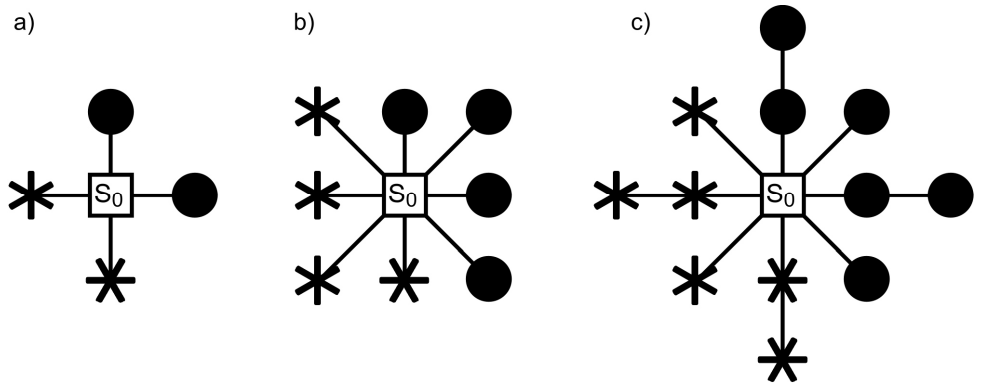
$$P_{0n} = \Phi(-\Delta_{0n}/2).$$

Šiuo atveju EBER nepriklauso nuo klasterizacijos parametro. ρ.

SLD atvejai tenkinantys šio pavyzdžio sąlygas yra pavaizduoti pav. 2.3

BE(*i*) yra EBER reikšmė, kuri apskaičiuojama tokiam STL, kuris suformuoja *i*-tosios eilės kaimynystę taškui s_0; čia *i*=1,2,3. Dviejų skirtingų dydžių STL palyginimas yra atliekamas panaudojant efektyvumo indeksą, apibrėžiamą kaip santykį $E(ij) = BE(i)/BE(j)$.

Tikimybių (EBER), apibrėžtų 2.49 formule, reikšmės ir efektyvumo indeksų reikšmės yra apskaičiuotos skirtingoms pločio parametro α reikšmėms, bet fiksuotam Δ_0=1. Gauti rezultatai pateikti lentelėje 1.



Pav. 2.3 Skirtingi erdvinių žymių planai su taškais $S^{(1)}$ ir $S^{(2)}$ pažymėtais • ir ∗. Atvejai a), b) ir c) atitinkamai atvaizduoja pirmos eilės, antros eilės ir trečios eilės kaimynystės taškui s_0 schemas.

Lentelė 2.5 Tikimybės EBER reikšmės ir efektyvumo indeksų reikšmės trims kaimynystėms ir skirtingoms pločio parametro α reikšmėms.

α	0,5	1	1,5	2	2,5	3
BE(1)	0,30260	0,27007	0,23612	0,20707	0,18255	0,16170
BE(2)	0,30237	0,26946	0,23562	0,20671	0,18229	0,16151
BE(3)	0,30236	0,26930	0,23512	0,20588	0,18120	0,16021
E(21)	0,99924	0,99774	0,99788	0,99826	0,99858	0,99882
E(13)	0,99921	0,99715	0,99576	0,99425	0,99260	0,99079
E(23)	0,99997	0,99941	0,99788	0,99599	0,99402	0,99195

Lentelėje 2.5 pateikti rezultatai patvirtina gana logiškas išvadas, kad EBER mažėja didėjant imties dydžiui. Analizuojant lentelės 2.5 eilutes su efektyvumo indeksų reikšmėmis, galime teigti, jog EBER mažėjimo greitis yra didesnis prie didesnių pločio parametro α reikšmių.

Taip pat yra tiriama statistinių parametrų įtaka tikimybei EBER. Čia nagrinėjamos STL formuojančios antros eilės kaimynystę taškui s_0 (t.y. $n = 8$).

Pažymėsime, jog antros eilės kaimynystę taškui $s_0 = (0,0)$ atitinka aibė

$$S_8 = \{(0,1),(1,1),(1,0),(1,-1),(0,-1),(-1,-1),(-1,0),(-1,1)\}.$$

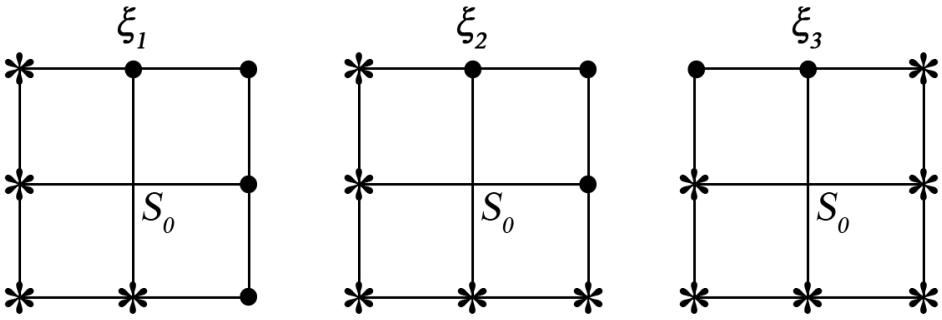
Tarkime, jog apriorinės tikimybės yra $\pi_2 = \pi_6 = 0{,}05$, $\pi_3 = \pi_5 = 0{,}15$, $\pi_4 = 0{,}6$ ir $\pi_j = 0$ visiems $j = 0{,}1{,}7{,}8$.

Tada tik tokie SLD, kurie atitinka $N_1 = 2{,}3{,}...{,}6$ yra reikšmingi.

Sekantys trys SLD

$$\xi_1 = \left\{ S^{(1)} = \{(0,1),(1,1),(1,0),(1,-1)\}, S^{(2)} = \{(0,-1),(-1,-1),(-1,0),(-1,1)\} \right\},$$
$$\xi_2 = \left\{ S^{(1)} = \{(0,1),(1,1),(1,0)\}, S^{(2)} = \{(1,-1),(0,-1),(-1,-1),(-1,0),(-1,1\} \right\},$$
$$\xi_3 = \left\{ S^{(1)} = \{(-1,1),(0,1)\}, S^{(2)} = \{(1,1),(1,0),(1,-1),(0,-1),(-1,-1),(-1,0)\} \right\},$$

gali būti laikomi kaip reikšmingų SLD pavyzdžiai, kadangi SLD ξ_1 atitinka situaciją, kai $N_1 = 4$, SLD ξ_2 atitinka situaciją, kai $N_1 = 3$, ir SLD ξ_3 atitinka situaciją $N_1 = 2$. Šie SLD yra pavaizduoti pav. 2.4



Pav. 2.4 Skirtingi SLD su $S^{(1)}$ ir $S^{(2)}$ taškais, atitinkamai pažymėtais • ir *

Aukščiau aprašytomis sąlygomis yra apskaičiuotos EBER reikšmės skirtingoms klasterizacijos ir pločio parametrų reikšmėms. Lentelėse 2.6 ir 2.7 yra pateikti rezultatai atvejais, $\Delta_0 = 0.2$ ir $\Delta_0 = 1$, atitinkamai.

Lentelėje 2.6 yra matyti, kad EBER silpnai atskirtoms klasėms ($\Delta_0 = 0{,}2$) yra monotoniškai mažėjanti didėjant α reikšmei prie fiksuotos parametro ρ reikšmės. Taip pat galima teigti, kad stebiniai su stipresne erdvine priklausomybe gali būti klasifikuojami tiksliau.

Panašios tendencijos EBER priklausomybėje nuo pločio parametro α ir klasterizacijos parametro ρ griežtai atskirtoms klasėms ($\Delta_0 = 1.0$) yra pavaizduotos lentelėje 2.7.

Lentelė 2.6 Reikšmės P_{0n} su $\Delta_0 = 0{,}2$ ir skirtingomis α, ρ reikšmėmis

ρ \ α	0.5	1	1.5	2	2.5	3
0	0.45877	0.45110	0.44271	0.43503	0.42805	0.42166
0.4	0.44698	0.44095	0.43394	0.42725	0.42100	0.41518
0.8	0.42014	0.41655	0.41191	0.40712	0.40240	0.39783
1.2	0.38851	0.38656	0.38372	0.38055	0.37725	0.37392
1.6	0.35633	0.35537	0.35374	0.35173	0.34950	0.34715
2	0.32529	0.32490	0.32405	0.32285	0.32140	0.31979

Lentelė 2.7 Reikšmės P_{0n} su $\Delta_0 = 1$ ir skirtingomis α, ρ reikšmėmis

ρ \ α	0,5	1	1,5	2	2,5	3
0	0,30237	0,26946	0,23562	0,20671	0,18229	0,16151
0,4	0,30178	0,26900	0,23525	0,20641	0,18204	0,16129
0,8	0,30005	0,26761	0,23415	0,20551	0,18129	0,16065
1,2	0,29728	0,26539	0,23237	0,20405	0,18006	0,15961
1,6	0,29362	0,26242	0,22999	0,20209	0,17842	0,15822
2	0,28923	0,25884	0,22711	0,19971	0,17642	0,15651

Skaitinė analizė atlikta mažoms mokymo imtims parodė, jog didesnė priklausomybė tarp klasių žymių ir stipresnė erdvinė koreliacija tarp požymių stebinių užtikrina mažesnes EBER reikšmes. Taigi galima tikėtis panašių priklausomybių ir kitiems erdvinės koreliacijos požymių modeliams ir labiau sudėtingiems žymių skirstiniams.

2.10. Antrojo skyriaus išvados

- Vidutinė Bajeso klaidos tikimybė mažėja didėjant imties dydžiui.
- EBER mažėjimo greitis yra didesnis prie didesnių erdvinės koreliacijos pločio parametro α reikšmių.
- EBER silpnai atskirtoms klasėms ($\Delta_0 = 0,2$) yra monotoniškai mažėjanti didėjant α reikšmei prie fiksuotos parametro ρ reikšmės. Taip pat

galima teigti, kad stebiniai su stipresne erdvine priklausomybe gali būti klasifikuojami tiksliau.

- Panašios tendencijos EBER priklausomybėje nuo pločio parametro α ir klasterizacijos parametro ρ griežtai atskirtoms klasėms $(\Delta_0 = 1.0)$ yra pavaizduotos lentelėje 2.7.
- Skaitinė analizė atlikta mažoms mokymo imtims parodė, jog didesnė priklausomybė tarp klasių žymių ir stipresnė erdvinė koreliacija tarp požymių stebinių užtikrina mažesnes EBER reikšmes.Taigi galima tikėtis panašių priklausomybių ir kitiems erdvinės koreliacijos požymių modeliams ir labiau sudėtingiems žymių skirstiniams.

3

Pasiūlytos metodikos taikymas ir eksperimentiniai rezultatai

Skyrelyje yra taikoma pasiūlyta metodika vaizdams, sugadintiems erdvėje koreliuoto triukšmo. Čia atliekamas juodai balto vaizdo, sugadinto su erdvėje koreliuotu triukšmu, rekonstravimo pavyzdys. Taip pat pateiktas palydovinės nuotraukos vaizdo klasifikavimas. Skyriaus tematika yra paskelbti keturi autorės straipsniai [2A], [4A], [5A]. Šis skyrius yra papildytas naujais skaičiavimais. Papildomai yra atliktas klasifikavimas, nuotolinio stebėjimo vaizdo, natūraliai sugadinto debesimis.

