VILNIAUS UNIVERSITETAS

Indrė Žliobaitė

ADAPTYVUS MOKYMO IMTIES FORMAVIMAS

Daktaro disertacijos santrauka
Fiziniai mokslai, informatika (09P)

Vilnius, 2010

Disertacija rengta 2006 - 2010 metais Vilniaus universitete bendradarbiaujant su Bangoro universiteto (Didžioji Britanija) ir Eindhoveno technologijų universiteto (Nyderlandai) mokslininkais.

Mokslinis vadovas:

prof. habil. dr. Šarūnas Raudys (Vilniaus universitetas, fiziniai mokslai, informatika - 09P).

Disertacija ginama Vilniaus universiteto Informatikos mokslo krypties taryboje:

Pirmininkas

prof. dr. Algimantas Juozapavičius (Vilniaus universitetas, fiziniai mokslai, informatika 09P),

Nariai:

doc. dr. Algirdas Bastys (Vilniaus universitetas, fiziniai mokslai, informatika - 09P),

prof. habil. dr. Henrikas Pranevičius (Kauno technologijos universitetas, fiziniai mokslai, informatika - 09P),

prof. habil. dr. Rimvydas Simutis (Kauno technologijos universitetas, technologijos mokslai, informatikos inžinerija - 07T),

dr. Julius Žilinskas (Matematikos ir informatikos institutas, fiziniai mokslai, informatika - 09P).

Oponentai:

doc. dr. Minija Tamošiūnaitė (Vytauto Didžiojo universitetas, fiziniai mokslai, informatika - 09P),

doc. dr. Pranas Vaitkus (Vilniaus universitetas, fiziniai mokslai, matematika - 01P).

Disertacija bus ginama viešame Informatikos mokslo krypties tarybos posėdyje 2010 m. balandžio mėn. 1 d. 11 val.

Adresas: Vilniaus universiteto Matematikos ir informatikos fakulteto Nuotolinių studijų centras, Šaltinių 1A, LT-03225 Vilnius.

Disertacijos santrauka išsiuntinėta 2010 m. kovo mėn. 1 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje.

VILNIUS UNIVERSITY

Indrė Žliobaitė

ADAPTIVE TRAINING SET FORMATION

Summary of doctoral dissertation
Physical sciences, informatics (09P)

Vilnius, 2010

The dissertation work was carried out at Vilnius University from 2006 to 2010 in co-operation with Bangor University (UK) and Eindhoven University of Technology (the Netherlands) researchers.

Scientific supervisor:

prof. habil. dr. Šarūnas Raudys (Vilnius University, physical sciences, informatics - 09P).

The defence council:

Chairman

prof. dr. Algimantas Juozapavičius (Vilnius University, physical sciences, informatics - 09P),

Members:

doc. dr. Algirdas Bastys (Vilnius University, physical sciences, informatics - 09P),

prof. habil. dr. Henrikas Pranevičius (Kaunas University of Technology, physical sciences, informatics - 09P),

prof. habil. dr. Rimvydas Simutis (Kaunas University of Technology, technological sciences, informatics engineering - 07T),

dr. Julius Žilinskas (Institute of Mathematics and Informatics, physical sciences, informatics - 09P).

Opponents:

doc. dr. Minija Tamošiūnaitė (Vytautas Magnus University, physical sciences, informatics - 09P),

doc. dr. Pranas Vaitkus (Vilnius university, physical sciences, mathematics - 01P).

The dissertation will be defended at the public meeting of the council on the 1st of April, 2010 at 11:00.
Adress: VU MIF Distance Learning Center, Šaltinių 1A, LT-03225 Vilnius.

The summary of the dissertation was distributed on the 1st of March, 2010.
The dissertation is available at the library of Vilnius University.

Turinys

1 Įvadas

Šiandieninėje, dinamiškai besikeičiančioje aplinkoje reikalingi adaptyvūs duomenų gavybos metodai. Nepageidaujamų elektroninių laiškų klasifikatoriai, įsilaužimų į kompiuterinius tinklus aptikimo, rinkodaros bei rekomendavimo, verslo rodiklių prognozavimo bei sprendimų priėmimo sistemos turi nuolat persimokyti, reaguoti į besikeičiančius duomenis. Stacionarioje aplinkoje kuo daugiau mokymo duomenų - tuo tikslesnis modelis. Besikeičiančioje aplinkoje seni duomenys blogina tikslumą. Tokiu atveju, vietoje visų turimų istorinių duomenų panaudojimo, gali būti tikslingai išrenkama tik tam tikra jų dalis, pvz. naudojamas mokymo langas (tik naujausi duomenys).

Disertacijoje nagrinėjama adaptyvaus mokymo besikeičioje aplinkoje mokslinė problema, siekiant pagerinti klasifikavimo bei prognozavimo tikslumą esant koncepcijos pokyčiams [1]. Darbas priskiriamas duomenų gavybos [2] mokslinei sričiai.

1.1 Tyrimų objektas

Darbo tyrimo objektas yra adaptyvūs mokymo su mokytoju [3] metodai, kurie remiasi specializuotu mokymo imties formavimu, leidžiančiu išsaugoti ir panaudoti aktualią istorinę informaciją. Pagrindiniu vertinimo kriterijumi laikome klasifikavimo bei prognozavimo tikslumą.

1.2 Darbo tikslai ir uždaviniai

Adaptyvaus mokymo besikeičiančioje aplinkoje problema nagrinėjama per trijų pokyčių tipų prizmę: staigių pokyčių, palaipsnių pokyčių bei pasikartojančių koncepcijų. **Darbo tikslas yra pagerinti klasifikavimo bei prognozavimo tikslumą besikeičiančioje aplinkoje, patobulinant adaptyvaus mokymo metodus ir sukuriant specializuotus mokymo algoritmus šiems metodams.** Tikslumo pagerinimas reiškia aplinkybių, kurioms esant siūlomi metodai veiktų tiksliau nei baziniai metodai (iš viso

[1] angl. concept drift
[2] angl. data mining
[3] angl. supervised learning

netaikant adaptyvaus mokymo) bei rinktiniai žinomi metodai iš naujausios literatūros, įdentifikavimą.

Darbo tikslui pasiekti keliami šie uždaviniai:

1 uždavinys (RQ1[4]): Ištirti: (1) nuo ko priklauso optimalus mokymo lango ilgis esant staigiam koncepcijos pokyčiui; (2) kokiais atvejais pokyčio taškas nesutampa su mokymo lango pradžia; (3) kaip šį skirtumą panaudoti adaptyvaus mokymo tikslumo pagerinimui.

2 uždavinys (RQ2): Sukurti mokymo imties parinkimo metodą sujungiant du kriterijus: panašumo laike ir požymių erdvėje. Nustatyti, kaip panašumo laike ir erdvėje kriterijų sujungimas įtakoja mokymo tikslumą esant palaipsniams pokyčiams.

3 uždavinys (RQ3): Sukurti kontekstinį mokymo metodą, kuris susietų mokymo imties parinkimą su kontekstu (istorinių duomenų „elgesio“ tipais). Metodą ištestuoti sprendžiant maisto produktų pardavimų prognozės uždavinį, kuriame vyksta dažni koncepcijų pasikartojimai.

4 uždavinys (RQ4): Praplėsti adaptyvaus mokymo teorinius samprotavimus nagrinėjant pramoninio katilo masės judėjimo prognozavimo uždavinį, ir sukuriant jo sprendimo būdą, apimantį mokymo imties patrinkimo metodiką laikui bėgant.

1.3 Tyrimų metodika

Disertacijoje koncepcijos pokyčių problema mokyme su mokytoju nagrinėjama analitiškai, naudojant matematinius modelius, paremtus daugiamate statistika, bei eksperimentiškai, naudojant generuotus bei realaus pasaulio duomenis. Tyrimo uždaviniai disertacijoje sprendžiami trimis pagrindiniais etapais. Pirmajame etape analizuojama dalykinės srities litertūra ir artimai susiję metodai, sprendžiantys tas pačias ar glaudžiai susijusias mokslines problemas. Antrajame etape įvertinamos problemos, jeigu tinkamų metodų nėra, esamų metodų trūkumai, suformuluojamos strategijos bei sprendimai trūkumams pašalinti. Sprendimai patikrinami prototipiniais eksperimentais paprastoms uždavinio formuluotėms. Trečiajame etape sukurti metodai įvertinami eksperimentiškai

[4]research question - tyrimo uždavinys

plačiam duomenų kiekiui, bei palyginami su artimais žinomais metodais bei su paprastais samprotavimais paremtais problemos sprendimais.

Didžiausia teorinė dalis sukoncentruota tyrimo uždavinyje RQ1. Tuo tarpu uždavinyje RQ4 vyrauja uždavinio ir sprendimo formulavimas bei eksperimentinių duomenų analizė. Tyrimo uždavinių RQ1-RQ4 sprendimo planas yra toks:

1. Literatūros apžvalga konkrečioje srityje, susijusioje su sprendžiamu uždaviniu: kintamo mokymo lango ilgio nustatymo metodai; naudojami mokymo imties išrinkimo kriterijai; mokymo imties formavimo strategijos tikintis pasikartojančių koncepcijų; pokyčio nustatymo mechanizmai taikomi sensoriniams duomenims.
2. Žinomų adaptyvaus mokyko metodų trūkumų nustatymas, juos taikant konkrečioje specifinėje srityje.
3. Sprendimo metodo suformulavimas ir jo intuityvus pagindimas.
4. Analitinis suformuluoto sprendimo pagrindimas, naudojant daugiamatės statistikos modelius.
5. Realių uždavinių metodų palyginimui pasirinkimas, duomenų gavimas ir paruošimas, lyginamųjų eksperimentų plano sudarymas.
6. Lyginamųjų bei sukurtojo metodų programinis realizavimas, testavimas, suplanuotų eksperimentų atlikimas bei gautų rezultatų analizė.

Rezultatams vertinti disertacijoje naudojami kiekybiniai ir kokybiniai metodai. Kiekybinio vertinimo pagrindiniu kriterijumi laikome klasifikavimo bei prognozavimo tikslumą. Generalizavimo klaida laikoma testavimo klaida, įvertinta naudojant progresinį mokymą, kuriame permokymas leidžiamas kiekviename žingsnyje, testavimui naudojami „ateities“ duomenys laike. Kokybinio įvertinimo tikslas yra identifikuoti sukurtų metodų pranašumus, trūkumus ir sąlygas, kurioms esant jie pasireiškia, didesnį dėmesį skiriant „netipinių“ bei „įdomių“ stebėjimų išsiaiškinimui.

1.4 Svarbiausi rezultatai

Koncepcijos pokyčio problemos nagrinėjimas per skirtingų pokyčių tipų prizmę yra naujas. Susistemintos mokymo imties parinkimo strategijos esant koncepcijos pokyčiams

srityje iki šiol nebuvo.

Pagrindiniai disertacijos rezultatai duomenų gavybos sričiai yra tokie. Patobulintos žinomos mokymo strategijos esant staigiems, palaipsniams ir pasikartojantiems pokyčiams. Sukurti ir eksperimentiškai aprobuoti keturi adaptyvaus mokymo imties formavimo algoritmai (WR*, FISH, CAPA, OMFP), kurie leidžia pagerinti klasifikavimo bei prognozavimo tikslumą besikeičiančiose aplinkose, esant atitinkamai kiekvienam iš trijų pokyčių tipų, lyginant su žinomais algoritmais bei pasyviomis strategijomis (naudojant visus istorinius duomenis).