3.1. Juodai balto vaizdo rekonstravimo pavyzdys, naudojant BDF ir PBDF

Šiame skyrelyje yra lyginamas klasifikavimo taisyklių, susietų su diskriminantinėmis funkcijomis (formulės 2.4, 2.5, 2.7 ir 2.8), veikimas. Rezultatai pateikiami skaitiškai ir vizualiai. Yra aptariamas skaičiaus paveiksliuko, sugadinto su stacionariu GRF ir su izotropine eksponentine kovariacija, klasifikavimo pavyzdys. Laikoma, kad sąlyginis skirstinys $Y(s_0)$ duotam $Y=y$ priklauso tik nuo kaimyninių žymių lokacijose $N_0=NN(4)$, t.y.,

$$\pi_1(y) = 1\Big/\Big(1+\exp\big(\rho\big(1-2j/4\big)\big)\Big), \quad j=0,1,\ldots,4,$$

kur ρ – klasterizacijos parametras, o j – lokacijų skaičius iš N_0 su žymėmis, lygiomis 1 (Stabingienė *et al.* 2010).

Pagrindinė problema, kaip buvo minėta, yra stacionaraus GRF stebinio Z_0 klasifikavimas. Ši problema sprendžiama pasitelkiant Bajeso diskriminantines

funkcijas. Primenama, jog formulės 2.4 ir 2.5 – tai Bajeso diskriminantinės funkcijos, klasifikuojančios stebinį, *nepriklausomą* nuo mokymo imties. Pirmoji, 2.4, taikoma atveju, kai populiacijos parametrai žinomi, o antroji, 2.5, – nepilnai žinomų parametrų atveju. Likusios formulės 2.7 ir 2.8 – tai BDF, klasifikuojančios stebinį *priklausomą* nuo TS, atitinkamai, su tikrais parametrais ir su įvertiniais.

Šioje dalyje realizuojami ankstesniuose skyriuose aptarti metodai, siekiant parodyti BDF-jų, atsižvelgiančių į klasifikuojamo stebinio erdvinę priklausomybę nuo TS, pranašumą prieš kitas, ignoruojančias šią priklausomybę. Juodai balto vaizdo rekonstravimo pavyzdys atliekamas su griežtai atskirtų klasių paveiksliuku, ant kurio kaip papildomas triukšmas uždedamas erdvėje koreliuotas laukas. Klasifikavimo tikslumui nusakyti eksperimentų metu naudojamas klaidingo klasifikavimo tikimybių įvertinys.

Eksperimentams pademonstruoti naudojamas statistinis paketas R. Kadangi tiriami metodai yra nauji, tai standartinių paketų jiems realizuoti nėra. Visi metodai užprogramuoti naudojant standartines statistinio paketo R funkcijas. Papildomai naudojami šie paketai:

geoR – darbe naudojamas Gauso atsitiktinių laukų generavimui.

rtiff – skirtas darbui su tiff formato paveiksliukais, kurie dažnai naudojami palydovinių nuotraukų saugojimui. Šis formatas yra geras tuo, jog šiuo formatu išsaugotas paveiksliukas nepraranda informacijos, kas nutinka saugant JPEG formatu. Taip pat svarbu ir tai, jog programos R aplinkoje šio tipo duomenis galima nesudėtingai nuskaityti ir naudoti tolimesniuose skaičiavimuose.

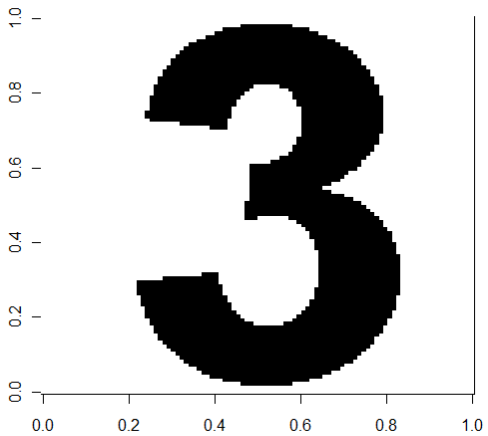
pixmap – tai būtina rtiff paketo dalis, kuri skirta papildomam darbui su paveiksliukais (paveiksliukų atvaizdavimui).

RSEIS – veiksmų su matricomis palengvinimui. Šis paketas naudojamas matricos transformavimui, kai reikia *y* ašį apversti simetriškai ikso ašies atžvilgiu. Tai būtina norint tinkamai apdoroti paveiksliukus, kadangi nuskaityto tiff formato paveiksliuko ordinačių ašis yra nukreipta iš viršaus į apačią, o laukai yra generuojami ant normalios Dekarto koordinačių sistemos.

Paveiksliukų iškirpimui ir išsaugojimui tiff formatu galima naudoti bet kurį grafinį redaktorių, kuris turi išsaugojimo tiff formatu galimybę.

Eksperimento tiklas yra ištirti įprastai naudojamų Bajeso diskriminantinių funkcijų ir pasiūlytos metodikos veikimą. Tai atkiekama tokiais etapais: pradinio paveikslėlio parinkimas, GRF generavimas, duomenų sujungimas, TS sudarymas, klasifikavimas su pasiūlyta metodika ir su įprastai naudojamomis BDF, klaidų vertinimas.

Šio eksperimento metu tiriamas klasifikavimo tikslumas panaudojant sugadintą paveiksliuką, kuris sugadinamas uždedant ant jo erdvėje koreliuotą Gauso atsitiktinį lauką. Pradinis paveiksliukas naudojamas su griežtai atskirtomis dvejomis klasėmis. Kadangi ankstesniuose skyriuose aprašytas klasifikavimas yra taikomas dviejų klasių atvejui, tai eksperimentavimui puikiai tinka skaičiaus paveiksliuko, taškų klasifikavimas. Čia naudojamas skaičiaus „3“ paveiksliukas, kurį klaidingai suklasifikavus galima gauti kitus skaičius, tokius kaip „8“ ar 9“. Be to, jis naudojamas daugelyje vaizdų klasifikavimo, vaizdų atstatymo (restauravimo), ir vaizdų atpažinimo straipsniuose. Tam kad būtų pakankamai didelis skaičius klasifikuojamų taškų, pradinį paveiksliuką atvaizduojame 100x100 taškų dydžio, tai sudaro 10000 taškų, kuriuos reikės klasifikuoti. Pradinis paveiksliukas pavaizduotas 3.1 paveikslėlyje



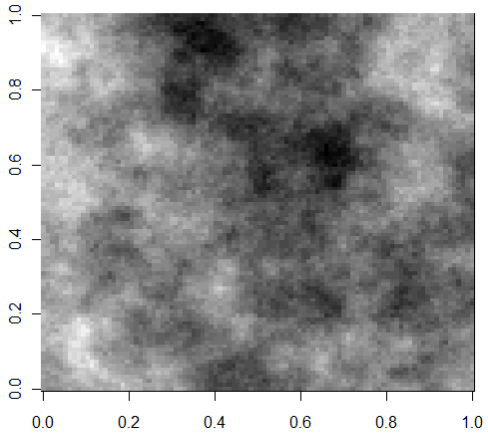
3.1 Pav. Pradinis paveiksliukas

Paveiksliukas yra sukuriamas grafinio apdorojimo programa ir išsaugomas TIFF formatu. Su *rtiff* praplėtimu nuskaičius paveiksliuko duomenis, gauname

tris matricas, kurių kiekviena atitinka skirtingų RGB spalvų reikšmių matricas. Kadangi naudojamas juodai baltas paveiksliukas, tai visų šių trijų spalvų reikšmių matricos sutampa, todėl galima naudoti bet kurią iš jų. Kaip buvo minėta, rtiff priemonėmis nuskaityto paveiksliuko koordinačių sistema skiriasi nuo įprastos Dekarto koordinačių sistemos, todėl, po duomenų nuskaitymo, jas transformuojame.

Pateiktas paveiksliukas yra juodai baltas, todėl naudojama tik vienos spalvos duomenų matrica, kurios kiekviename elemente, nuskaičius duomenis, yra reikšmės nuo 0-io iki 1-o. Viso šis intervalas yra padalintas į 256 vienodo dydžio lygius. Kuo taško spalva yra šviesesnė, tuo šį tašką atitinkanti matricos elemento reikšmė yra arčiau vieneto. Taigi, 3.1 iliustracijoje pavaizduotame paveiksliuke, skaičiaus taškus atitinka reikšmės lygios nuliui, o aplink jį visos reikšmės lygios vienetui.

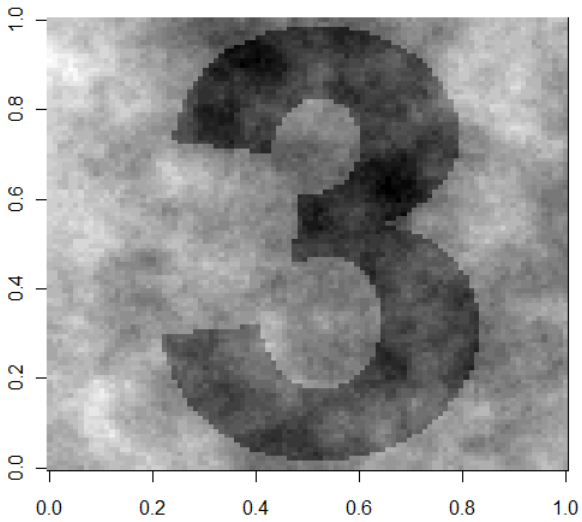
Pasirinkus pradinį paveiksliuką, sekanti užduotis – GRF generavimas. Kadangi paveiksliukas yra 100 x 100 taškų dydžio tai ir sugeneruotas Gauso laukas turi atitikti šį dydį. Pasinaudojant geoR paketo funkcija grf sugeneruojamas 100 x 100 taškų dydžio atsitiktinis Gauso laukas. Sugeneruotas laukas pateiktas 3.2 paveikslėlyje. Atsitiktinis Gauso laukas sugeneruotas su parametrais: μ vidurkis = 0,2; σ^2 =0,03; koreliacijos plotis α =2; naudojamas eksponentinis kovariacijos modelis. Taigi, šis laukas, sujungtas su ankstesniu paveiksliuku atitiks atsitiktinę modelio dalį



3.2 Pav. Sugeneruotas atsitiktinis Gauso laukas

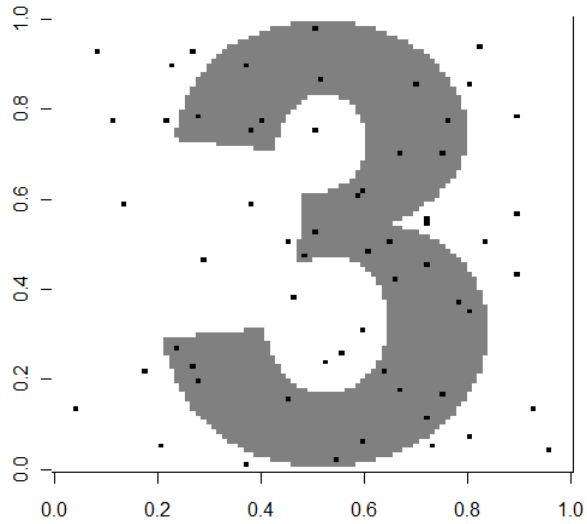
Taip pat svarbu paminėti, jog laukas grf funkcija nevisiškai tiksliai sugeneruoja atsitiktinį lauką, todėl jis generuojamas kelis kartus, tam, kad jame vizualiai matytųsi erdvinė priklausomybė. Kadangi toliau eksperimentuojama keletą kartų, tinkamai sugeneruotą lauką būtina išsisaugoti ir toliau jį naudoti kas kartą nuskaitant iš išsaugoto failo.