1.4.1 Mokslinis naujumas

Pagrindiniai šio disertacinio darbo mokslinio naujumo aspektai yra tokie:

1. Duomenų gavybos strityje iki šiol buvo laikoma, kad nustačius staigų pokytį senų mokymo duomenų atsisakoma iš karto. Disertacijoje teoriškai atskirtas mokymo langas nuo pokyčio taško. Pademonstruota, jog skirtumo reikšmė modelio tikslumui auga didėjant duomenų sudėtingumui. Remiantis pokyčio taško ir mokymo lango teoriniu atskyrimu sukurtas ir eksperimentiškai aprobuotas naujas kintamo mokymo lango ilgio nustatymo algoritmas WR*, pademonstruotas tikslumo pagerėjimas lyginant su žinomais lango ilgio nustatymo algoritmais, neatskiriančiais pokyčio nuo mokymo lango.

2. Iki šiol srityje mokymo imčiai parinkti buvo naudojamas arba tik panašumo laike kriterijus (mokymo langai), arba tik požymių erdvėje. Vykstant palaipsnams pokyčiams aktualūs abu kriterijai. Disertacijoje sukurtas naujo tipo matas mokymo imčiai parinkti ne tik pagal laiką, bet ir *kartu* pagal panašumą požymių erdvėje. Parodyta analitiškai ir pagrįsta taikomaisiais pavyzdžiais, kad jungtinis kriterijus yra naudingas. Tuo pagrindu sukurtas ir eksperimentiškai aprobuotas mokymo imties parinkimo algoritmas FISH, pademonstruotas tikslumo pagerėjimas lyginant su žinomais mokymo imties parinkimo algoritmais, naudojančiais tik laiko ar tik erdvės kriterijus.

3. Remiantis pardavimų kiekio prognozavimo uždaviniu, kuriam aktualūs koncepcijų pasikartojimai, sukurtas bei eksperimentiškai aprobuotas kontekstinis mokymo

imties formavimo metodas CAPA. CAPA identifikuoja objekto tipą ir pagal jį interaktyviai formuoja požymių erdvę bei parenka mokymo vektorius. Objekto tipo identifikavimas remiantis suformuluotais struktūriniais požymiais siekiant atitinkamai tipui suformuoti mokymo imtį yra moksliškai naujas.

1.4.2 Praktinis reikšmingumas

Disertacijoje sukurti metodai ištestuoti naudojant realius duomenis iš įvairių dalykinių sričių bei sprendžiant du pramoninius uždavinius: maisto produktų kiekio prognozavimo bei šildymo katilo masės kitimo įvertinimo.

CAPA metodas, sukurtas remiantis maisto produktų prognozavimo uždaviniu, gali būti taikomas ir kitiems kitiems prognozavimo uždaviniams, pvz. įvairių pardavimų, paklausos prognozavimui, nusikaltimų prognozavimui geografiškai, autobusų maršruto įveikimo laiko prognozavimui.

Pramoninio katilo uždaviniui sukurtas naujas masės judėjimo įvertinimo algoritmas OMFP, įvertinantis koncepcijos pokyčius, bei formuojantis mokymo imtį lango principu priklausomai nuo nustatyto pokyčio, reikalingas katilo kontrolės sistemai. OMFP gali būti adaptuotas įvairiems degimo ar kuro sunaudojimo uždaviniams, pvz. kuro sunaudojimo sekimas automobilyje priklausomai nuo eismo sąlygų, kuro tipo.

Sukurtieji algoritmai WR* ir FISH gali būti taikomi įvairiems klasifikavimo uždaviniams, kuriuose tikimasi atitinkamai staigių ar palaipsnių pokyčių laikui bėgant. Staigūs pokyčiai ypač aktualūs kompiuterinių tinklų įsilaužimų aptikimo, finansinių nusikaltimų prevencijos, navigacijos, paklausos pokyčių uždaviniams. Palaipsniai pokyčiai ypač aktualūs rinkodaros, rekomendavimo atsižvelgoiant į tikėtinus asmens interesus (pvz. filmų, knygų) uždaviniuose, adaptyvioms edukacinėms sistemoms, elektroninėms parduotuvėms.

Disertacinis darbas prisideda prie koncepcijos pokyčio problemos sprendimo duomenų gavyboje, sukuriami nauji aktualūs adaptyvaus mokymo imties formavimo metodai, metodai pritaikomi praktinių uždavinių sprendime.

1.5 Aprobavimas

Doktorantės rezultatai disertacijos tema publikuoti 11 mokslinių straipsnių, 2 išplėstinėse santraukose. Iš 11 straipsnių 7 yra periodiniai recenzuojami leidiniai (ISSN kodai) ir 4 yra neperiodiniai. Iš 11 straipsnių 3 patenka į ISI[5], 4 į ISI proceedings, 2 į Lietuvos mokslo tarybos patvirtintą tarptautinių duomenų bazių sąrašą. Vienas konferencijos straipsnis yra gavęs geriausio straipsnio apdovanojimą. Straipsnių sąrašas Vilniaus universiteto institucojos vardu pateikiamas 3 skyriuje, pilną doktorantės publikacijų sąrašą disertacijos tema galima rasti disertacijoje.

Autorė dalyvavo ir pristatė rezultatus šešiose tarptautinėse mokslinėse konferencijose: ICAISC 2006 (The 8^th International Conference on Artificial Intelligence and Soft Computing), MLDM 2007 (The 5^th International Conference on Machine Learning and Data Mining in Pattern Recognition), IWAPR 2007 (International Workshop on Advances in Pattern Recognition), FSKD 2008 (The 7^th International Conference on Fuzzy Systems and Knowledge Discovery), BNAIC 2009 (The 21^st Benelux Conference on Artificial Intelligence), ICMD 2009 (IEEE International Conference on Data Mining: the 1^st International Workshop on Transfer Mining (TM 2009) ir the 3^rd International Workshop on Domain Driven Data Mining (DDDM 2009)).

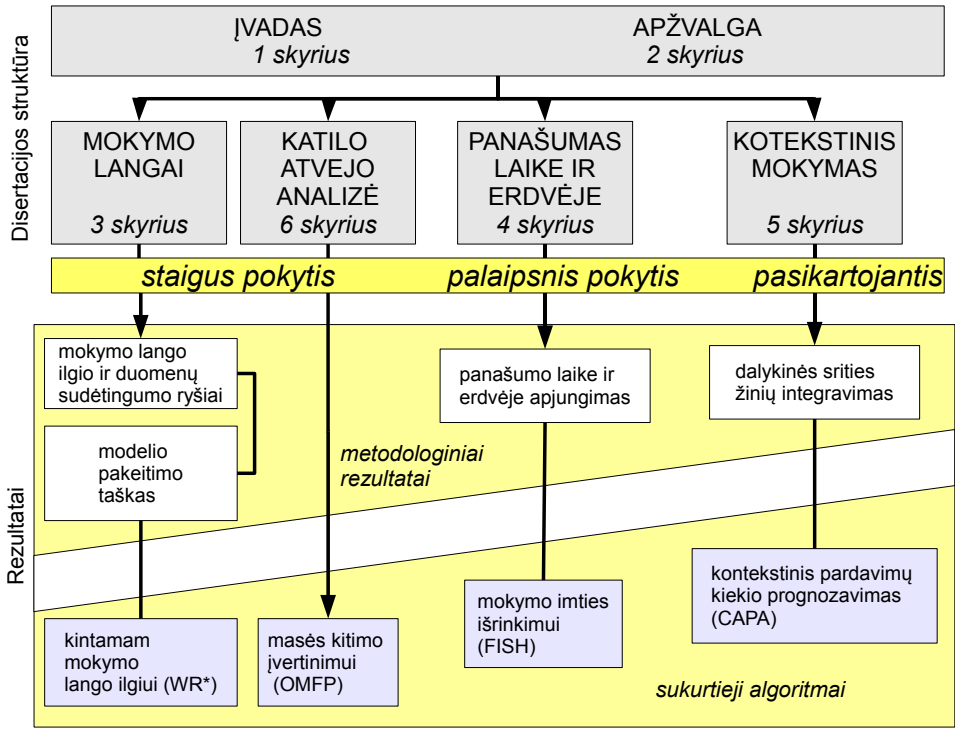
Disertacijos rezultatai taip pat pristatyti moksliniuose pranešimuose Bangoro universitete (Didžioji Britanija), Helsinkio Technologijų universitete (Suomija), Eindhoveno Technologijų universitete (Olandija), Vytauto Didžiojo universitete, Matematikos ir informatikos institute, Vilniaus universitete.

1.6 Disertacijos struktūra

Disertacinis darbas susideda iš septynių skyrių: įžangos, tyrimų srities aprašymo, keturių skyrių skirtų rezultatų pristatymui, išvadų, cituotos literatūros sąrašo bei keturių priedų.

Pirmame disertacijos skyriuje pristatoma koncepcijos pokyčių problema bei jos aktualumas, suformuluojamas darbo tikslas ir uždaviniai. Antrame skyriuje apžvelgiami susiję mokslo darbai, susisteminami žinomi algoritmai bei kategorizuojami taikomieji uždaviniai. Trečiame skyriuje pateikiami disertacijos rezultatai staigiam koncepcijos

[5] 2005 ir 2006 m. LNCS ir LNAI priklausė ISI.



1 pav.: Disertacinio darbo struktūra ir rezultatai.

pokyčiui, sprendžiamas kintamo mokymo lango ilgio nustatymo uždavinys, atskirai vertinant pokyčio tašką ir mokymo lango ilgį. Ketvirtame skyriuje pateikiami rezultatai palaipsniam pokyčių tipui, sprendžiant mokymo imties išrinkimo uždavinį, sujungiant panašumo laike ir erdvėje sąvokas. Penktame skyriuje pateikiami rezultatai pasikartojančioms koncepcijoms, sprendžiamas pardavimų kiekio prognozavimo uždavinys, sukuriamas kontekstinis mokymo imties formavimo metodas. Šeštame skyriuje pateikiami pramoninio katilo atvejo analizės rezultatai, sprendžiamas masės pokyčio įvertinimo uždavinys, sukuriamas algoritmas masės kiekio pokyčiui įvertinti kintamomis sąlygomis. Septintame skyriuje pristatomos tolesnių tyrimų kryptys ir pateikiamos disertacijos išvados.

Disertacijos struktūra ir rezultatai grafiškai pavaizduoti 1 pav.

1.7 Ginamieji teiginiai

1. Mokymo lango ir pokyčio taško teorinis atskyrimas parenkant kintamą mokymo lango ilgį leidžia pagerinti klasifikavimo tikslumą esant staigiam koncepcijos pokyčiui (3 disertacijos skyrius).

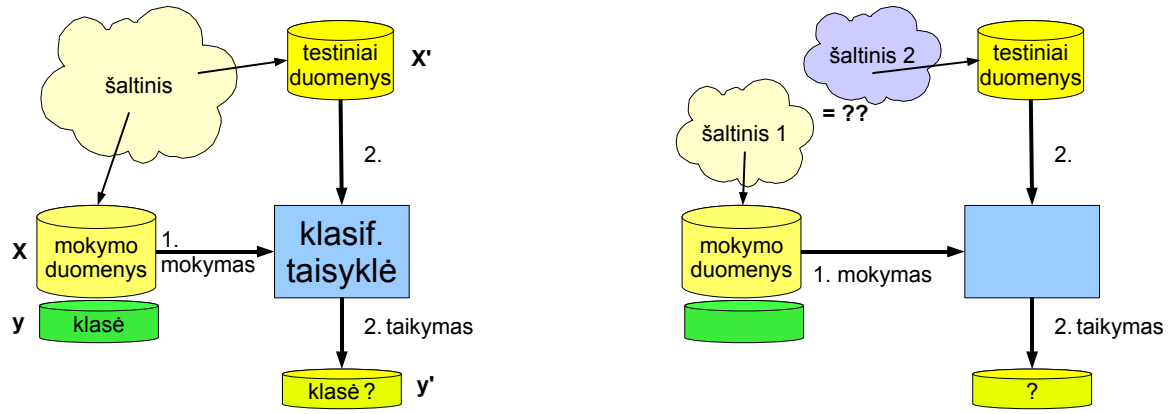
2. Panašumo požymių erdvėje įtraukimas kartu su laiku į mokymo imties parinkimo procedūrą leidžia pagerinti klasifikavimo tikslumą, lyginant su atskirų panašumo kriterijų naudojimu, esant palaipsniams koncepcijos pokyčiams (4 disertacijos skyrius).

3. Konteksto panaudojimas mokymo imčiai parinkti, surišant istorinių pardavimų tipus su mokymo imties formavimo strategijomis, bei išmokstant atskirti tipus vykdymo metu naudojant struktūrinius požymius, leidžia pagerinti prognozavimo tikslumą sprendžiant maisto produktų pardavimo kiekio prognozavimo uždavinį, kur tikimasi koncepcijų pasikartojimo bei pokyčių (5 disertacijos skyrius).

4. Sukurtasis adaptyvus masės kitimo įvertintinimo metodas pramoniniam katilui, veikiančiam kintamomis kuro tipų ir kuro padavimo sąlygomis, leidžia pasiekti tikslesnius įverčius, nei nenaudojant adaptyvumo pokyčiams, ir tuo būdu patobulinti katilo kontrolės sistemą (6 disertacijos skyrius).

2 Rezultatai

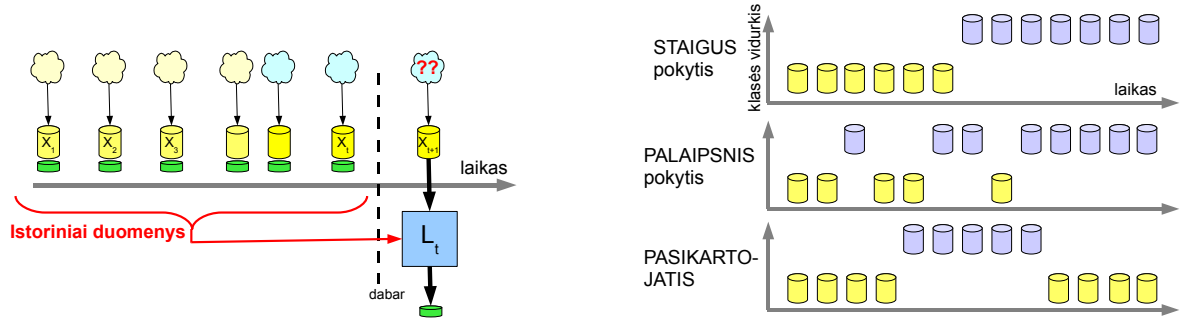
Mokymo su mokytoju [6] tikslas yra iš mokymo duomenų išmokti klasifikavimo (ar prognozavimo) taisyklę $\mathcal{L}$ tam, kad po to ją būtų galima taikyti nematytiems testavimo duomenims. Mokymo duomenys susideda iš objektų porų: vektorių $\mathbf{X} \in \Re^p$ p-matėje požymių erdvėje bei klasių numerių $\mathbf{y}$, kur $\mathbf{y} \in \mathcal{Z}^1$ klasifikavimo uždaviniams, $\mathbf{y} \in \Re^1$ - prognozavimo uždaviniams. Pagrindinis tikslas yra turint nematytus *testavimo vektorius* nustatyti jų klasės numerius, žr. 2(a) pav. Tam naudojama taisyklė $\mathcal{L}$ ($\mathbf{y} = \mathcal{L}_t(\mathbf{X})$), kurios parametrai buvo fiksuoti naudojant mokymo duomenis. Kaip taisyklė $\mathcal{L}$ gali būti naudojami įvarūs *baziniai klasifikatoriai*, pvz. Euklidinis klasifikatorius, sprendimų medis, atraminių vektorių klasifikatorius [5].



2 pav.: Mokymas su mokytoju: (a) stacionarus, (b) koncepcijos pokyčiai.

Duomenų šaltinis tai skirstinių $p(\mathbf{X}|c_i)$ ir klasių apriorinių tikimybių $P(c_i)$ rinkinys visoms klasėms $\mathbf{y} = c_1, \ldots, c_k$ duotame uždavinyje. Paprastai daroma prielaida, kad mokymo ir testavimo duomenų šaltinis yra tas pats (stacionarūs duomenys). Tačiau tam tikruose uždaviniuose (pavyzdžiai minėti 1 Skyriuje) šaltinis gali keistis laikui bėgant. Gali keistis visų arba dalies klasių apriorinės tikimybės $P(c_i)$, skirstiniai $p(\mathbf{X}|c_i)$ ir iš to sekančios posteriorinės klasių tikimybės $p(c_i|\mathbf{X})$, pagal kurias, naudojant Bajeso sprendimų teoriją [5], yra priimami klasifikavimo sprendimai. Nenumatytus duomenų šaltinių pasikeitimus laikui bėgant vadinsime *koncepcijos pokyčiais*. Nenumatyti reiškia, kad pokyčių galima tikėtis remiantis dalykinės srities žiniomis, tačiau nėra tiksliai žinoma, *kada* ir *ar tikrai* pokyčiai įvyks. Jei šaltinis keičiasi laikui bėgant, reikalingi modeliai, kurie sugebėtų adaptuotis prie koncepcijos pokyčių, žr. ilustraciją 2 (b) pav.

[6] angl. supervised learning



3 pav.: (a) Progresinis mokymas laiko momentu t. (b) Pagrindiniai pokyčių tipai.

Mokymas esant koncepcijos pokyčiams yra susijęs su klasifikavimo sprendimų priėmimu laike (paeiliui). Tai reiškia, kad laiko momentu t mokymo duomenys $(\mathbf{X}_1, \ldots, \mathbf{X}_t)$ ir testavimo duomenyes (ar duomuo) $\mathbf{X}_{t+1}$ yra išrikiuoti pagal laiką ir turi laiko atributus. Laiko momentu t reikia nustatyti testavimo duomenų (duomens) klasės numerius(-į) $\mathbf{y}_{t+1}$ Taigi, sprendimo priėmimo taisyklė $\mathcal{L}_\sqcup$ taip pat yra išmokstama konkrečiam laikui, naudojant mokymo duomenis - vektorius $(\mathbf{X}_1, \ldots, \mathbf{X}_t)$ ir jų klasių reikšmes $(\mathbf{y}_1, \ldots, \mathbf{y}_t)$. Mokymas klasifikavimo sprendimui priimti laiku t yra iliustruotas 3 (a) pav.

Kai išmokomas klasifikatorius $\mathcal{L}_\sqcup$ ir naudojnt jį priimamas klasifikavimo sprendimas bei vėliau sužinoma $\mathbf{y}_{t+1}$ tikroji reikšmė reikšmė, galima $\mathbf{X}_{t+1}$ priskirti prie mokymo duomenų sekančiam klasifikavimo žingsniui. Tai reiškia, kad laiko momentu t+1 mokymo duomenys yra $(\mathbf{X}_1, \ldots, \mathbf{X}_t, \mathbf{X}_{t+1})$ su $(\mathbf{y}_1, \ldots, \mathbf{y}_t, \mathbf{y}_{t+1})$, testavimo duomenys (duomuo) yra $\mathbf{X}_{t+2}$, reikia išmokyti klasifikatorių $\mathcal{L}_{t+1}$ tam, kad nustatyti klasės reikšmę $\mathbf{y}_{t+2}$. Tokią mokymo schemą (permokoma laikui bėgant) vadiname *progresiniu mokymu.*

Kai uždavinys stacionarus, bendruoju atveju kuo daugiau mokymo duomenų tuo tikslesnį klasifikatorių $\mathcal{L}_t$ galime išmokyti, todėl tikslinga naudoti visus istorinius duomenis $(\mathbf{X}_1, \ldots, \mathbf{X}_t)$. *Esant koncepcijos pokyčiui gali būti tikslinga naudoti ne visus istorinius duomenis, bet kryptingai išrinktą ir suformuotą jų dalį.* Kryptingas formavimas reiškia metodus, kurie įvertina kiek, kaip ir kokių istorinių duomenų tikslinga įtraukti į mokymo imtį, kad pasiekti mokymo adaptyvumą laikui bėgant. Tokie metodai ir yra šios disertacijos tyrimo objektas. Pagrindiniu metodo vertinimo kriterijumi laikome klasifikavimo tikslumą.

Disertacijoje mokymo imties formavimo metodus nagrinėjame pagal tai, kokio tipo koncepcijos pokyčių tikimasi uždavinyje. Išskiriame tris pagrindinius pokyčių tipus: staigus,

palaipsnis ir pasikartojantis, žr. 3 (b) pav.

Toliau apžvelgiame pagrindinius disertacijos rezultatus. Rezultatai pristatomi keturiuose skyriuose atitinkamai staigiems, palaipsniams bei pasikartojantiems pokyčiams, bei pramoninio katilo uždaviniui, jungiančiam kelis tipus.

2.1 Staigūs pokyčiai: mokymo lango nustatymas

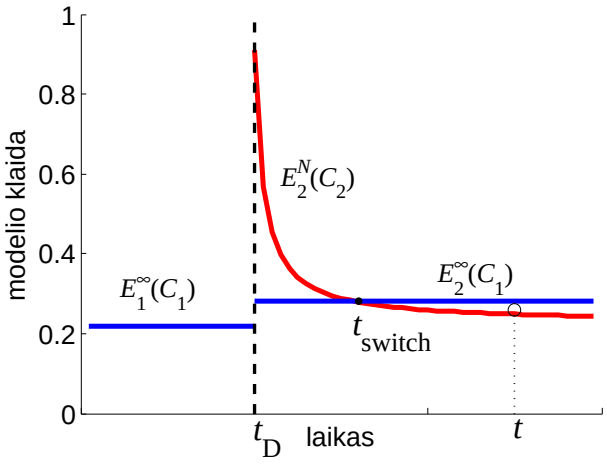
Esant staigiam koncepcijos pokyčiui vienas duomenų generavimo šaltinis S_1 staigiai pasikeičia į kitą S_2. Nėra tiksliai žinoma, kuriuo laiko momentu įvyko pokytis. Pokytį galima nustatyti naudojant pokyčio nustatymo metodus (apžvalga [3]). Staigaus pokyčio pavyzdžiai: kontrakto sąlygų pasikeitimas maisto produktų pardavimų prognozavimo uždavinyje, įsigytas automobilis tiesioginės rinkodaros uždavinyje, pasikeitęs kreditinės kortelės turėtojo atsiskaitymų aktyvumas finansinių nusikaltimų aptikimo uždavinyje.

Tokiu atveju naudojama mokymo imties parinkimo strategija, vadinama *mokymo langu*, t.y. į mokymo imtį laiko momentu t įtraukiami naujausi N duomenų $(\mathbf{X}_{t-N+1}, \ldots, \mathbf{X}_t)$ su $(\mathbf{y}_{t-N+1}, \ldots, \mathbf{y}_t)$. Pagrindinė mokslinė problema tokiu atveju, kokio ilgio turėtų būti langas, t.y. N, nuo ko priklauso lango ilgis ir kaip jį nustatyti laiko momentu t.