Tam, kad imituoti realų paveiksliuko panaudojimą, jį reikėtų suformuoti taip, kad matricos elementų reikšmės būtų intervale [0, 1]. Ši sąlyga nėra būtina klasifikavimui. Taigi, prieš susumuojant dviejų matricų (realaus paveikslio ir atsitiktinio Gauso lauko) elementų reikšmes, realaus paveikslio reikšmes padaliname iš 10. Po to prie šių duomenų pridedame atsitiktinio lauko atitinkamų matricos elementų reikšmes. Gautas sugadintas paveiksliukas pateikiamas pav. 3.3. Toks duomenų pertvarkymas yra svarbus tam, jog tokiu būdu atsitiktinio lauko įtaka paveiksliukui yra labai didelė. Jei paimtume nedidelę įtaką, su tokiu sugadintu paveiksliuku nesunkiai susidorotų daugelis klasikinių klasifikavimo metodų. Paveiksle 3.3 uždėtas laukas yra labai stiprus paveiksliuko atžvilgiu, nors vizualiai mes ir galime atskirti trejetą, tačiau programiniu apdorojimu jį tiksliai išskirti būtų labai sudėtinga, todėl taikysime 2.4, 2.5, 2.7 ir 2.8 formulėse aprašytus klasifikavimo su mokymu metodus.



3.3 Pav. Sugadintas paveiksliukas.

Prieš pradedant klasifikavimą sudaroma mokymo imtis iš 60 taškų, kur kiekvienai skirtingai klasei yra po 30 taškų. Žemiau esančiame paveiksle 3.4 pateikta mokymo imtis.

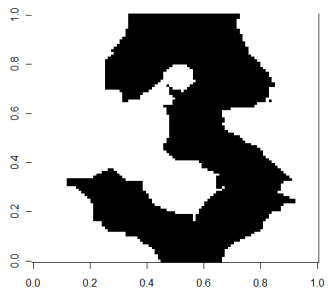
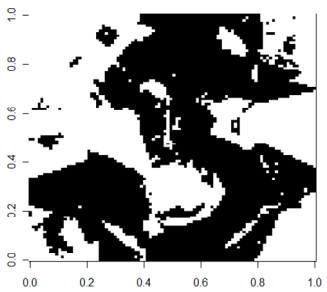
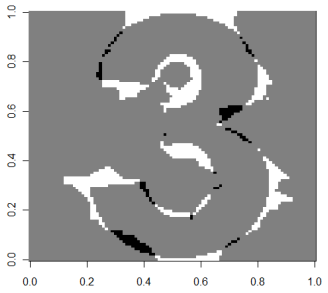
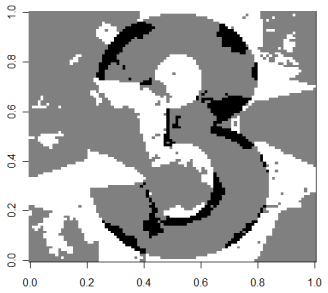


3.4 Pav. Mokymo imtis kiekvienai klasei po 30 taškų

Paveikslėliuose 3.5 yra pateiktas rezultatas, gautas suklasifikavus pagal 2.4 ir 2.5 formules. Klasifikavimas vykdomas pagal keturių artimiausių kaimynų NN4 modelį.

3.5 paveiksle pateiktas klasifikavimo rezultatas, kai populiacijos parametrai žinomi. Pirmoje eilutėje ir pirmame stulpelyje esantis paveiksliukas žymiai geriau suklasifikuotas už šalia esantį. Tai atlikta su BDF neignoruojančia erdvinės priklausomybės tarp klasifikuojamo stebinio ir mokymo imties. Žemiau esantys paveikslėliai pateikia klaidas vizualiai. Čia balta spalva reiškia neteisingai suklasifikuotų baltų taškų kiekį, o juoda spalva reiškia neteisingai suklasifikuotų juodų taškų skaičių, pilka parodo kiek taškų yra gerai suklasifikuota. Perdengimo paveikslėliai gaunami tokiu būdu: prie pradinio paveikslo pridedamas 1, tada iš jo atimamas suklasifikuotasis ir padalijama iš dviejų.

Vizualinis klaidų pateikimas nėra pakankamas, todėl tai nusakoma skaitiškai. Taigi klaidingo klasifikavimo vertinimui skaičiuojamos klaidų tikimybės: P(2|1), P(1|2). Šios tikimybės apibrėžtos skyrelyje 2.5. Pirmoji P(2|1) reiškia priskirti elementą antrai klasei, nors iš tikrųjų jis priklauso pirmajai. Tikimybė P(1|2) reiškia priskirti elementą pirmai grupei, nors jis priklauso antrai.

	BDF-formulė 2.7	BDFI-formulė 2.4
Suklasifikuota		
Perdengimas		

3.5 Pav. Klasifikavimo su BDF ir BDFI rezultatai

Taigi yra apskaičiuojamos tikimybės:

$$P(2|1)=P(0|1)=P\{\text{ Tikimybė priskirti }\textbf{\textit{juodam}}\text{, nors yra }\textbf{\textit{baltas}}\ \}=\frac{m_2}{n_2},$$

$$P(1|2)=P(1|0)=P\{\text{ Tikimybė priskirti }\textbf{\textit{baltam}}\text{, nors yra }\textbf{\textit{juodas}}\ \}=\frac{m_1}{n_1},$$

kur m_1 ir m_2 – neteisingai suklasifikuotų stebinių skaičius 1–je ir 2–je grupėse atitinkamai, n_1 ir n_2 – pirmos ir antros grupės elementų skaičius atitinkamai.

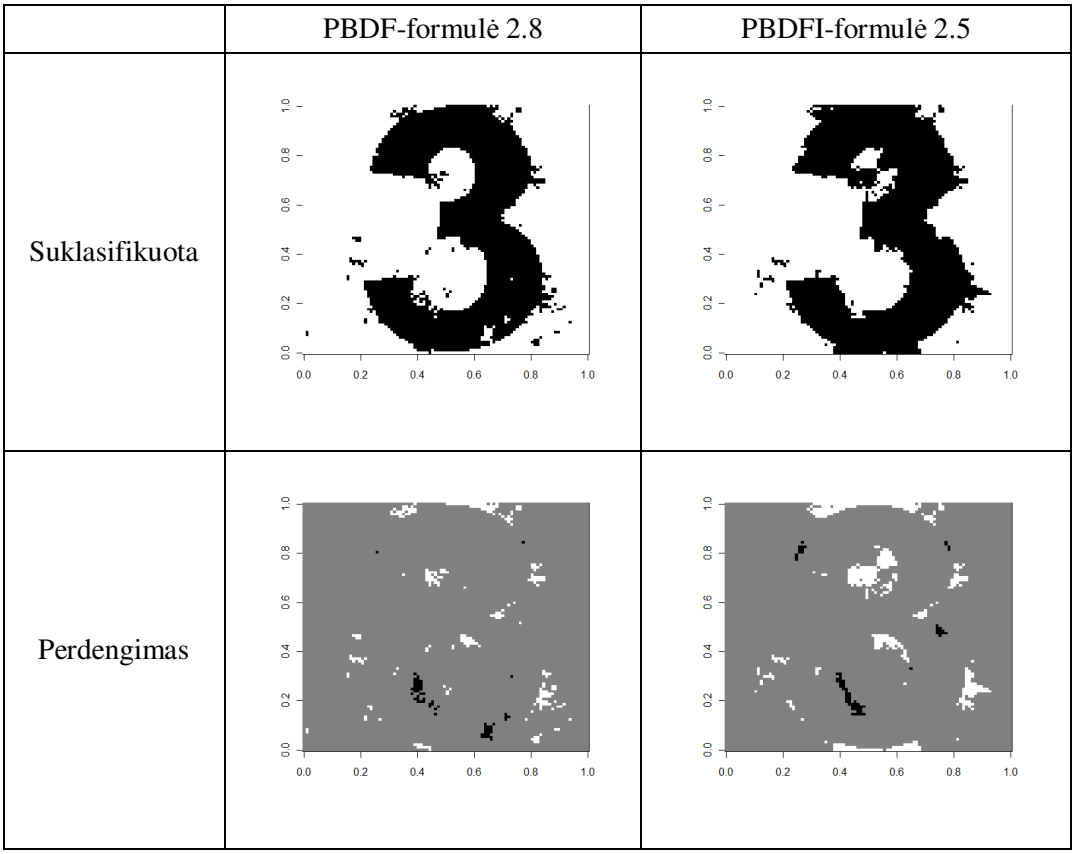
Iš 3.1 lentelės rezultatų matyti, kad erdvinės priklausomybės įvedimas į klasifikacijos problemą duoda gerus rezultatus.

Lentelė 3.1 Klaidingo klasifikavimo tikimybės

Formulė	BDF	BDFI
P(2|1)	0.0479	0.2165
P(1|2)	0.1251	0.4208

(Stabingienė 2010)

Aukščiau pateiktose realizacijose naudojamos tikrosios parametrų reikšmės. Tačiau praktiškai labai dažnai pasitaikanti situacija yra ta, jog tikrųjų parametrų reikšmių nežinome ir todėl yra reikalingi jų įvertiniai. Įterpus vidurkio ir dispersijos įvertinius į BDF yra gaunama PBDF.



3.6 Pav. Klasifikavimo su PBDF ir PBDFI rezultatai

Paveikslėlyje 3.6 yra pavaizduotas klasifikavimo rezultatas, naudojant PBDF dviem atvejais. Pirmu atveju, yra laikoma, kad klasifikuojamas taškas yra priklausomas nuo mokymo imties, o kitu atveju – nepriklausomas nuo mokymo imties. Matomas pranašumas atveju, kai yra atsižvelgiama į erdvinę priklausomybę.

Lentelė 3.2 Klaidingo klasifikavimo tikimybės

Formulė	PBDF	PBDFI
P(2|1)	0.0206	0.0236
P(1|2)	0.0388	0.0771

(Stabingienė 2010)

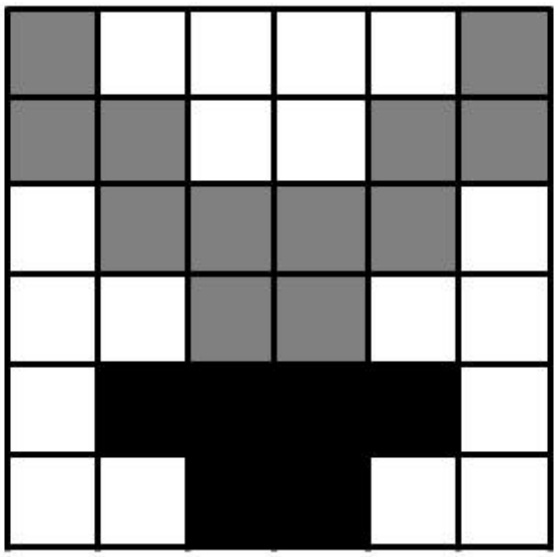
Iš lentelės 3.2 rezultatų matyti, kad klaidingo klasifikavimo tikimybės mažesnės tuo atveju, kai naudojame PBDF. Vadinasi erdvinės priklausomybės įvedimas į klasifikacijos problemą yra akivaizdžiai pasiteisinantis.

3.2. Klasifikavimas, paremtas pilkumo lygio pasikartojimų matricomis

Šiame skyrelyje aptariamas klasifikavimo be mokymo metodas, paremtas pilkumo lygio pasikartojimų matricomis (GLCM) (*angl. grey level co-occurrence matrix*). GLCM – tai taškų šviesumo (pilkumo) lygio reikšmių skirtingų kombinacijų pasikartojimo paveiksliuke skaičių lentelė (matrica). Šios matricos dažnai taikomos vaizdų analizėje. Tad remiantis šiomis matricomis darbe atliekamas klasifikavimas be mokymo realaus vaizdo atpažinimui. Lygiagrečiai tas pats vaizdas klasifikuojamas, naudojant pasiūlytą metodiką (klasifikavimo su mokymu).