Disertacijos 3 skyriuje teoriškai ir eksperimentiškai nagrinėjamas mokymo lango ir klasifikavimo klaidos ryšys parametriniams modeliams. Pokyčio taškas atskirtas nuo mokymo lango sąvokos, tai ir iš to sekantis metodas bei algoritmas yra pagrindinis 3 skyriaus mokslinis naujumas. Sukurtas teorinis metodas nustatyti, kuriuo laiko momentu tikslinga pakeisti seną klasifikavimo modelį į išmokytą naudojant naujus duomenis. Metodo pagrindu sukurtas algoritmas WR* kintamam mokymo lango ilgiui N nustatyti.

2.1.1 Modelio pakeitimo taškas

Tarkime laiko momentu t_D įvyko staigus koncepcijos pokytis, dabar yra laiko momentas t. Pradžiai tarkime, kad t_D yra žinomas, bet nežinomi duomenų šaltiniai (skirstiniai) atitinkamai S_1 iki ir S_2 po pokyčio. Tarkime C_1 yra klasifikatorius išmokytas naudojant istorinius duomenis $(\mathbf{X}_1, \ldots, \mathbf{X}_{tD-1})$ su $(\mathbf{y}_1, \ldots, \mathbf{y}_{tD-1})$, kurie yra iš šaltinio S_1, C_2 yra klasifikatorius išmokytas naudojant naujo šaltinio istorinius duomenis $(\mathbf{X}_{tD}, \ldots, \mathbf{X}_t)$ su $(\mathbf{y}_{tD}, \ldots, \mathbf{y}_t)$. Kadangi nauji duomenys ateina laikui bėgant, iš karto po pokyčio laiku t_D



4 pav.: *Modelio perjungimo taškas.*

naujų duomenų bus mažai ir klasifikatorius C_2 išmokytas naudojant tik naujus duomenis gali būti netikslus.

Disertacijoje teoriškai nustatėme kiek laiko po pokyčio dar tikslinga naudoti seną modelį, priklausomai nuo duomenų parametrų (panaudojome Fišerio bazinį klasifikatorių [5] dviems Gausinių duomenų klasėms su vienoda apriorine tikimybe). Gavome *modelio perjungimo taško išraišką* t_{switch} ($t_{switch} > t_D$). Problema grafiškai pavaizduota 4 pav.

Fukunaga ir Hayes [6] parodė, kad bet kokiam parametriniam klasifikatoriui C klasifikavimo klaida išreiškiama:

$$E^N(C) \approx E(C) + \frac{1}{N} f(C), \tag{1}$$

kur N yra mokymo imties dydis, $E(C) = \lim_{N\to\infty} E^N(C)$ yra asimptotinė klasifikatoriaus C klaida, $E^N(C)$ yra klasifikatoriaus išmokyto naudojant N duomenų klaida, $f(C)$ yra funkcija, kuri priklauso nuo klasifikatoriaus tipo, duomenų skirstinių, bet nepriklauso nuo N.

Pasinaudodami šia išraiška randame *modelio perjungimo tašką* $t_{switch} = t_D + N^*$, kuris iliustruotas 4 pav.

$$N^* = \frac{f(C_2)}{E_2(C_1) - E_2(C_2)}. \tag{2}$$

Rasti $f(C)$ skirtingiems klasifikatoriams galima pasinaudojant metodika nurodyta [6, 13]. Klaidos $E_i(C_i)$ išraiškas skirtingiems klasifikatorių tipams galima rasti [13]. Disertacijoje išvesta *perjungimo taško* nustatymui reikalinga išraiška $E_i(C_j)$ kai $i \neq j$ Fišerio

įvesties duomenys: istorinių duomenų seka su žinomomis klasių reikšmėmis.

1. Istorinės pokyčio tikimybės $P(\text{change}|j)$ kai $j = 1, \ldots, t$.
2. Nustatomas pokyčio taškas $t_D = \arg\max_{j=1}^{t} P(\text{change}|j)$.
3. Naudojant (3) randamas $N^{WR^*} = N(t_D)$.

rezultatas: mokymo lango ilgis N^{WR^*}.

5 pav.: Mokymo lango nustatymo algoritmas (WR*)

klasifikatoriui, taip pat parametrų įverčių korekcijos.

2.1.2 Kintamojo lango ilgio nustatymo algoritmas WR*

Naudodami formulę (2) ir žinodami t_D, mokymo lango ilgį laiko momentu t nustatome tokiu būdu:

$$N(t) = \begin{cases} t, & \text{if } t < t_{\text{switch}}, \\ t - t_D + 1, & \text{if } t \geq t_{\text{switch}}. \end{cases} \tag{3}$$

Pokyčio taškui t_D nustatyti disertacijoje sukurtas metodas, kuris remiasi Hotelling T^2 testu [8], tačiau galima naudoti ir kitą pokyčio nustatymo metodą. Tuomet pagal išraišką (3) mokymo langas N^{WR^*} nustatomas:

$$t_D = \arg\max_{j=1}^{t} P(\text{change}|j) \tag{4}$$

$$N^{WR^*} = N^*(t_D). \tag{5}$$

WR* algoritmo žingsniai iliustruoti 5 pav.

2.1.3 Eksperimentai ir rezultatai

3 skyriuje atlikti eksperimentai turėjo tris pagrindinius tikslus.

- Nustatyti kaip mokymo lango ilgis priklauso nuo uždavinio sudėtingumo ir nuo pokyčio stiprumo.

- Ištirti kokią įtaką pokyčio taško ir lango ilgio atskyrimas daro klasifikavimo tikslumui galiojant ir negaliojant modelio prielaidoms.
- Palyginti WR* algoritmo tikslumą su kitais kintamo lango nustatymo algoritmais iš naujausios literatūros.

Iš viso skyriaus eksperimentams naudojome 8 generuotus ir 10 realių duomenų. WR* tikslumo tyrimui naudojome 3 generuoti duomenys ir 10 realių, charakteristikos pateikiamos 1 lentelėje.

1 lentelė: WR* tyrimui naudoti duomenys.

	Dimensijos	**Kiekis**	**Klasių proporcijos**	**Duomenų tipas**
Stagger	9	120	kinta	generuoti
Gaus2	7	400	0.5:0.5	generuoti
Hyper2	2	250	0.5:0.5	generuoti
WRaustralian	14	690	0.56:0.44	realūs
WRbreast	30	596	0.64:0.36	realūs
WRcylinder	36	540	0.58:0.42	realūs
WRgerman (num)	24	1000	0.70:0.30	realūs
WRhepatitis	19	155	0.79:0.21	realūs
WRionosphere	34	351	0.64:0.36	realūs
WRStatlog heart	13	270	0.56:0.44	realūs
WRSPECT heart	22	267	0.79:0.21	realūs
WRsonar	60	208	0.53:0.47	realūs
WRvote	16	435	0.61:0.39	realūs

WR* tikslumas palygintas su su trim žinomais algoritmais naudojančiais mokymo langą KLI [11], BIF [2], GAM [7] ir baziniu mokymo algoritmu be adaptyvumo, naudojančiu visą mokymo istorija ALL. Palyginimui taip pat naudotas WR algoritmas, kuris irgi yra disertacinio darbo rezultatas, jis nenaudoja modelio perjungimo taško, tačiau naudoja disertacijoje sukurtą pokyčio nustatymo metodą. Mokymo lango strategijų nustatymo palyginamumo tikslais Fišerio klasifikatorius naudotas kaip bazinis modelis visiems algoritmams.

WR* algoritmo tikslumo palyginimo rezultatai su tikrais duomenimis pateikiami 2 lentelėje. Algoritmų rezultatai palyginti skaičiuojant vidutinį reitingą. Konkretiems duomenims

2 lentelė: Palyginamųjų algoritmų testavimo klaidos.

Duomenys	**WR***	**KLI**	**WR**	**ALL**	**BIF**	**GAM**
WRaustralian	35.34	34.33	35.92	35.34	35.34	35.34
WRbreast	11.71	11.88	11.71	11.71	11.71	11.71
WRcylinder	44.62	47.22	48.89	44.62	48.89	44.62
WRgerman	38.29	37.89	38.29	38.59	38.59	38.59
WRStatlog heart	38.48	39.22	38.48	39.22	38.85	39.22
WRSPECT heart	26.50	25.00	29.14	25.75	25.75	25.75
WRhepatitis	42.53	36.69	42.53	43.18	43.18	43.18
WRionosphere	25.86	25.00	26.14	27.57	27.57	28.43
WRsonar	37.92	41.30	37.92	37.92	37.92	37.92
WRvote	11.18	12.10	11.64	11.87	11.87	11.87
reitingas	2.60	3.20	3.50	3.80	3.95	3.95
skirtumas statistiškai reikšmingas	$\alpha =$	20%	10%	4%	3%	3%

geriausią tikslumą pasiekęs algoritmas gauna reitingą 1, blogiausią - 6. Statistinis rezultatų reikšmingumas vertintas naudojant Bonferroni-Dunn testą [4].

Gautas tikslumo klasifikavimo pagerėjimas naudojant WR* algoritmą, kuris naudoja pokyčio taško ir mokymo lango atskyrimo strategiją.

2.1.4 Pagrindiniai skyriaus rezultatai

Teoriškai atskirtas mokymo langas nuo pokyčio taško. Parodyta kaip ir kodėl šios sąvokos skiriasi, pademonstruota, jog skirtumo reikšmė modelio tikslumui auga didėjant duomenų sudėtingumui. Teoriškai nustatytas ryšys tarp pokyčio, mokymo lango ir duomenų sudėtingumo parametriniams modeliams.

Remiantis pokyčio taško ir mokymo lango teoriniu atskyrimu sukurtas ir eksperimentiškai aprobuotas naujas kintamo mokymo lango ilgio nustatymo algoritmas WR*, kuris nustato lango ilgį remiantis teoriniais generalizavimo klaidos įverčiais. Pademonstruotas tikslumo pagerėjimas lyginant su žinomais lango ilgio nustatymo algoritmais, neatskiriančiais pokyčio nuo mokymo lango.

Mokymo lango ir pokyčio taško teorinis atskyrimas parenkant kintamą mokymo lango ilgį leidžia pagerinti klasifikavimo tikslumǩ esant staigiam koncepcijos pokyčiui.

2.2 Palaipsniai pokyčiai: panašumo laike ir erdvėje sujungimas

Esant palaipsniams pokyčiams vienu laiko momentu t gali būti aktyvus daugiau nei vienas duomenų šaltinis. Tarkime, iki laiko momento t_1 duomenis generuoja šaltinis S_I. Nuo laiko momento $t_2 + 1$ duomenis generuoja S_{II}, kuris pilnai pakeičia prieš tai buvusį šaltinį. Laiko tarpu $(t_1 + 1, t_2)$ abu šaltiniai yra aktyvūs ir duomenys gali būti generuoti tiek vieno, tiek kito su tam tikra tikimybe. Tikimybė, kad duomenis generuos šaltinis S_{II} laikui bėgant didėja. Pavyzdžiui, naujienų rekomendavimo sistemose vartotojas gali pradžioje domėtis mėsos kainomis, tačiau nekilnojamojo turto interesas didėja laikui bėgant ir ilgainiui tampa pagrindiniu. Rekomendavimo sistema turi klasifikuoti duotą straipsnį kaip įdomų arba neįdomų. Įdomumas keičiasi laikui bėgant palaipsniui vis mažiau grįžtant į seną temą.