Paprastai naudojamas spalvotas skaitmeninis paveiksliukas susideda iš trijų intensyvumo lygių, kurių kiekvienas atitinka vieną iš pagrindinių spalvų (raudona, žalia ir mėlyna). Esant didesniam intensyvumui spalva šviesesnė. Suvidurkinus kiekvieno paveiksliuko taško skirtingų spalvų intensyvumus yra gaunama pilka spalva (galimi ir kitokie spalvingumo panaikinimo būdai). Tokiems pilkiems paveiksliukams analizuoti dažniausiai ir yra taikoma pasikartojimų matricų analizė, bei jomis paremtas vaizdų bei vaizdų taškų klasifikavimas.

Praktikoje naudojami paveiksliukai yra koduojami diskrečiomis intensyvumo reikšmėmis. Spalva tamsesnė, kuo intensyvumo reikšmė mažesnė. Dažniausiai praktikoje naudojami paveiksliukai, kuriems kiekvienai spalvai atvaizduoti naudojamas 8-ių bitų kodavimas (256 skirtingi intensyvumo lygiai). Jei praktikoje būtų naudojami visi 256 lygiai, tai pasikartojimų matrica būtų didžiulė (256×256). Esant tokiai didelei matricai operacijos taptų sudėtingesnės ir tam tikrų perėjimų pasikartojimai pasitaikytų labai retai, todėl praktiškai taikant lygių skaičius sumažinamas (dažnai naudojami 32 skirtingi pilkumo lygiai). Pavyzdžiui 3 pilkumo lygių paveiksliukas atrodo taip, kaip pateikta paveiksle 3.7.



3.7 Pav. Paveikslas triejų pilkumo lygių

$$D = \begin{pmatrix} 1 & 2 & 2 & 2 & 2 & 1 \\ 1 & 1 & 2 & 2 & 1 & 1 \\ 2 & 1 & 1 & 1 & 1 & 2 \\ 2 & 2 & 1 & 1 & 2 & 2 \\ 2 & 0 & 0 & 0 & 0 & 2 \\ 2 & 2 & 0 & 0 & 2 & 2 \end{pmatrix}$$

Pasikartojimų kombinacijos fiksuojamos judant per paveikslą tam tikra kryptimi. Dažniausiai praktikoje pasikartojimų matricos skaičiuojamos visomis kryptimis atskirai. Vertinant charakteristikas ar klasifikuojant atsižvelgiama į kiekvieną iš jų. Kai kurie autoriai krypties įvedimą vertina kaip erdvinės informacijos įvedimą (Haralick *et al.* 1973). Dažniausiai naudojamos keturios skirtingos kryptys (0^0, 45^0, 90^0, 135^0 laipsnių kryptimis). Judant tam tikra kryptimi nustatomas pilkumo lygio pasikeitimas ir atitinkamas pasikartojimų matricos elementas padidinamas vienetu. Pavyzdžiui, jei aukščiau pateikto paveiksliuko matricai D skaičiuotume pasikartojimų matricą kryptimi *iš dešinės į kairę*, tai gautume sekančią pradinę pasikartojimų matricą (3×3 dydžio):

$$B = \begin{pmatrix} 4 & 0 & 2 \\ 0 & 6 & 4 \\ 2 & 4 & 8 \end{pmatrix}$$

Pradinės pasikartojimų matricos b_{ij} elementas parodo kiek kartų iš vieno taško einant į kitą tašką pasirinkta kryptimi pilkumo lygis pasikeitė iš vienos spalvos į kitą. Pavyzdžiui, matricoje elementas b_{11}=4 reiškia, kad perėjimų iš juodos celės į juodą celę yra keturi, kryptimi iš dešinės į kairę. Atitinkamai ta pačia kryptimi pereinant iš juodo į pilką bus 0, t.y., b_{12}=0 ir t.t. Charakteristikoms, kurios vertinamos pagal sudarytą pasikartojimų matricą, yra reikalinga simetrinė matrica, todėl pradinė pasikartojimų matrica yra simetrizuojama:

$$B_{sim} = B + B' .$$

Simetrizuota pradinė pasikartojimų matrica B_{sim} atvaizduoja pasikartojimus dviem priešingomis kryptimis (čia iš dešinės į kairę ir iš kairės į dešinę). Pasikartojimų tikimybėms nusakyti simetrizuota pradinė pasikartojimų matrica B_{sim} normuojama:

$$p_{ij} = \frac{b_{ij}}{\sum_{k,l=0}^{N-1} b_{kl}}$$

kur p_{ij} – normuotos pasikartojimų matricos elementas, b_{ij} – simetrizuotos pradinės pasikartojimų matricos elementas, N – pasikartojimų matricos eilučių arba stulpelių skaičius. Matrica su elementais p_{ij} vadinama pasikartojimų matrica. Šios matricos kiekvienas elementas p_{ij} parodo kiek yra tikėtinas perėjimas iš i-tojo pilkumo lygmens į j-jį pilkumo lygmenį, einant pasirinkta kryptimi. Gavus pasikartojimų matricą skaičiuojamos įvairios paveikslėlio charakteristikos. Žemiau pateiktos dažniausiai naudojamos charakteristikos (Adan *et al.* 2003), (Hall-Beyer 2007), (Haralick *et al.* 1973). Visoms charakteristikoms skaičiuoti pateiktos formulės atveju, kai paveikslėlio pilkumo lygiai pradedami skaičiuoti nuo 0–io.

Kontrastas (*angl. contrast*) parodo gretimų taškų skirtingumo lygį; kuo kontrastas didesnis, tuo didesni intensyvumų skirtumai tarp gretimų taškų. Kontrasto charakteristika apskaičiuojama pagal formulę (Hall-Beyer 2007):

$$\sum_{i,j=0}^{N-1} p_{i,j} \left(i - j\right)^2 . \qquad (3.1)$$

Nepanašumas (*angl. dissimilarity*) panaši charakteristika į kontrastą, ji priskiriama prie kontrasto charakteristikų grupės. Apibrėžiama tokiu būdu (Hall-Beyer 2007):

$$\sum_{i,j=0}^{N-1} p_{i,j} \left| i - j \right|. \tag{3.2}$$

Homogeniškumas (*angl.* homogeneity) – dar viena kontrasto grupės charakteristika:

$$\sum_{i,j=0}^{N-1} \frac{p_{i,j}}{1+\left(i-j\right)^2}. \tag{3.3}$$

Kampinis antrasis momentas (*angl. angular second moment* (ASM)):

$$\sum_{i,j=0}^{N-1} p_{i,j}^2. \tag{3.4}$$

Entropija:

$$\sum_{i,j=0}^{N-1} p_{i,j} \left(-\ln p_{i,j} \right). \tag{3.5}$$

Stulpelių vidurkis:

$$\mu_j = \sum_{i,j=0}^{N-1} j \left(p_{i,j} \right). \tag{3.6}$$

Eilučių vidurkis:

$$\mu_i = \sum_{i,j=0}^{N-1} i \left(p_{i,j} \right) \tag{3.7}$$

Stulpelių dispersija:

$$\sigma_j^2 = \sum_{i,j=0}^{N-1} p_{i,j} \left(j - \mu_j \right) \tag{3.8}$$

Eilučių dispersija:

$$\sigma_i^2 = \sum_{i,j=0}^{N-1} p_{i,j} \left(i - \mu_i \right) \tag{3.9}$$

Koreliacija skaičiuojama pagal formulę:

$$r = \sum_{i,j=0}^{N-1} p_{i,j} \left[\frac{(i-\mu_i)(j-\mu_j)}{\sqrt{(\sigma_i^2)(\sigma_j^2)}} \right]. \qquad (3.10)$$

Klasifikavime taikant pasikartojimų matricas yra tokia tvarka:

1. Parenkami pavyzdiniai klases atitinkantys paveikslėliai.
2. Pasirenkamos kryptys, pagal kurias skaičiuojamos pasikartojimų matricos. Galima naudoti visų krypčių pasikartojimų matricas.
3. Paimtiems paveikslėliams apskaičiuojamos pasikartojimų matricos.
4. Gautoms pasikartojimų matricoms apskaičiuojamos charakteristikos, kiekvienam pavyzdiniam paveikslėliui atskirai.
5. Apskaičiuojamos klasifikuojamų paveiksliukų arba jų fragmentų pasikartojimų matricos, bei nustatomos charakteristikos.
6. Klasifikuojamo taško aplinkos charakteristikos lyginamos su pavyzdinėmis ir pagal labiausiai atitinkančias reikšmes, visomis pasirinktomis kryptimis, klasifikuojamas paveiksliukas priskiriamas atitinkamai klasei.

Pasikartojimų matricos taikomos dviem atvejais, kada visas paveiksliukas priskiriamas tam tikrai klasei ir kada viso paveiksliuko atskiri taškai priskiriami tam tikrai klasei. Pastarasis metodas labai dažnai taikomas palydovinės informacijos klasifikavimui (kiekvienas taškas priskiriamas tam tikram teritorijos tipui). Šio klasifikavimo principas yra toks: apie klasifikuojamą tašką paimamas tam tikras kiekis taškų ir pagal juos atliekami pasikartojimų matricų charakteristikų vertinimai, ir, tuom remiantis tam taškui priskiriama klasė. Dažniausiai klasifikavimui naudojamas 7×7 lango dydis, tačiau jo dydžio pasirinkimas priklauso ir nuo kitų faktorių, tokių kaip: paveiksliuko rezoliucija (kaip detaliai pateikiamas vaizdas), klasifikavimui naudojamos charakteristikos (skirtingos charakteristikos geriau klasifikuoja prie skirtingų lango dydžių (Hall-Beyer 2007)) ir pan.

Taigi, skyrelyje aptartas metodas naudojamas tik palyginimui, o visas dėmesys darbe sutelktas į klasifikavimą su mokymu, paremtą Bajeso diskriminantinėmis funkcijomis.

3.3. Palydovinės nuotraukos vaizdo klasifikavimas

Vaizdų klasifikavime dažnai sutinkame situaciją, kai, tam tikru lygiu vaizdai yra sugadinti triukšmo. Toks triukšmas vaizdų klasifikavime gali būti modeliuojamas Gauso atsitiktiniais laukais. Statistiniame vaizdų klasifikavime naudojami metodai su mokymu ir be mokymo. Šiame skyriuje lyginami pasiūlyti klasifikavimo su mokymu metodai, paremti įterptomis Bajeso diskriminantinėmis funkcijomis (*see* Dučinskas 2009 and Stabingienė *et al.* 2010), su klasifikavimo be mokymo metodu, paremtu GLCM (Haralick *et al*. 1979 and Adan *et al*. 2003). Klasifikavimui naudojamas palydovinės nuotraukos vaizdas iš NASA palydovo LANDSAT 7 (USGS Earth Explorer). Paveikslėlis vaizduoja vakarų Lietuvos teritoriją. Taip pat sugeneruojami GRF su skirtingais koreliacijos pločiais ir uždedami ant palydovinės nuotraukos. Tokia situacija gali natūraliai susidaryti degant miškui, kai gaisro dūmai uždengia tam tikrą dalį teritorijos. Šie paveiksliukai naudojami klasifikavimo tikslumo tyrimui.

Lentelė 3.3 Eksperimente naudojamo Landsat 7 palydovo žemėlapio informacija.

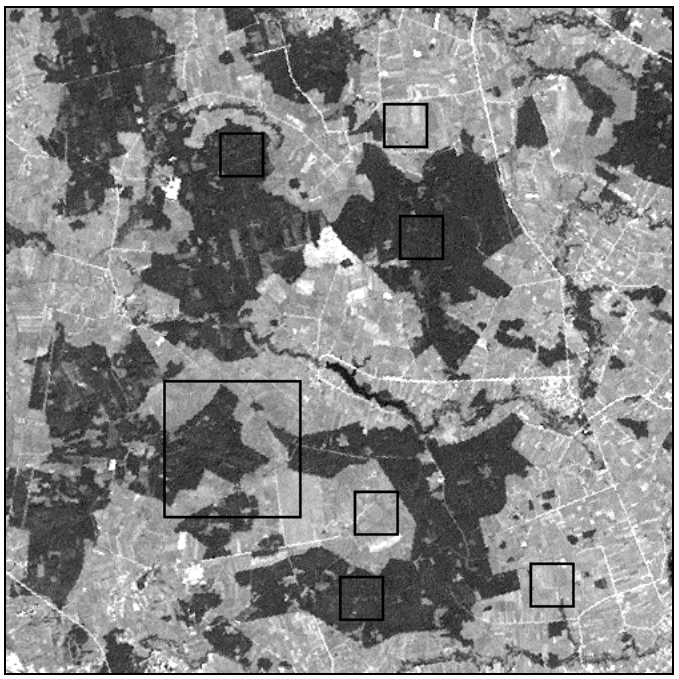
Žemėlapio identifikacinis numeris	ELP188R021_7T20010729
Centro koordinatės	°54'52.14"N, 22°31'04.47"E
Žemėlapio nufotografavimo data	2001.07.29

LANDSAT 7 palydovas nuotraukas daro septyniose skirtingų dažnių juostose. Jo nuotraukos viename taške užfiksuojama 30m×30m pločio teritorija. Taip pat su žemėlapiais papildomai yra pateikiamas dar vienas paveiksliukas, kurį uždėjus ant šių žemėlapių vaizdas tampa dvigubai detalesnis (vienas žemėlapio taškas tada atvaizduoja 15m×15m teritoriją). Darbe eksperimentai atliekami su detalizuotu žemėlapiu ir kadangi PBDF bei PBDFI formulės kol kas pritaikytos tik vienmačiam atvejui, tai naudojamas tik vienos dažnių juostos

žemėlapis. Sekančioje lentelėje pateikiama tiksli naudojamo žemėlapio informacija.

Kadangi eksperimentų tikslas nėra suklasifikuoti visą žemėlapį, o tik ištirti įvairių metodų veikimą prie tam tikrų sąlygų, tai eksperimentams naudojamas iš viso žemėlapio iškirptas 500×500 taškų dydžio gabaliukas (pav. 3.8), kuriame nemažą dalį teritorijos sudaro miškai. Šis paveiksliukas vėliau naudojamas kaip pradinis paveiksliukas ant kurio uždedami sugeneruoti erdvėje koreliuoti laukai. Šio eksperimento metu Gauso laukai generuojami naudojant eksponentinę koreliacinę funkciją, o erdvinės koreliacijos pločio parametrai $\alpha\in\{1,10,50\}$. Kadangi paveiksliuke taškai skaičiuojami vienetais, tai reiškia, jog žemėlapiui ant kurio uždėtas laukas su pločiu lygiu 50, lauko erdvinė priklausomybė išlieka per 50 paveiksliuko taškų. Paveiksliukų sujungimas atliekamas analogiškai ankstesniam eksperimentui. Generuojamas GRF laukas yra 500×500 dydžio.

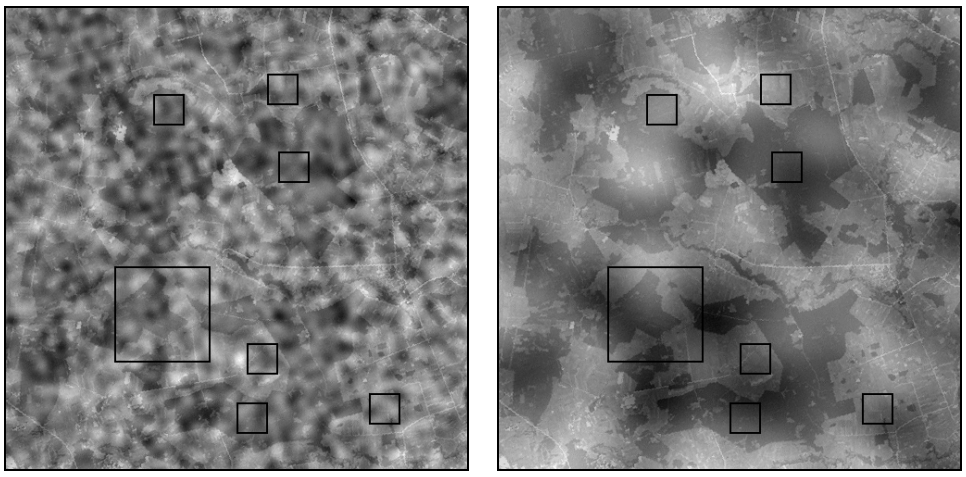
Šio eksperimento metu klasifikavimas atliekamas pagal tuos pačius klasifikavimo metodus, kaip ir ankstesniame eksperimente. Visi šie metodai priklauso klasifikavimo su mokymu metodams. Šiame eksperimente taip pat yra taikomas ir klasifikavimo be mokymo metodas paremtas pasikartojimų matricomis, kuris buvo aptartas aukščiau esančiame skyrelyje. Klasifikavimui be mokymo yra reikalingi klasių pavyzdiniai paveiksliukai. Šie paveiksliukai yra iškerpami iš to pačio 500×500 taškų dydžio paveiksliuko. Iš šio pradinio paveiksliuko, iš fiksuotos vietos, taip pat iškerpama ir 100×100 taškų dydžio paveiksliukai naudojami klasifikavimui (3 iš miško ir 3 iš pievos). Pradinis 500×500 taškų dydžio paveiksliukas ir pavyzdiniai paveiksliukai, kurie bus naudojami GLCM metodui, pavaizduoti pav. 3.8.



Pav.3.8 Eksperimente naudojama viso paveiksliuko dalis 500×500 taškų. Didysis kvadratas žymi klasifikavimui naudojamą paveiksluko dalį, o mažieji kvadratai žymi pavyzdinius klasių paveiksliukus, naudojamus metodui, paremtam GLCM.

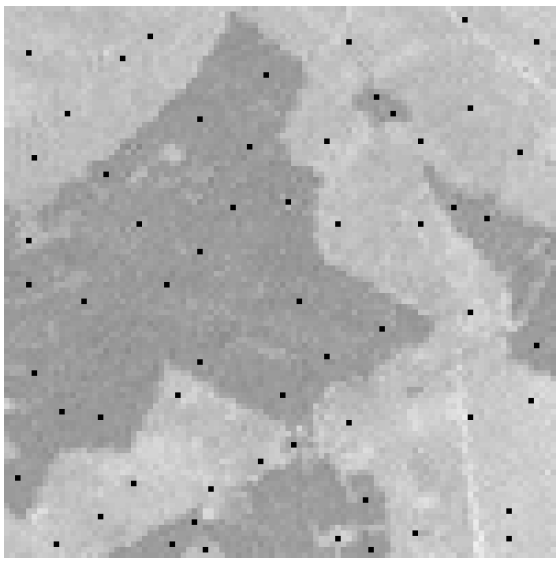
Kaip buvo minėta, generuojamas GRF laukas yra 500×500 dydžio ir jis sujungiamas su pradiniu paveiksliuku. Klasifikavimui naudojami paveiksliukai iškerpami jau iš šių sujungtų paveiksliukų (GLCM metodui naudojami pavyzdiniai paveiksliukai iškerpami taip pat iš šių, erdvėje koreliuotu lauku perdengtų paveiksliukų). Šioje vietoje reikėtų paminėti, jog GLCM metodas labiau remiasi paveiksliuko tekstūros požymiais ir klasifikuojama yra pagal tai, kurios klasės pavyzdinė tekstūra yra panašesnė į klasifikuojamo taško aplinką. Atsižvelgiant į tai būtų galima taikyti vien tik pradinio paveiksliuko pavyzdinių paveiksliukų informaciją, tačiau persidengęs laukas taip pat pakenkia ir tekstūrai, ji tampa šviesesnė, pasikeičia perėjimų informacija. Taigi šiame darbe klasifikavimas paremtas GLCM atliekamas pavyzdinius klasių paveiksliukus paimant iš jau sugadintų paveiksliukų. Tokie pavyzdiniai paveiksliukai gali būti formuojami ir taikant praktikoje, jei žinome kad teritoriją dengia tam tikras triukšmas. Žemiau pateikti pradiniai paveiksliukai su sugeneruotais skirtingos erdvinės koreliacijos pločio α laukais (pav. 3.9)

Klasifikavimo su mokymu metodams reikalinga mokymo imtis pavaizduota pav. 3.10. Kiekvienai klasei, miškui ir pievai, yra suformuojama po 30 mokymo imties taškų. Mokymo imtis (TS) iš 60 taškų. Klasifikavimas su mokymu, naudojant PBDF ir PBDFI metodus, atliekamas naudojant NN8 aštuonių artimiausių kaimynų schemą.



$\alpha = 10$ $\alpha = 50$

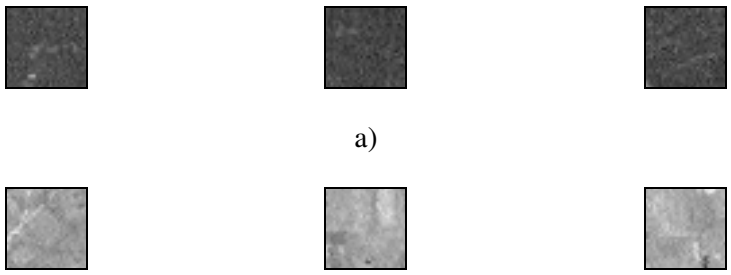
Pav. 3.9 Paveiksliukai su skirtingais laukais.



Pav. 3.10 Mokymo imties taškai, naudojami klasifikavimo su mokymu metodams.

Klasifikavimui be mokymo, naudojant GLCM, iškirpti pavyzdiniai klasių paveiksliukai, realaus paveiksliuko (be uždėto erdvėje koreliuoto lauko), yra pavaizduoti pav. 3.11. Kadangi šio eksperimento metu lyginami klasifikavimo

su mokymu ir be mokymo metodai, tai papildomai atliekamas realaus paveiksliuko klasifikavimas. Ši eksperimento dalis yra naudinga todėl, jog šio eksperimento metu pradinės paveiksliuko klasės nėra labai griežtai atskirtos ir tokio (pradinio) paveiksliuko suklasifikavimo rezultatų palyginimas yra naudingas analizuojant metodų veikimą.



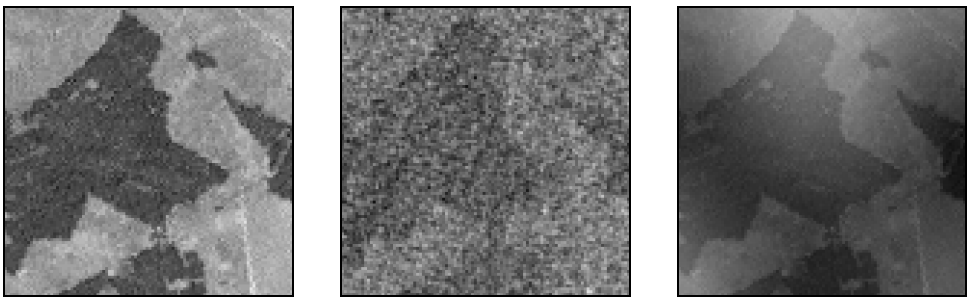
a)

b)

Pav. 3.11 GLCM paremtiems metodams naudojami pavyzdiniai klasių paveiksliukai. a) miško klasės paveiksliukai, b) pievos klasės paveiksliukai.