Palaipsnių pokyčių atvejais siekiant modelio adaptyvumo neužtenka vien mokymo lango strategijos. Tikslinga naudoti mokymo imties išrinkimą iš istorinių duomenų, t.y. mokymo imtis gali būti sudaroma imant mokymo vektorius nebūtinai iš eilės laike. Tikslinga naudoti ir duomenų panašumo erdvėje kriterijų. Pagrindinį tikslą formuluojame: parinkti mokymo imtį taip, kad ji kuo tiksliau atitiktų duomenų $\mathbf{X}_{t+1}$ šaltinį, t.y. parinkti kuo panašesnius mokymo duomenis į $\mathbf{X}_{t+1}$ laiko ir erdvės aspektais.

Srityje žinomi adaptyvūs metodai mokymo imtį parinkdavo arba laike (langai) arba erdvėje. Disertacijos 4 skyriuje sujungti abu kriterijai mokymo imčiai sudaryti. To pagrindu sukurtas mokymo imties parinkimo algoritmas FISH, kuris gali išmokti parametrus vykdymo eigoje (panašumo proporcijas ir mokymo imties dydį).

2.2.1 Panašumo laike ir erdvėje sujungimas mokymo imčiai parinkti

Progresiniu mokymo scenariju, kuris pristatytas skyriaus pradžioje, turimi mokymo duomenys $(\mathbf{X}_1, \ldots, \mathbf{X}_t)$ su $(\mathbf{y}_1, \ldots, \mathbf{y}_t)$ ir gaunamas vektorius $\mathbf{X}_{t+1}$, kuriam reikia nustatyti klasę $\mathbf{y}_{t+1}$, naudojant mokymo taisyklę $\mathcal{L}_t$. Nors klasės reikšmė $\mathbf{y}_{t+1}$ yra nežinoma, tačiau vektorius $\mathbf{X}_{t+1}$ yra duotas ir galima lyginti jo panašumą su istoriniais mokymo

vektoriais $(\mathbf{X}_1, \ldots, \mathbf{X}_t)$ požymių erdvėje, pavyzdžiui naudojant Euklidinį atstumą. Daugiau atstumo erdvėje matų yra apžvelgta [1, 10].

Disertacijoje apibrėžiame jungtinį panašumą laike ir erdvėje tarp $\mathbf{X}_j$ ir $\mathbf{X}_i$ per atstumo funkciją

$$\mathcal{D}(\mathbf{X}_i, \mathbf{X}_j) = f(d_{ij}^{(S)}, d_{ij}^{(\mathcal{T})}), \tag{6}$$

čia $d_{ij}^{(S)}$ yra atstumas tarp vektorių $\mathbf{X}_i$ ir $\mathbf{X}_j$ erdvėje, $d_{ij}^{(\mathcal{T})}$ yra atstumas laike. Kuo mažesnis atstumas, tuo vektoriai panašesni.

Panašumą tarp $\mathbf{X}_i$ ir $\mathbf{X}_j$ laike šiame darbe apibrėžiame:

$$d_{ij}^{(\mathcal{T})} = f(|i - j|). \tag{7}$$

Tačiau reikiant galima naudoti ir sudėtingesnę (netiesinę) laiko funkciją.

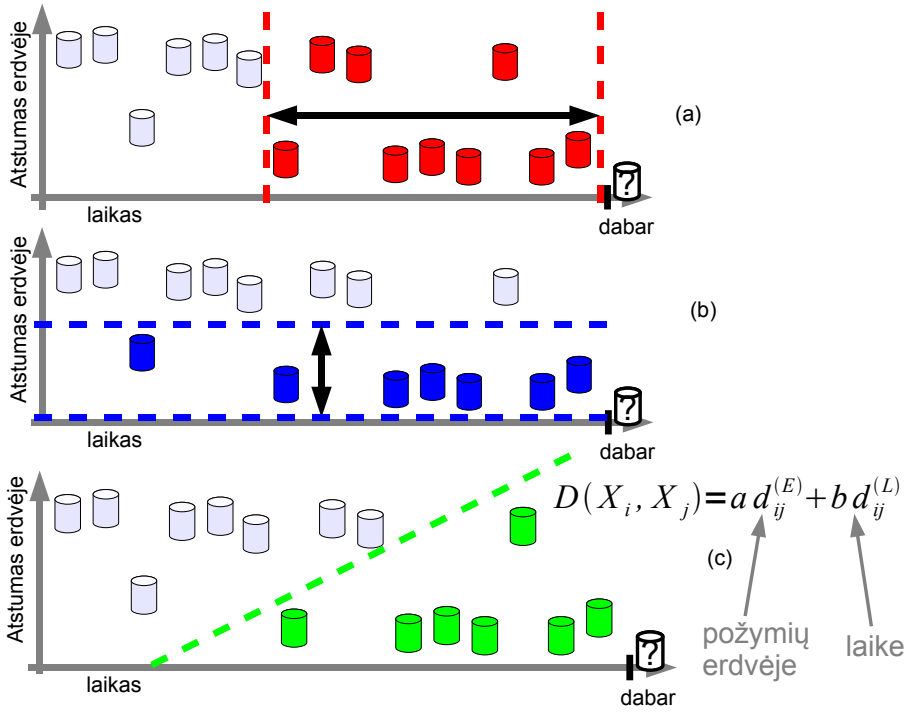
Suformuluotas panašumo laike ir erdvėje kriterijus apima ir srityje naudojamas strategijas (kurias čia vadiname ribinėmis): mokymo imties parinkimą tik laike (langas) bei tik požymių erdvėje, žr. iliustracijas 6 pav. (a) ir (b) . To paties pav. (c) iliustruotas sukurtasis kriterijus sujungiantis abu panašumus.

Šiame darbe apsiribojama tiesine panašumo laike ir erdvėje sujungimo funkcija $\mathcal{D}$, tačiau priklausomai nuo dalykinės srities ir uždavinio sudėtingumo, galima būtų panaudoti ir kitokias funkcijas. Tiesinė kombinacija atrodo taip:

$$\mathcal{D}(\mathbf{X}_i, \mathbf{X}_j) = \alpha_1 d_{ij}^{(S)} + \alpha_2 d_{ij}^{(\mathcal{T})}, \tag{8}$$

čia α_1 ir α_2 yra panašumo proporcijos. Jei $\alpha_1 = 0$, gauname mokymo langą (6(a) pav). Jei $\alpha_2 = 0$, gauname panašumą tik požymių erdvėje (6(b)pav.). α_1, α_2 gali būti fiksuoti naudojant validavimo duomenis arba išmokstami vykdymo metu.

Turint panašumo kriterijų D, kaip nurodyta formulėje (8), istoriniai mokymo duomenys $(\mathbf{X}_1, \ldots, \mathbf{X}_t)$ gali būti išrikiuoti pagal panašumą į $\mathbf{X}_{t+1}$. Tą padarius lieka kitas ne mažiau svarbus mokymo imties parinkimo klausimas - *kokį kiekį* N panašiausių duomenų paimti į mokymo imtį. Užfiksavus panašumo proporcijas α_1, α_2 galima parinkti N iš anksto arba taip pat išmokti vykdymo metu. $\alpha_1 = 0$ atveju N parinkimas atitinka kintamo mokymo lango ilgio nustatymą.



6 pav.: Mokymo imties parinkimas: (a) tik laike (langas), (b) tik erdvėje, (c) naujasis kriterijus laike ir erdvėje

2.2.2 FISH algoritmas mokymo imčiai parinkti

Disertacijoje sukurta FISH algoritmų šeima, susidedanti iš trijų algoritmų FISH1, FISH2 ir FISH3. Šie algoritmai naudoja suformuluotą panašumo laike ir erdvėje kriterijų mokymo imčiai parinkti. Algoritmai skiriasi parametrų α_1, α_2 ir N parinkimu. FISH1 naudoja iš anksto fiksuotas panašumo proporcijas α_1, α_2 ir mokymo imties dydį N. FISH2 naudoja fiksuotas panašumo proporcijas, bet parenka kintamą N vykdymo metu. FISH3 parenka vykdymo metu ir α_1, α_2, ir N.

Šioje santraukoje pristatome tik FISH2. Remiantis eksperimentais manome, kad dažniausiai α_1, α_2 išlieka daugmaž pastovūs konkretiems duomenims laikui bėgant, tačiau N kinta aktyviai laike. Dėl to FISH2 laikome pagrindiniu algoritmu FISH šeimoje, ir su FISH2 disertacijoje atlikta plačiausi eksperimentai.

FISH2 žingsniai iliustruoti 7 pav. Algoritmas mokymo imties dydį N parenka kryžminio validavimo būdu, naudodamas k panašiausių istorinių duomenų kaip validavimo imtį.

Mokymo imties parinkimo algoritmas FISH2

įvesties duomenys
Istoriniai duomenys $\mathbf{X}_1, \ldots, \mathbf{X}_t$ su $\mathbf{y}_1, \ldots, \mathbf{y}_t$, vektorius $\mathbf{X}_{t+1}$. Parametrai: kaimynø kiekis k, $A = \frac{\alpha_2}{\alpha_1}$ $\alpha_1 \neq 0$.
ALGORITMAS

1. Suskaičiuoti atstumus laike ir erdvėje $\mathcal{D}_i^*$ (formulė (8)) $i = 1 : t$.
2. Išrūšiuoti atstumus nuo mažiausio $\mathcal{D}_{z1}^* < \mathcal{D}_{z2}^* < \ldots < \mathcal{D}_{zt}^*$.
3. *Ciklas* $N = k : step : t$ mokymo imties dydžiui parinkti
   (a) išrinkti N mažiausių atstumų $\mathcal{D}$,
   (b) naudojant kryžminį validavimą [a] išmokyti klasifikatorių $\mathcal{L}^N$ naudojant mokymo duomenis $(\mathbf{X}_{z1}, \ldots, \mathbf{X}_{zN})$ su $(\mathbf{y}_{z1}, \ldots, \mathbf{y}_{zN})$,
   (c) testuoti $\mathcal{L}^N$ naudojant k artimiausių kaimynų $(\mathbf{X}_{z1}, \ldots, \mathbf{X}_{zk})$, gauti testavimo klaidà e_N.
4. Išrinkti klasifikatorių kuris davė mažiausią klaidą $\mathcal{L}^{N*}$, čia $N^* = \arg\min_{N=k}^{t}(e_N)$.
5. Gauti indeksus $\{z1, \ldots, zN^*\}$.

REZULTATAS
Indeksai $\mathcal{I}_t = \{z1, \ldots, zN^*\}$ kuriuos naudojant išrenkama mokymo imtis $(\mathbf{X}_{z1}, \ldots, \mathbf{X}_{zN*})$ su $(\mathbf{y}_{z1}, \ldots, \mathbf{y}_{zN*})$.

[a] testuojant $\mathbf{X}_{zk}$ šis vektorius neįeina į validavimo imtį

7 pav.: FISH2 algoritmas.

2.2.3 Eksperimentai ir rezultatai

4 disertacijos skyriuje atlikti eksperimentai turėjo du pagrindinius tikslus.

- Ištirti ar jungtinis laiko ir erdvės kriterijus reikalingas (FISH1).
- Įvertinti FISH algoritmų tikslumą
  - išmokstant vykdymo metu tik mokymo imties dydžio parametrą N (FISH2),
  - išmokstant vykdymo metu ir N, ir panašumo proporcijas α_1, α_2.