Klasifikavimas su GLCM buvo atliktas pagal keletą skirtingų charakteristikų, tačiau geriausius rezultatus pavyko pasiekti klasifikuojant pagal eilučių vidurkių charakteristiką (formulė 3.7). Klasifikavimo tikslumas nustatomas vertinant empirines klaidingo klasifikavimo tikimybes.

Paveiksle 3.12 yra parodyti paveiksliukai, kuriems buvo atliktas klasifikavimas abiem aptartais atvejais.



Realus paveiksliukas α=10 α=50

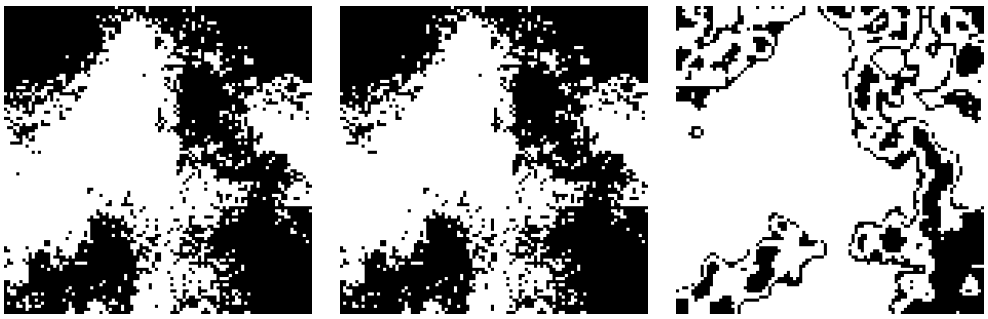
Pav. 3.12 Klasifikavimui naudojami paveiksliukai.

Paveiksle 3.13 yra pateikti gauti vaizdiniai rezultatai atlikus klasifikavimą su PBDF (siūloma metodika), įprastai naudojamomis PBDFI ir su klasifikavimo be mokymo metodu, paremtu pilkumo lygio pasikartojimų matricomis.

Analogiškai atliktas klasifikavimas su aptartais metodais, esant vaizdui sugadintam erdvėje koreliuoto triukšmo su erdvinės koreliacijos pločiu α=1. Vertinat vizualiai yra matyti, kad metodas paremtas GLCM klasifikuoja prasčiau.

PBDF PBDFI GLCM

Pav. 3.13 Klasifikavimo rezultatai, kai nėra uždėto jokio papildomo koreliuoto lauko.



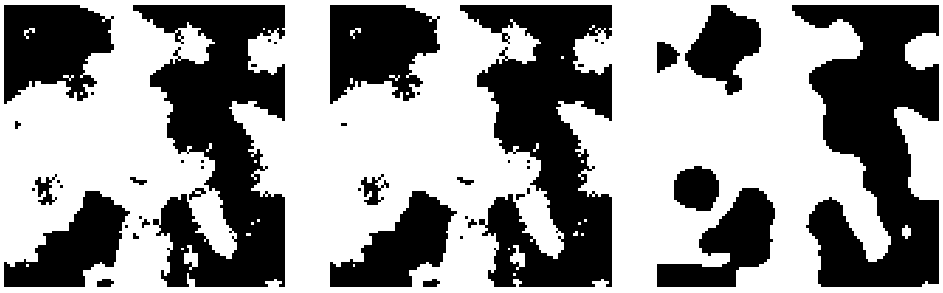
PBDF PBDFI GLCM

Pav. 3.14 Klasifikavimo rezultatai, kai uždėtas erdvėje koreliuotas laukas su α=1.

Sekančiame paveikslėlyje 3.15 pateikti klasifikavimo rezultatai, kai vaizdas sugadintas erdvėje koreliuoto triukšmo su koreliacijos pločio parametru α=10. Iš čia matyti, kad klasifikavimo su mokymu metodai elgiasi panašiai, tačiau vizualiai vertinti kokybę yra sunkoka. Grubiai tariant, galima sakyti, kad PBDF ir PBDFI „elgiasi“ geriau prie didesnės α reikšmės, t.y., prie α=10, nei α=1 (pav. 3.15)

Paveiksle 3.16 pateikti klasifikavimo rezultatai vaizdo, sugadinto su erdvėje koreliuotu triukšmu, esant koreliacijos pločiui 50. Akivaizdžiai matyti, kad

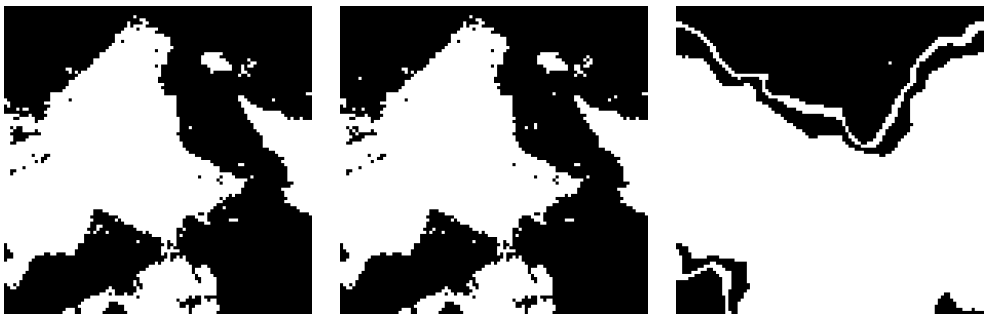
klasifikavimo su mokymu metodas lenkia metodą, paremtą GLCM. Esant dideliam lauko koreliacijos pločiui α=50 PBDF ir PBDFI metodai beveik nebereaguoja į uždėto lauko keliamą triukšmą.



PBDF PBDFI GLCM

Pav. 3.15 Klasifikavimo rezultatai, kai uždėtas erdvėje koreliuotas laukas su α=10.

GLCM matricomis paremtas metodas yra labai jautrus tuo atveju kada atsiranda papildomas erdvėje koreliuotas triukšmas.



PBDF PBDFI GLCM

Pav. 3.16 Klasifikavimo rezultatai, kai uždėtas erdvėje koreliuotas laukas su α=50.

Iš lentelėje 3.4 pateiktų rezultatų matosi PBDF metodo pranašumas prieš PBDFI metodą. PBDF metodo klasifikavimo tikslumas yra didesnis, o klaidingo klasifikavimo tikimybių įverčiai mažesni. Nors skirtumai nėra labai dideli, tačiau iš to galime teigti, jog atsižvelgiant į klasifikuojamo taško erdvinę priklausomybę su mokymo imtimi gaunami geresni klasifikavimo rezultatai.

Taip pat galime pastebėti, jog šiek tiek gerėja klasifikavimas, naudojant PBDF ir PBDFI, kai erdvinės koreliacijos plotis didėja, tačiau kadangi gene-

ruojami laukai skiriasi vienas nuo kito ne vien erdvinės koreliacijos pločiu, bet ir skirtingų teritorijų nevienodu perdengimu (nevienodu sugadinimu skirtingose vietose), tai klasifikavimo tikslumas gali būti įtakojamas ir kitų faktorių.

Dar svarbu pastebėti, jog klasifikuojant realų paveiksliuką, su užduotimi puikiai susidorojo ir GLCM metodai, tačiau duomenyse atsiradus erdvėje koreliuotam triukšmui šie metodai nebetenka prasmės, nes jie pradeda klasifikuoti ne paveiksliuko teritoriją, o lauko informaciją.

Lentelė 3.4 Empirinės klaidingo klasifikavimo klaidos. OI – originalus vaizdas, be GRF.

α	PBDF		PBDFI		GLCM	
	$\hat{P}_{(2\|1)}$	$\hat{P}_{(1\|2)}$	$\hat{P}_{(2\|1)}$	$\hat{P}_{(1\|2)}$	$\hat{P}_{(2\|1)}$	$\hat{P}_{(1\|2)}$
1	0.213	0.093	0.210	0.093	0.439	0.019
10	0.217	0.073	0.222	0.077	0.323	0.142
50	0.150	0.065	0.154	0.067	0.477	0.265
OI	0.036	0.015	0.036	0.015	0.050	0.035

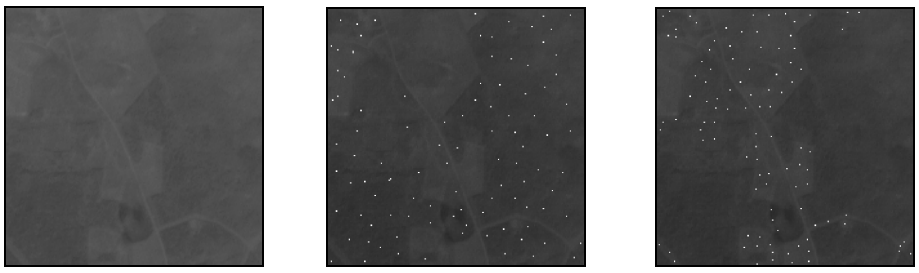
(Stabingienė *et al.* 2011)

Esant dideliam lauko koreliacijos pločiui α PBDF ir PBDFI metodai beveik nebereaguoja į uždėto lauko keliamą triukšmą.

3.4. Klasifikavimas, realaus nuotolinio stebėjimo vaizdo, padengto debesimis

Šiame pavyzdyje siūloma metodika yra pritaikoma realiam nuotolinio stebėjimo vaizdui, kuris yra natūraliai padengtas debesimis. Šis paveiksliukas taip pat, kaip ir ankstesniame pavyzdyje, aprašytame 3.3 skyrelyje, yra gautas iš Landsat7 palydovo (USGS Earth Explorer), kuriame yra Lietuvos teritorijos vaizdas. Eksperimente naudojama tik dalis viso paveikslėlio (200x200 taškų). Jame pavaizduotos dvi klasės (pirmoji klasė yra miškas, o antroji ne miškas) (pav. 3.16a). Originalus paveiksliukas yra natūraliai padengtas (sugadintas) debesimis, ir šis, debesis atitinkantis triukšmas, yra modeliuojamas Gauso atsitiktiniu lauku, su nuliniu vidurkiu ir eksponentine erdvinės koreliacijos funkci-

ja, aprašoma formule $r(h) = \exp\left\{-|h|^2/\alpha\right\}$. Čia α yra erdvinės koreliacijos pločio parametras, kurį reikia įvertinti. Šio parametro įvertinimui naudojama geoR paketo variofit komanda, R programoje (r-project). Klasifikavimo su mokymu PBDF ir PBDFI metodams paimama mokymo imtis iš n_1=n_2=100 taškų, kurie pateikti paveikslėliuose 3.16 a) ir 3.16 b). Klasifikavimui naudojamos 4-ių, 8-ių ir 12-os artimiausių kaimynų kaimynystės schemos (pav. 3.18).



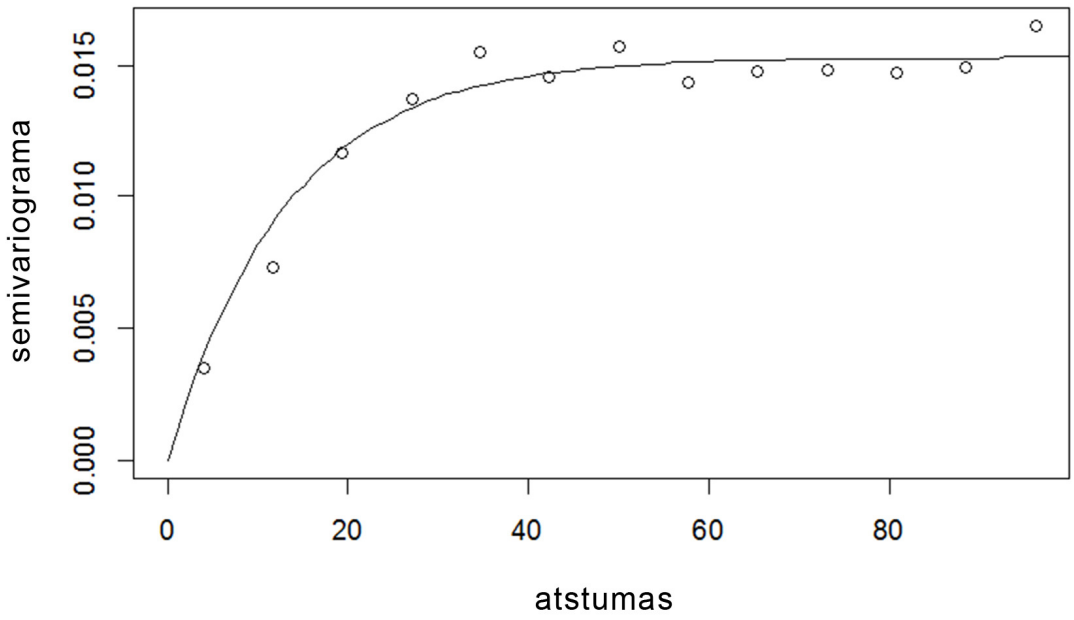
a) b) c)

Pav. 3.16 a) Paveiksliukas, natūraliai padengtas debesimis, kuris naudojamas klasifikavimui; b) miško klasės mokymo imties taškai; c) ne miško klasės mokymo imties taškai.

Kadangi nuotolinio stebėjimo paveiksliukas, naudojamas klasifikavimui, yra sugadintas natūraliai, tai tikslus koreliacijos pločio parametras α nėra žinomas. Šis parametras yra įvertinamas remiantis mokymo imties taškais, pasinaudojant R (r-project) sistemos paketu geoR.

Atliekant koreliacijos pločio parametro vertinimą, vienu metu naudojami abiejų klasių mokymo imties taškai. Tam, kad sumažinti skirtingų klasių įtaką vertinimo tikslumui, iš kiekvieno mokymo imties taško reikšmės atimamas šį tašką atitinkančios klasės vidurkis. Taip transformuoti taškai, su jų koordinatėmis, kurios atitinka paveiksliuko taškų pozicijas, paveikliuko reikšmių matricoje, toliau yra naudojami empirinės semivariogramos apskaičiavimui. Tai atliekama su geoR paketo komanda variog. Tada, prie empirinės semivariogramos taškų priglodinamas parametrinis semivariogramos modelis, panaudojant variofit komandą, kuri priglodinimui naudoja mažiausių kvadrtų

metodą. Šio eksperimento metu, buvo atlikta keletas skirtingų parametrinio modelio priglodinimo bandymų, kurių metu buvo naudojami skirtingų tipų semivariogramų modeliai. Modelių tipai buvo parinkti pagal empirinės semivariogramos formą. Šiai konkrečiai situacijai buvo nustatya, jog labiausiai tinkamas buvo eksponentinis modelis. Toks modelių parametrų įvertinimas dažnai taikomas gegrafinėse informacinėse sistemose. Pritaikyto modelio empirinė ir parametrinė semivariogramos (modelis) yra pateiktos paveikslėlyje 3.17.

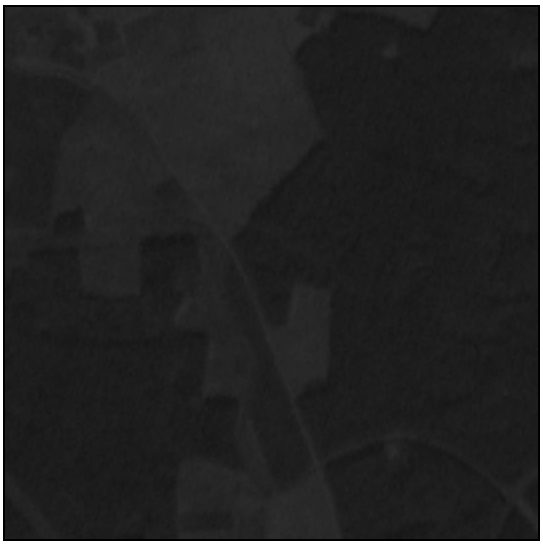


Pav. 3.17 Pagal eksperimento duomenis geriausiai tinkantis eksponentinės semivariogramos modelis.

Įvertinus koreliacijos pločio parametrą, buvo gauta, jog $\alpha = 13.0305$. Šis parametras naudojamas tolesniame klasifikavime. Klasifikavimo rezultatai pateikti skaitiškai lentelėje 3.5 ir vizualiai paveikslėlyje 3.18. Lentelėje 3.5 yra peteiktas bendras empirinis klasifikavimo tikslumas. Jis yra apskaičiuojamas lyginant gautus klasifikavimo rezutatus su tos pačios teritorijos palydoviniu vaizdu, gautu iš karto po triejų mėnesių, kai jame debesies nebėra (USGS Earth Explorer). Tos pačios teritorijos vaizdas iš Landsat 7 palydovo gali būti gautas tik po 90 parų (Gudritienė 2007).

Paveikslėlyje 3.19 pateikti paveiksliukai parodo klasifikavimo rezultatus, naudojant du skirtingus metodus, ir, naudojant 12-os artimiausių kaimynų

schemą. Nors PBDFI metodu gautas paveiksliukas (pav. 3.19 b)) yra glodesnis, tačiau PBDF metodas (pav. 3.19 a)) geriau išskiria jautresnes vietas. Geriau suklasifikuoja tokias vietas, kur klasifikuojamame paveiksliuke debesimi uždengtas miškas praktiškai susilieja su ne miško teritorija.



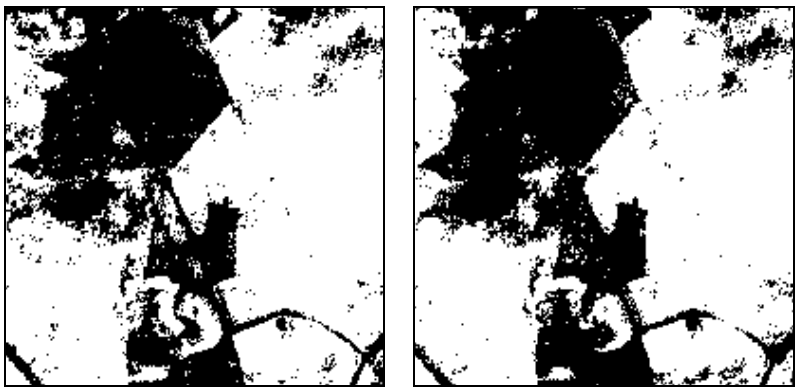
Pav. 3.18 Teritorijos, kuri buvo klasifikuojama, vaizdas po 90 parų.

Lentelė 3.5 Bendras klasifikavimo tikslumas, naudojant skirtingus klasifikavimo metodus, su skirtingomis kaimynystės schemomis.

Artimiausių kaimynų schema	Metodai	
	PBDF	PBDFI
4	0.9033	0.9020
8	0.8817	0.8747
12	0.8779	0.8623

Remiantis natūraliai sugadinto paveiksliuko klasifikavimo rezultatais, dar kartą galima teigti, jog erdvinės priklausomybės įvedimas tarp klasifikuojamo stebinio ir mokymo imties, leidžia gauti geresnį rezultatą.

Kadangi, modeliuojant realią situaciją, koreliacijos pločio parametras buvo gautas $\alpha = 13.0305$, galima teigti, jog debesims būdinga erdvinė koreliacija, todėl juos galima modeliuoti Gauso atsitiktiniu lauku.



a) b)

Pav. 3.19 Klasifikavimo rezultatai pagal PBDF (a) ir PDBFI (b) metodus, naudojant 12-os artimiausių kaimynų kaimynystės schemą.

3.5. Skyriaus išvados

- Eksperimentų metu išryškėjo PBDF metodo pranašumas prieš PBDFI metodą. Tai reiškia, jog klasifikuojant naudinga atsižvelgti į klasifikuojamų stebinių erdvinę priklausomybę su mokymo imtimi.
- Esant dideliam lauko koreliacijos pločiui α PBDF ir PBDFI metodai beveik nebereaguoja į uždėto lauko keliamą triukšmą.
- PBDF ir PBDFI metodams reikalinga mokymo imtis, tačiau šiuo atveju ji buvo naudota santykinai nedidelė lyginant su klasifikuojama teritorija (0,6 % stebinių).
- Rezultatai parodė, jog didėjant erdvinės koreliacijos pločiui duomenyse, Bajeso įterptomis diskriminantinėmis funkcijomis paremtų klasifikavimo metodų rezultatai tampa tikslesni.
- Pasiūlyta metodika gali būti naudinga klasifikuojant vaizdus, sugadintus erdvėje koreliuotu triukšmu.
- Klasifikuojant realų paveiksliuką su GLCM gauti rezultatai labai geri, bet duomenyse atsiradus erdvėje koreliuotam triukšmui šie metodai nebetenka prasmės, nes jie pradeda klasifikuoti ne paveiksliuko teritoriją, o lauko informaciją.

- Remiantis natūraliai sugadinto paveiksliuko klasifikavimo rezultatais (skyrius 3.4), dar kartą galima teigti, jog erdvinės priklausomybės įvedimas tarp klasifikuojamo stebinio ir mokymo imties, leidžia gauti geresnį rezultatą.
- Kadangi, modeliuojant realią situaciją (skyrius 3.4), koreliacijos pločio parametras buvo gautas $\alpha = 13.0305$, galima teigti, jog debesims būdinga erdvinė koreliacija, todėl juos galima modeliuoti Gauso atsitiktiniu lauku.
- Skaičiavimai atliekami žymiai greičiau PBDF ir PBDFI metodais, nei GLCM paremtu metodu.

Bendrosios išvados

- Visų eksperimentų metu PBDF metodas buvo pranašesnis už PBDFI metodą. Tai reiškia, jog klasifikuojant naudinga atsižvelgti į klasifikuojamų stebinių erdvinę priklausomybę tarp klasifikuojamo stebinio ir mokymo imties.
- Atliktų eksperimentų metu gauti rezultatai parodė, jog didėjant erdvinės koreliacijos pločiui duomenyse, Bajeso įterptomis diskriminantinėmis funkcijomis paremtų klasifikavimo metodų rezultatai tampa tikslesni. Tuo tarpu, kitų metodų, kurie buvo lyginami su PBDF ir PBDFI, klasifikavimo tikslumas negerėja.
- Ištyrus BDF klaidų tikimybių priklausomybę nuo statistinių parametrų reikšmių, gauta, jog didesnė priklausomybė tarp klasių žymių ir stipresnė erdvinė koreliacija tarp požymių stebinių, užtikrina mažesnes reikšmes. Taip pat galima teigti, kad stebiniai su stipresne erdvine priklausomybe gali būti klasifikuojami tiksliau (pagal pasiūlytą metodiką).
- Atlikus realaus nuotolinio stebėjimo vaizdo, padengto debesimis, klasifikavimą, dar kartą galima teigti, jog erdvinės priklausomybės įvedimas tarp klasifikuojamo stebinio ir mokymo imties, leidžia gauti geresnį rezultatą. Kadangi, šiuo atveju koreliacijos pločio parametras buvo gautas $\alpha = 13.0305$, galima teigti, jog debesims būdinga erdvinė koreliacija, todėl juos galima modeliuoti Gauso atsitiktiniu lauku.