Skyriaus eksperimentams naudojome realius duomenisiš šešių dalykinių sričių, kuriose tikėtini palaipsniai pokyčiai. Duomenų charakteristikos pateikiamos 3 lentelėje.

3 lentelė: FISH tikslumo tyrimui naudoti duomenys.

	Dimensijos	**Kiekis**	**Klasių proporcijos**	**Duomenų tipas**
Luxembourgh	31	1901	0.51:0.49	realūs
Ozone	72	2534	0.94:0.06	realūs
Electricity	6	2956	0.57:0.43	realūs
German	23	1000	0.70:0.30	realūs
Vote2	16	435	0.61:0.39	realūs
Iono2	43	435	0.61:0.39	realūs

FIHS2 tikslumas palygintas su su dviem žinomais algoritmais mokymo imčiai parinkti naudojančiais tik vieną iš panašumo kriterijų: KLI [11] naudoja laiko kriterijų (langą), TSY [14] naudoja erdvės kriterijų ir baziniu mokymo algoritmu be adaptyvumo, naudojančiu visą mokymo istorija ALL.

Siekiant kuo įvairiapusiškiau ištirti pasiūlyto jungtinio kriterijaus efektą klasifikavimo tikslumui, atlikti alternatyvūs eksperimentai su keturiais baziniais klasifikatoriais: Euklidiniu (NMC), k artimiausių kaimynų (kNN), Parzeno lango (PWC) ir sprendimų medžiu (TREE) (palčiau apie šiuos bazinius klasifikatorius galima rasti [5]). Taip pat panaudoti du alternatyvūs atstumo erdvėje matai: Euklidinis $d^E(\mathbf{X}_j, \mathbf{X}_l) = \sqrt{\sum_{i=1}^{p} |\mathbf{x}_j^{(i)} - \mathbf{x}_l^{(i)}|^2}$ ir kosinuso $d^C(\mathbf{X}_j, \mathbf{X}_l) = \cos(\mathbf{X}_j, \mathbf{X}_l) = \frac{\sum_{i=1}^{p} \mathbf{x}_j^{(i)} \mathbf{x}_l^{(i)}}{\sqrt{\sum_{i=1}^{p} (\mathbf{x}_j^{(i)})^2} \sqrt{\sum_{i=1}^{p} (\mathbf{x}_l^{(i)})^2}}$, čia $\mathbf{x}_j^{(i)}$ yra i^{tasis} požymis

4 lentelė: Testavimo klaidos, Euklidinis atstumas erdvėje. • žymi statistiškai reikšmingą skirtumą FISH2 naudai, ∘ - reikšmingą skirtumą FISH2 nenaudai, − reiškia statistiškai nereikšmingą skirtumą, kai $\alpha = 0.05$.

	base	**Luxe**	**Ozon**	**Elec**	**Cred**	**Vote**	**Iono**	**REITINGAS**
FISH2	NMC	11.89	34.31	15.16	36.94	8.53	17.43	*1.33*
KLI		30.89•	22.90∘	19.97•	36.24−	11.29•	21.71•	*2.08*
TSY		35.89•	37.23•	15.47−	40.64•	11.29•	20.57−	*2.75*
ALL		39.68•	86.70•	24.84•	37.84−	11.52•	31.71•	*3.83*
FISH2	kNN	14.63	7.03	15.06	30.13	8.76	22.00	*2.08*
KLI		15.74−	7.11−	18.98•	30.03−	9.68−	22.86−	*3.00*
TSY		28.79•	7.03−	13.16∘	31.43−	10.60−	23.14−	*3.25*
ALL		11.84∘	6.99−	19.86•	28.83−	8.29−	22.29−	*1.67*
FISH2	PWC	12.37	70.79	41.08	34.33	8.99	12.86	*1.75*
KLI		14.42•	38.81∘	46.06•	34.63−	10.37−	15.14•	*3.00*
TSY		26.42•	54.72∘	43.62•	36.54−	9.68−	19.71•	*3.25*
ALL		11.68−	84.88•	43.62•	34.43−	8.53−	12.86−	*2.00*
FISH2	tree	0.37	9.99	13.54	31.03	7.37	18.00	*1.42*
KLI		0.37−	11.69•	17.36•	36.34•	9.68−	20.86−	*3.25*
TSY		0.37−	12.63•	8.97∘	37.04•	10.14−	20.29−	*3.08*
ALL		0.37−	10.50−	16.99•	32.83−	7.83−	18.57−	*2.25*

vektoriaus $\mathbf{X}_j$, p yra požymių kiekis. Taigi iš viso atlikti 4 $times2$ eksperimentiniai algoritmų palyginimai.

FISH2 algoritmo tikslumo palyginimo rezultatai su tikrais duomenimis pateikiami lentelėse 4 ir 5. Algoritmų rezultatai palyginti skaičiuojant vidutinį reitingą. Konkretiems duomenims geriausią tikslumą pasiekęs algoritmas gauna reitingą 1, blogiausią - 5. Statistinis rezultatų reikšmingumas vertintas naudojant McNeamar testą [12].

Gautas klasifikavimo tikslumo pagerėjimas naudojant FISH2 algoritmą, kuris remiasi jungtiniu panašumo laike ir erdvėje kriterijumi.

2.2.4 Pagrindiniai skyriaus rezultatai

Iki šiol srityje mokymo imčiai parinkti buvo naudojamas arba tik panašumo laike kriterijus (mokymo langai), arba tik požymių erdvėje. Vykstant palaipsnams pokyčiams aktualūs abu kriterijai. Disertacijoje sukurtas naujo tipo matas mokymo imčiai parinkti

5 lentelė: Testavimo klaidos, kosinuso atstumas erdvėje.

	base	Luxe	Ozon	Elec	Cred	Vote	Iono	REITINGAS
FISH2	NMC	12.68	35.25	15.57	38.14	8.76	16.86	*1.67*
KLI		30.89•	22.90○	19.97•	36.24−	11.29•	21.71•	*2.08*
TSY		35.89•	37.23−	15.47−	40.64−	11.29•	20.57−	*2.58*
ALL		39.68•	86.70•	24.84•	37.84−	11.52•	31.71•	*3.67*
FISH2	kNN	14.79	6.99	15.19	29.93	8.53	21.71	*1.75*
KLI		15.74−	7.11−	18.98•	30.03−	9.68−	22.86−	*3.17*
TSY		28.79•	7.03−	13.16○	31.43−	10.60•	23.14−	*3.33*
ALL		11.84○	6.99−	19.86•	28.83−	8.29−	22.29−	*1.75*
FISH2	PWC	12.68	72.25	39.26	34.83	8.29	13.34	*2.00*
KLI		14.42•	38.81○	46.06•	34.63−	10.37•	15.14−	*2.83*
TSY		26.42•	54.72○	43.62•	36.54−	9.68−	19.71•	*3.25*
ALL		11.68○	84.88•	43.62•	34.43−	8.53−	12.86−	*1.92*
FISH2	tree	0.37	10.03	12.79	31.34	7.60−	17.71	*1.71*
KLI		0.37−	11.69•	17.36•	36.34•	9.68−	20.86−	*2.83*
TSY		0.37−	12.63•	8.97○	37.04•	10.14−	20.29−	*3.06*
ALL		0.37−	10.50−	16.99•	32.83−	7.83−	18.59−	*2.40*

ne tik pagal laiką, bet ir *kartu* pagal panašumą požymių erdvėje. Parodyta analitiškai ir pagrįsta taikomaisiais pavyzdžiais, kad jungtinis kriterijus yra naudingas.

Tuo pagrindu sukurtas ir eksperimentiškai aprobuotas mokymo imties parinkimo algoritmas FISH, kuris nustato mokymo imties dydį ir išrenka mokymo vektorius naudojant jungtinį panašumo laike ir erdvėje kriterijų. Algoritmas gali išmokti panašumo laike ir erdvėje proporcijas bei mokymo imties dydį vykdymo metu. Pademonstruotas tikslumo pagerėjimas lyginant su dviem žinomais mokymo imties parinkimo algoritmais: naudojančiu tik laiko ir tik erdvės kriterijų.

Panašumo požymių erdvėje įtraukimas kartu su laiku į mokymo imties parinkimo procedūrą leidžia pagerinti klasifikavimo tikslumą, lyginant su atskirų panašumo kriterijų naudojimu, esant palaipsniams koncepcijos pokyčiams.

2.3 Pasikartojimai: kontekstinis mokymas

Tikintis pasikartojančių pokyčių modelio tikslumą galima pagerinti laiku atpažįstant žinomas apibrėžtas situacijas (gal būt jau buvusias praeityje). Suformuluotus situacijų

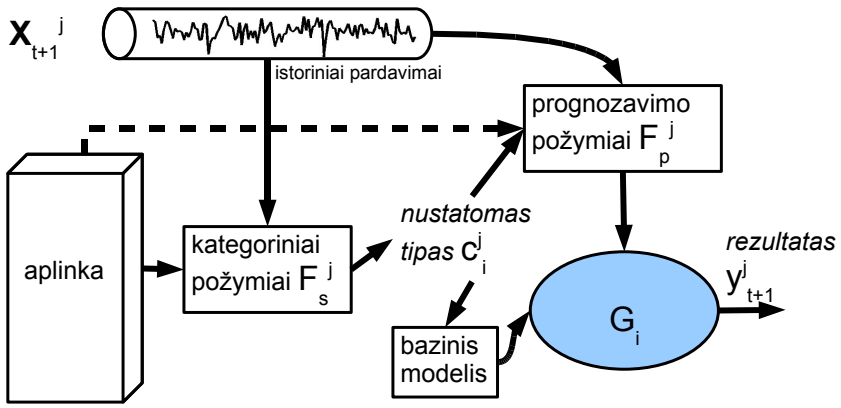
tipus vadinsime *kontekstu*. Pavyzdžiui, maisto produktai turintys sezoninius pardavimus galėtų sudaryti vieną tipą, proginiai produktai - kitą tipą. Tipai nebūtinai yra fiksuoti, jie gali kisti laikui bėgant. Pavyzdžiui, produkto išdėstymo vietos pakeitimas parduotuvėje gali pakeisti pardavimų tipą, taip pat kontraktų su tiekėjais pasikeitimas, ilgalaikės rinkodaros kampanijos.

5 disertacijose skyriuje, remiantis pardavimų kiekio prognozavimo uždaviniu, sukurta mokymo imties formavimo metodika tikintis pasikartojančių koncepcijų (situacijų). Ankstesniuose skyriuose imtis formuota tik vektorių erdvėje. Šiame skyriuje mokymo imties formavimas liečia ir vektorių, ir požymių erdvę. Mokymo imtį formuojame priklausomai nuo istorinių pardavimų „elgesio“ tipo. Metodika ir sukurtas algoritmas CAPA apima tipų suformavimą, ir mokymo imčių strategijų susiejimą bei išmokimą nustatyti tipus vykdymo metu, tam kad atitinkamai būtų suformuota mokymo imtis. Metodas nėra apribotas vien pardavimų kiekio prognozavmo uždaviniu.

Ankstesniuose skyriuose mokymo duomenys turėjo du matavimus: požymių erdvę ir laiko erdvę, t.y. $\mathbf{X}_t$ su $\mathbf{y}_t$ buvo p-matis vektorius (su žinoma klase) ir jis stebėtas laiku t. Maisto produktų pardavimų kiekio prognozavimo uždavinyje atsiranda dar vienas matavimas - produktas j. Taigi, šiuo atveju $\mathbf{X}_t^j$ bus p-matis požymių vektorius laiku t produktui j, vienu laiko momentu stebime daugelio produktų požymius, kai kurios reikšmės yra vienodos visiems produktams (pvz. ar Kalėdos), kai kurios skiriasi (pvz. ar vykdoma šio produkto rinkodaros kampanija). Reikia prognozuoti kiekvieno produkto pardavimų kiekį atskirai sekančiai savaitei. Tačiau stebint lygiagrečiai laike daug produktų, siekiant pagerinti prognozavimo tikslumą, galima suformuoti ir išmokti atskirti jų elgesio kategorijas.

2.3.1 Kontekstinis mokymo metodas

Metodą iš esmės galima vadinti mokymo imties formavimo strategijų ansambliu. Gavus testavimo vektorių $\mathbf{X}_{t+1}^j$ CAPA veikimas parodytas 8 pav. Pirmiausia išskiriami struktūriniai požymiai F_s^j, apibūdinantys produkto „elgesį“ (pvz. istorinių pardavimų vidurkis, standartinis nuokrypis), pagal šiuos požymius atpažįstamas produkto j tipas laikui t, tarkime c_i, kur $c_i \in (c_1, \ldots, c_m)$ yra fiksuota tipų aibė. Modelyje tipas c_i yra susietas su konkrečiu baziniu klasifikatoriumi bei mokymo imties parinkimo strategija, vadinsime $\mathcal{G}_i$. Tuomet jau galima mokyti konkretų klasifikatorių G_i^j atitinkamai parenkant mokymo



8 pav.: CAPA veikimas.

imtį. Jam mokyti ir sprendimui $\mathbf{y}_{t+1}^j$ priimti naudojama prognozavimo požymių erdvė F_p nėra ta pati, kaip struktūrinių požymių erdvė F_s naudota tipui atpažinti. Galutinis sprendimas priimamas: $\mathbf{y}_{t+1}^j = G_i^j(\mathbf{X}_{t+1}^j)$.

Svarbiausios CAPA paruošimo konkrečiam uždaviniui dalys yra šios:

- apibrėžti mokymo imties formavimo strategijų ir baz. klasifikatorių aibę $(\mathcal{G}_1, \ldots, \mathcal{G}_m)$,
- apibrėžti tipų aibę $(c_1, \ldots, c_m)$ struktūrinių požymių erdvė F_s bei kaip tipai bus atpažįstami (pvz. naudojant meta klasifikatorių),
- susieti kiekvieną tipą c_i su kuria nors iš strategijų $\mathcal{G}_l$: $c_i \to \mathcal{G}_l$.

2.3.2 Eksperimentai ir rezultatai

Eksperimentams panaudojome realius Olandjos didmeninės prekybos tinklo Sligro Food Group N.V. duomenis. Šiuo metu įmonė prognozavimui ir prekių atsargų valdymui naudoja 6 savaičių pardavimų istorijos slenkantį vidurkį, kurį paskui koreguoja atsakingi darbuotojai ekspertiniu būdu. Eksperimentuose CAPA rezultatus lyginome su šiuo būdu MA6: $\mathbf{y}_{t+1} = (\mathbf{y}_t + \ldots + \mathbf{y}_{t-5})/6$ bei su naivia prognoze, vadinama „rytoj bus taip kaip šiandien“ MA1: $\mathbf{y}_{t+1} = \mathbf{y}_t$. Abu šie metodai nenaudoja jokių papildomų požymių, tik pardavimų istoriją. CAPA naudoja papildomų požymių erdvę.

Eksperimentiniai duomenys susideda iš 538 produktų pardavimų istorijos. Tai daugiau nei 2 metai (120 savaičių). Papildomi duomenys požymių erdvėms tai: rinkodaros ak-

6 lentelė: Prognozvimo klaidos.

Tipas	MA1	MA6	CAPA	vidutinis dydis
Mokymo grupė				
„atsitiktiniai“	1.000	1.341	1.730	323
„prognozuojami“	1.000	0.987	**0.940**	115
Testavimo grupė				
„atsitiktiniai“	1.000	1.383	1.762	80
„prognozuojami“	1.000	0.970	**0.950**	20
Atsitiktinis skirstymas į tipus				
„atsitiktiniai“	1.000	1.306	1.604	50
„prognozuojami“	1.000	1.296	**1.593**	50

cijos, tinklo suminių pardavimų istorija, kalendorinės šventės, orų, kritulių duomenys. Prognozavimui absoliučios pardavimų reikšmės suskirstytos į 8 lygius atskirai kiekvienam produktui. Kaip bazinis modelis naudojama tiesinė regresija.

Tikslumą vertinome pagal sąlyginę klaidą [9]: $MASE = \frac{1}{n}\Sigma_{t=1}^{n}|\frac{e_t}{MAE(Baseline)}|$, kur e_t yra absoliuti prognozavimo klaida laiku t, $MAE(Baseline)$ yra absoliuti vidutinė bazinio metodo klaida. Įvertinimui baziniu laikome MA1.

Santraukoje pateikiame rezultatus kai produktai suskirstyti į du tipus, kuriuos įvardinome kaip „prognozuojamą“ ir „atsitiktiniai“Atpažinus šiuos tipus pirmajam taikyta regresija su plačia išorinių požymių erdve, antrajam - MA1 naudojantis tik vieną požymį (istorinius pardavimus). Pateikiame prognozavimo rezultatus laikant, kad tipas c yra fiksuotas konkrečiam produktui, 6 pav. Rezultatai parodyti atskirai mokymo ir testavimo grupei reiškia produktų padalinimą tipams išmokti. Galutinis prognozavimas visais atvejais buvo atliekamas progresinio mokymo principu kuris pristatytas 2 skyriuje. Lentelėje parodyti ir rezultatai taikant atsitiktinį skirstymą į tipus, t.y. nesimokant kategorizuoti. Kadangi išmokimas kategorizuoti yra CAPA metodo esmė, šis atsitiktinis kategorizavimas įtrauktas tam, kad būtų matomas efektas galutiniam tikslumui. Matosi, kad jeigu nebūtų išmokstamo kategorizavimo, slenkančio vidurkio metodas būtų tikslesnis.

2.3.3 Pagrindiniai skyriaus rezultatai

Remiantis pardavimų kiekio prognozavimo uždaviniu, kuriam aktualūs koncepcijų pasikartojimai, sukurtas bei eksperimentiškai aprobuotas kontekstinis mokymo imties formavimo metodas CAPA, kurio esmė išmokstamas prekės kategorijos ir iš to sekantis mokymo imties formavimo strategijos susiejimas. CAPA identifikuoja objekto tipą ir pagal jį interaktyviai formuoja požymių erdvę bei parenka mokymo vektorius. Objekto tipo identifikavimas remiantis suformuluotais struktūriniais požymiais siekiant atitinkamai tipui suformuoti mokymo imtį yra moksliškai naujas. Eksperimentiškai pademonstruotas prognozavimo tikslumo pagerėjimas 5% yginant su baziniu metodu.

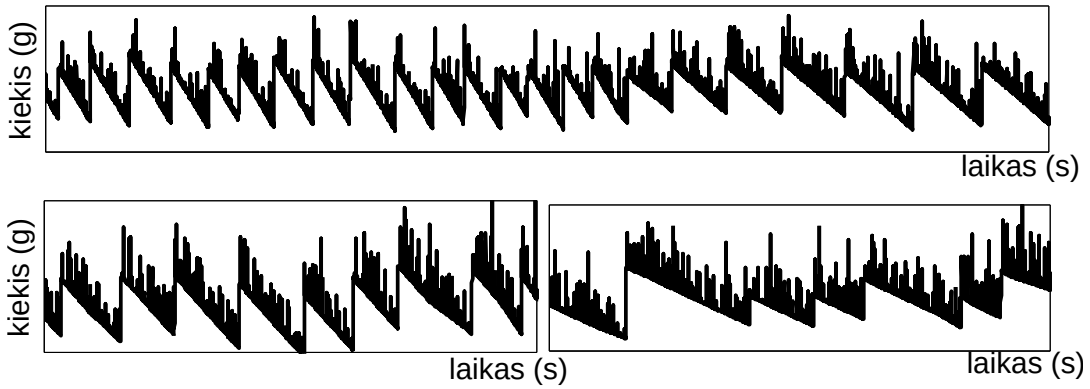
Konteksto panaudojimas mokymo imčiai parinkti, surišant istorinių pardavimų tipus su mokymo imties formavimo strategijomis, bei išmokstant atskirti tipus vykdymo metu naudojant struktūrinius požymius, leidžia pagerinti prognozavimo tikslumą sprendžiant maisto produktų pardavimo kiekio prognozavimo uždavinį, kur tikimasi koncepcijø pasikartojimo bei pokyčių.

2.4 Pramonio katilo atvejo analizė

6 skyriuje sprendžiame masės kitimo įvertinimo uždavinį pramoniniam katilui. Įvertinimas reikalingas katilo kontrolės sistemai.

Katilas skirtas šildymui ir gali kūrenti skirtingą kurą įvairiomis proporcijomis, iš to kyla koncepcijos pokyčiai. Ypač dėl biokuro, kuris negali būti pilnai vienalytis. Taip pat gali būti naudojami keli kuro tipai vienu metu, pvz. biokuras su anglimi. Kitas koncepcijos pokyčių šatinis yra kuro padavimas. Jis nėra pilnai automatizuotas ir padavimo „stilius“ priklauso nuo operatoriaus (žmogaus). Vienas gali paduoti greičiau, kitas lėčiau su pertraukėlėmis. Dėl mechaninių katilo dalių judėjimo daviklių signalas gaunamas iškraipytas. Be to pasitaiko staigių signalo nuokrypių aukštyn dėl smulkių dalelių įstrigimo matavimo procese.

Signalo pavyzdys pateikiamas 9 pav. Signalas yra vienmatis $(\mathbf{x}_1, \ldots, \mathbf{x}_t, \ddots)$. Yra dvi veikimo fazės: kuro padavimo ir kūrenimo. Kūrenimas nesustoja ir kuro pardavimo metu. Kontrolės mechanizmui reikalinga masės kitimo prognozė arba įvertis realiu laiku. Taip pat reikia nustatyti, kurioje fazėje yra sistema: kuro padavimo ar kūrenimo (pokyčio



9 pav.: Masės signalas gautas iš katilo daviklių.

nustatymas). Signalo nuokrypiai gali būti labai panašūs į kuro padavimo fazės pradžią, algoritmas turėtų sugebėti juos atskirti.

Disertaciniame darbe sukurtas masės kitimo įvertinimo algoritmas OMFP, kurį sudaro: signalo modelis, nuokrypių eliminavimo ir pokyčių nustatymo mechanizmas. Nustačius pokytį parenkamas tinkamas mokymo langas ir signalo modelis permokomas, fiksuojami nauji parametrai.

Signalo modelis. Turime pradinį signalą $\mathbf{x} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_t, \ldots, \mathbf{x}_n)$, $\mathbf{x} \in \Re^1$. Reikalingas įvertis
$hat\mathbf{y} = \mathcal{F}(\mathbf{x})$, pasirinktas modelis:

$$\hat{\mathbf{y}}_t = a_t^{(2)}\mathbf{x}_t^2 + a_t^{(1)}\mathbf{x}_t + a_t^{(0)}, \tag{9}$$

čia $a_t^{(2)}, a_t^{(1)}, a_t^{(0)}$ yra koeficientai kuriuos reikia išmokti iš istorinių duomenų. Prieš išmokstant koeficientus reikia išvalyti mokymo duomenis nuo triukšmo (nuokrypių) ir nustatyti mokymo lango ilgį. Koeficientai įvertinami mažiausių kvadratų metodu.