Literatūra ir šaltiniai

Adan, M., Barcelo, J. A., Pijoan-Lopez, J., Toselli, A. (2003). Spatial statistics in archaeological texture analysis. Computer applications and quantitative archaeology methods.

Alder, M. (2001). An Introduction to Pattern Recognition.

Atkinson, P. M., Lewis, P. (2000). Geostatistical classification for remote sensing: an introduction. Computers and geosciences, 26, 361-371.

Atkinson, P. M. (2004). Spatially weighted supervised classification for remote sensin. International Journal of Applied Earth Observation and Geoinformation, 5, 277-291.

Atkinson, P. M., Naser, D. K. (2010). A geostatistically weighted k-NN classifier for remotely sensed imagery. Geographical analysis, 42, 204-225.

Aksoy, S. (2011). Introduction to Pattern Recognition. Department of Computer Engineering (available at http://faculty.chicagobooth.edu/hedibert.lopes/ teaching/41903-Spring2008/spatialmodels.pdf).

Barnsley, M. J., Barr, S. L. (1996) Inferring Urban Land Use from Satellite Sensor Images Using Kernel-Based Spatial Reclassification. Photogrammetric Engineering & Remote Sensing, 62(8), 949-958.

Bäse, A. M. (2004). Pattern Recognition for Medical Imaging. Elsevier Academic Press.

Besag, J. (1986). On the statistical analysis of dirty pictures. J. R. Statist. Soc. B, 48(3), 259-302.

Cressie, N. (1993). Statistics for Spatial Data revised edition. New York: Wiley

Čekanavičius, V., Murauskas, G. (2008). Statistika ir jos taikymas 2. Vilnius.

Deng, H., Clausi, D. A. (2004). Gaussian MRF Rotation Invariant Features for Image classification. IEEE Trans. on Pattern Anal. And Machine Intell., Vol. 26, No. 7, 951-955.

Diggle, P. J., Ribeiro, P. J. (2007). Model-based Geostatistics. Springer.

Dučinskas, K., Šaltytė-Benth, J. (2003). Erdvinė Statistika. Klaipėda.

Dučinskas, K. (2009). Approximation of the expected error rate in classification of the Gaussian random field observations. Statistics and Probability Letters, Nr. 79, 138-144.

Dučinskas, K., Stabingienė, L. (2011). Expected Bayes error rate in supervised classification of spatial gaussian data. Informatica. Volume 22, No. 3, 371-381. ISSN 0868-4952.

Dučinskas, K., Stabingienė, L., Stabingis, G. (2011). Image classification based on Bayes discriminant functions. In Proceedings of the 1th International Conference „Spatial Statistics 2011-Mapping Global Change". Procedia Environmental Sciences, Volume 7, selected papers, 218-223. Elsevier. ISSN 1878-0296.

Dudani, S. A. (1976). The Distance Weighted k-Nearest Neighbour Rule. IEEE Transactions on Systems, Man, and Cybernetics, 6, 325-327.

Dzemyda, G., Kurasova, O., Žilinskas, J. (2008). Daugiamačių duomenų vizualizavimo metodai. Mokslo aidai. Vilnius.

Fix, E., Hodges, J. L. (1951). Discriminatory Analysis, Nonparametric Discrimination: Consistency Properties. (available at http://www.dtic.mil/cgi-bin/GetTRDoc?AD=ADA800276).

Fukunaga, K. (1990). Introduction to statistical pattern recognition. Second ed. Academic press. New York.

Foody, G. M. (2002). Status of Land Cover classification accuracy assessment. Remote Sensing of Environment, 80, 185-201.

Gudritienė, D., Aleknavičius, A. (2007). Skaitmeninė fotogrametrija. LŽŪU.

Gupta, M., Rajaram, S., Petrovic, N., Huang, T. (2005). Restoration and recognition in a loop. Preceedings of the 2005 IEEE Computer society conference on computer vision and pattern recognition (CVPR'05), Nr. 1063-6919/05.

Gupta, M., Rajaram, S., Petrovic, N., Huang, T. (2009). Models for patch-based image restoration. EURASIP Journal on image and video processing. Nr. 10, 1155/2009/641804.

Haralick, R. M., Shanmugam, K., Dinstein, I. (1973). Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics 3(6):610-21.

Haralick, R. M. (1979). Statistical and structural approaches to texture. Proceedings of the IEEE 67, 786-804.

Hall-Beyer, M. (2007). (žiūrėta 2011 gegužė http://www.fp.ucalgary.ca/ mhallbey/tutorial.htm).

Härdle, W., Simar, L. (2003). Applied Multivariate Statistical Analysis. Method & Data Technologies.

Kettig, R. L., Landgrebe, D. A. (1976). Classification of Multispectral Image Data by Extraction and Classification of Homogeneous Objects. IEEE Transactions on Geoscience Electronics. 14, 19-26.

Li, S. Z. (2009). Markov Random Field Modelling in Image Analysis. Third ed. Springer.

Liu, J. G., Mason, P. J. (2009). Essential image processing and GIS for remote sensing. Wiley-Blackwell. UK.

Lopes, H. F. (2008). Applied Econometrics 41903-01 (available at http://faculty.chicagobooth.edu/hedibert.lopes/teaching/41903-Spring2008/spatialmodels.pdf).

Mahalanobis, P. C. (1936). On the generalized distance in statistics. Proc. Nat. Inst. Sci. 2, 49-55.

Mardia, K. V. (1984). Spatial discrimination and classification maps. Comm. Statist. Theory Methods. 13(18), 2181–2197.

Mardia, K. V. (1988). Multidimensional multivariate Gaussian Markov random fields with application to image processing. J. Multivariate Anal. 24, 265-284.

McLachlan, G. J. (2004). Discriminant Analysis and Statistical Pattern Recognition. Wiley & Sons. New York.

NASA (žiūrėta 2011 gegužė http://landsat.usgs.gov/index.php).

NASA (žiūrėta 2011 gegužė http://edcsns17.cr.usgs.gov/NewEarthExplorer).

Nishii, R. (2003). A Markov random field-based approach to decision level fusion for remote sensing image classification. IEEE Trans. Geosci. Remote Sensing 41. 41(10), 2316-2319.

Nishii, R., Eguchi, S. 2006. Image classification based on Markov random field model with Jeffreys divergence. Journal of Multivariate Analysis 97(9):1997-2008.

Raudys, S. (1976). On dimensionality, learning sample size and complexity of classification algorithms. In Proceedings of the 3rd International Conference on Pattern Recognition, pages 166-169. IEEE Computer Society, Long Beach, CA.

Raudys, S., Pikelis, V. (1980). On dimensionality, sample size, classification error, and complexity of classification algorithms in pattern recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2:242-252.

Ripley, B. D. (1996). Pattern Recognition and Neural Networks. Cambridge university press.

Schmaltz, J. (2006). NASA (žiūrėta 2011 balandis http://visibleearth.nasa.gov/view_rec.php?id=20943. NASA visible earth).

Shekhar, S., Schrater, P. R., Vatsavai, R. R., Wu, W., Chawla, S. (2002). Spatial contextual classification and prediction models for mining geospatial data. IEEE Trans. on Multimedia, 4(2), 174-188.

Stabingienė, L., Dučinskas, K. (2009). Exact error rates of the supervised classification based on Markov random fields. In Proceedings of the 13th International Conference „Applied Stochastic Models and Data Analysis“ (ASMDA-2009): selected papers, 120-123, Vilnius: Technika. ISBN 978-9955-28-463-5.

Stabingienė, L., Dučinskas, K. (2010). Error rates in spatial classification of Gaussian data with random labelling. Lietuvos matematikos rinkinys. LMD darbai, Volume 51, 426-430. ISSN 0132-2818.

Stabingienė, L., Stabingis, G., Dučinskas, K. (2010). Comparison of linear discriminant functions in image classification. Lietuvos matematikos rinkinys. LMD darbai, Volume 51, 227-231. ISSN 0132-2818.

Stabingienė, L., Stabingis, G., Dučinskas, K. (2011). Comparison of images modeled by Gaussian Random Fields. Lietuvos matematikos rinkinys. LMD darbai, Volume 52, 200-204. ISSN 0132-2818.

Switzer, P. (1980). Extensions of linear discriminant analysis for statistical classification of remotely sensed satellite imagery. Math. Geol., 12(4), 367–376.

Townshend, J. R. G. (1992). Land cover. International Journal of Remote Sensing, 13, 1319-1328.

Tso, B., Mather, P. M. (2001). Classification methods for remotely sensed data. New York: Taylor & Francis.

The R Project for Statistical Computing. (available at http://www.r-project.org/).

Webb, A. R. (2002). Statistical Pattern Recognition. Second edition. John Wiley & Sons, Ltd. England.

Winkler, G. (2006). Image Analysis, Random Fields and Markov Chain Monte Carlo Methods. Second ed. Springer.

Wu, J., Ruan, Q., An, G. (2010). Exemplar-Based Image Completion Model Employing PDE Corrections. Informatica, 21(2), 259-276.

Autoriaus publikacijos disertacijos tema

Straipsniai recenzuojamuose mokslo žurnaluose

Dučinskas, K., Stabingienė, L. (2011). Expected Bayes error rate in supervised classification of spatial gaussian data. Informatica. Volume 22, No. 3, 371-381. ISSN 0868-4952.

Stabingienė, L., Stabingis, G., Dučinskas, K. (2011). Comparison of images modeled by Gaussian Random Fields. Lietuvos matematikos rinkinys. LMD darbai, Volume 52, 200-204. ISSN 0132-2818.

Stabingienė, L., Dučinskas, K. (2010). Error rates in spatial classification of Gaussian data with random labelling. Lietuvos matematikos rinkinys. LMD darbai, Volume 51, 426-430. ISSN 0132-2818.

Stabingienė, L., Stabingis, G., Dučinskas, K. (2010). Comparison of linear discriminant functions in image classification. Lietuvos matematikos rinkinys. LMD darbai, Volume 51, 227-231. ISSN 0132-2818.

Straipsniai konferencijų pranešimų recenzuojamuose leidiniuose

Dučinskas, K., Stabingienė, L., Stabingis, G. (2011). Image classification based on Bayes discriminant functions. In Proceedings of the 1th International Conference „Spatial Statistics 2011-Mapping Global Change“. Procedia Environmental Sciences, Volume 7, selected papers, 218-223. Elsevier. ISSN 1878-0296.

Stabingienė, L., Dučinskas, K. (2009). Exact error rates of the supervised classification based on Markov random fields. In Proceedings of the 13th International Conference „Applied Stochastic Models and Data Analysis“ (ASMDA-2009): selected papers, 120-123, Vilnius: Technika. ISBN 978-9955-28-463-5.

.

LIJANA STABINGIENĖ

VAIZDŲ ANALIZĖ NAUDOJANT BAJESO DISKRIMINANTINES FUNKCIJAS

Daktaro disertacija

Fiziniai mokslai,

Informatika (09P)

LIJANA STABINGIENĖ

IMAGE ANALYSIS USING BAYES DISCRIMINANT FUNCTIONS

Doctoral Dissertation

Physical Sciences,

Informatica (09P)