Nuokrypių eliminavimui naudojame slenkstį Tr_{out}, kuris išmokstamas iš anksto naudojant validavimo duomenis. Dėl mechaninio triukšmo prieš naudodami slenkstį pritaikome signalui slenkantį vidurkį. Nustatytas nuokrypis yra pakeičiamas artimiausių signalų vidurkiu.

Pokyčių ir fazės nustatymui taip pat naudojamas iš anksto išmokstamas slenkstis Tr_{ch}, prieš tai eliminavus signalo trendus. Mokymo lango ilgis dažniausiai sutampa su pokyčio tašku. Esant neužtikrintumui dėl pokyčio, naudojamos mokymo lango taisyklės,

7 lentelė: Vidutinės prognozavimo klaidos.

Vėlavimas	**0**	**2**	**4**	**0**	**2**	**4**	**0**	**2**	**4**
Duomenys		**A**			**B**			**C**	
OMFP	**29.4**	**27.8**	**27.6**	**20.9**	**16.6**	**16.3**	**13.0**	**10.3**	**10.1**
MA3	64.0	66.4		47.2	46.9		35.6	35.2	
MA5	51.9		39.9	45.3		41.7	33.9		32.5
MA10	54.8			53.7			37.2		
win50	45.0	44.4	44.4	34.3	32.0	32.0	16.7	15.2	15.2
all	1269	1267	1265	1310	1308	1306	1021	1019	1019
known	32.0	30.6	31.3	47.9	45.1	44.6	16.5	15.7	16.3

kurios išmokstamos iš anksto naudojant validavimo duomenų dalį.

Eksperimentams panaudoti realūs prototipinio katilo duomenys A, B ir C. A ir B naudotas biokuras, C - anglis. A duomenų ilgis 50977 atskaitymų, B ir C po 25177 atskaitymų. A duomenyse yra 24 atskiros kuro padavimo fazės, B - 9, C - 6. Duomenys pavaizduoti 9 pav.

Sukurto OMFP algoritmo tikslumas lygintas su slenkančiais vidurkiais (3,5 ir 10 s), fiksuoto ilgio mokymo langu (win50), visos istorijos naudojimu (all) ir mokymo lango nukirtimu ties žinomu pokyčio tašku (known). A duomenys naudoti slenksčiams išmokti.

Santraukoje pateikiami tik galutiniai rezultatai atitinkamai su skirtingais priimtinais vėlavimais 7 lentelėje. OMFP pasiekia geriausią tikslumą lyginant su kitais išvardintais metodais.

Sukurtas OMFP algortimas pritaikytas masės judėjimui degimo proceso metu įvertinti, mokymo imtis parenkama „lango“ principu.

Sukurtasis masės kitimo įvertintinimo metodas pramoniniam katilui, leidžia pasiekti tikslesnius įverčius, nei nenaudojant adaptyvumo pokyčiams, ir tuo būdu patobulinti katilo kontrolės sistemą.

2.5 Ginamosios išvados

Disertaciniame darbe patobulintos žinomos mokymo strategijos esant staigiems, palaipsniams ir pasikartojantiems pokyčiams. Sukurti ir eksperimentiškai aprobuoti keturi adaptyvaus mokymo imties formavimo algoritmai (WR*, FISH, CAPA, OMFP), kurie leidžia pagerinti klasifikavimo bei prognozavimo tikslumą besikeičiančiose aplinkose, esant atitinkamai kiekvienam iš trijų pokyčių tipų, lyginant su žinomais algoritmais bei pasyviomis strategijomis (naudojant visus istorinius duomenis).

Mokymo lango ir pokyčio taško teorinis atskyrimas parenkant kintamą mokymo lango ilgį leidžia pagerinti klasifikavimo tikslumą esant staigiam koncepcijos pokyčiui.

Panašumo požymių erdvėje įtraukimas kartu su laiku į mokymo imties parinkimo procedūrą leidžia pagerinti klasifikavimo tikslumą, lyginant su atskirų panašumo kriterijų naudojimu, esant palaipsniams koncepcijos pokyčiams.

Konteksto panaudojimas mokymo imčiai parinkti, surišant istorinių pardavimų tipus su mokymo imties formavimo strategijomis, bei išmokstant atskirti tipus vykdymo metu naudojant struktūrinius požymius, leidžia pagerinti prognozavimo tikslumą sprendžiant maisto produktų pardavimo kiekio prognozavimo uždavinį, kur tikimasi koncepcijų pasikartojimo bei pokyčių.

Sukurtasis adaptyvus masės kitimo įvertintinimo metodas pramoniniam katilui, veikiančiam kintamomis kuro tipų ir kuro padavimo sąlygomis, leidžia pasiekti tikslesnius įverčius, nei nenaudojant adaptyvumo pokyčiams, ir tuo būdu patobulinti katilo kontrolės sistemą.

3 Doktorantės publikacijos disertacijos tema su VU prieskyra

Pateikiami moksliniai straipsniai disertacijos tema Vilniaus universiteto vardu. Pilną I. Žliobaitės mokslinių straipsnių disertacijos tema periodiniuose ir neperiodiniuose leidiniuose sąrašą galima rasti disertacijoje.

Publikacijos periodiniuose leidiniuose

1. Kuncheva, L.I. and Žliobaitė, I. (2009). On the Window Size for Classification in Changing Environments. Intelligent Data Analysis 13(6), p. 861-872. ISSN:1088-467X [ISI]

2. Žliobaitė, I. (2009). Combining Time and Space Similarity for Small Size Learning under Concept Drift. Proceedings of ISMIS 2009, the 18th international symposium on Methodologies for Intelligent Systems, book: Foundations of Intelligent Systems, Zhong, N.; Ras, Z.W.; Tsumoto, S.; Suzuki, E. (Eds.), (Lecture Notes in Computer Science LNCS 5722), p. 412-421. ISSN: 0302-9743 [ISI proceedings]

3. Žliobaitė, I. (2007). Introduction of New Expert and Old Expert Retirement under Concept Drift. Book: Progress in Pattern Recognition, series: Advances in Pattern Recognition. S. Singh, M. Singh (Eds.) 2007, XIII, p.64-74. ISSN: 1617-7916 [ISI proceedings]

4. Žliobaitė, I. (2007). Ensemble Learning for Concept Drift Handling the Role of New Expert. Poster proceedings of the 5th international conference on on Machine Learning and Data Mining in Pattern Recognition MLDM 2007, book: Machine Learning and Data Mining in Pattern Recognition, Petra Perner (Ed.). IBaI publishing, p. 251-260. ISSN: 1864-9734

5. Raudys, Š., Žliobaitė, I. (2006). The Multi-Agent System for Prediction of Financial Time Series. Proceedings of the 8th international conference on Artificial Intelligence and Soft Computing ICAISC 2006 (Lecture Notes in Artificial Intelligence LNAI 4029), p. 653-662. ISSN: 0302-9743 [LNAI publikavimo metu buvo ISI, šiuo metu ISI proceedings]

6. Raudys, Š., Žliobaitė, I. (2005). Prediction of Commodity Prices in Rapidly Changing Environments. Pattern Recognition and Data Mining, proceedings of the 3rd international conference on Advances in Pattern Recognition, ICAPR 2005 (Lecture Notes in Computer Science LNCS 3686), p. 154-163. ISSN: 0302-9743 [LNCS publikavimo metu buvo ISI, šiuo metu ISI proceedings]

Publikacijos recenzuojamuose konferencijų leidiniuose

7. Žliobaitė, I., Kuncheva, L. (2009). Determining the Training Window for Small Sample Size Classification with Concept Drift. Proceedings of 2009 IEEE international conference on Data Mining Workshops, the 1st international workshop on Transfer Mining (TM-09), p. 447-452. ISBN: 978-8-7695-3895-2 [LMT database: IEEE/IEE]

8. Žliobaitė, I. (2008). Expected Classification Error of the Euclidean Linear Classifier under Sudden Concept Drift. Proceedings of the 5th international conference on Fuzzy Systems and Knowledge Discovery (FSKD 2008). IEEE Computer Society: vol 2, p. 29-33. ISBN: 978-0-7695-3305-6 [ISI proceedings]

9. Kuncheva, L., Žliobaitė, I. (2008). Linear Discriminant Classifier (LDC) for Streaming Data with Concept Drift. Proceedings of SSPR/SPR 2008, joint IAPR international workshop (Lecture Notes in Computer Science LNCS 5342), p: 4. ISSN: 0302-9743 [LMT database: SpringerLINK]

4 Trumpos žinios apie doktorantę

Indrė Žliobaitė 2000 m. baigė Vilniaus licėjų. 2003 m. baigė Stokholmo Aukštąją Ekonomikos mokyklą Rygoje (SSE Rīga), kur gavo ekonomikos ir verslo bakalauro laipsnį (BSc). 2006 m. įgijo informatikos magistro laipsnį (MSc) Vilniaus universitete. Nuo 2006 iki 2010 m. - Vilniaus universiteto doktorantė.

I. Žliobaitė stažavosi Bangoro Universitete Jungtinėje Karalystėje (3 mėn.) bei Eindhoveno Technologijų Universitete Nyderlanduose (3 + 1 mėn.). Atliko mokslinę praktiką Helsinkio informacinių technologijų institute Suomijoje (3 mėn.). Visose trijose institucijose skaitė kviestinius pranešimus adaptyvaus mokymo duomenų gavyboje tema.

I. Žliobaitė recenzuoja straipsnius šiems žurnalams: Pattern Recognition, Pattern Recognition Letters, European Journal of Operational Research, Journal of Computational and Graphical Statistics, Journal of Pattern Recognition Research.

5 Santrauka

Šiandieninėje, dinamiškai besikeičiančioje aplinkoje reikalingi adaptyvūs duomenų gavybos metodai. Nepageidaujamų laiškų klasifikatoriai, rekomendavimo bei rinkodaros, įsilaužimų į kompiuterinius tinklus aptikimo, verslo rodiklių prognozavimo bei sprendimų priėmimo sistemos turi nuolat persimokyti reaguoti į besikeičiančius duomenis. Stacionarioje aplinkoje kuo daugiau mokymo duomenų - tuo tikslesnis modelis. Besikeičiančioje aplinkoje seni duomenys blogina tikslumą. Tokiu atveju, vietoje visų turimų istorinių duomenų panaudojimo, gali būti tikslingai išrenkama tik tam tikra jų dalis, pvz. naudojamas mokymo langas (tik naujausi duomenys).

Tiriamojo darbo objektas yra adaptyvūs mokymo metodai, kurie remiasi kryptingu mokymo imties formavimu. Patobulintos žinomos mokymo strategijos esant staigiems, palaipsniams ir pasikartojantiems pokyčiams. Sukurti ir eksperimentiškai aprobuoti keturi adaptyvaus mokymo imties formavimo algoritmai, kurie leidžia pagerinti klasifikavimo bei prognozavimo tikslumą besikeičiančiose aplinkose, esant atitinkamai kiekvienam iš trijų pokyčių tipų. Naudojant generuotus bei realius duomenis, eksperimentiškai parodytas klasifikavimo bei prognozavimo tikslumo pagerėjimas, lyginant su visų istorinių duomenų naudojimu mokymui, bei žinomais šioje srityje naudojamais adaptyviais mokymo algoritmais. Sukurta metodika pritaikyta pramoninio katilo atvejui, jungiančiam kelis aplinkos pokyčių tipus.

6 Summary

We live in a dynamic world, where changes are a part of everyday life. When there is a shift in data, the classification or prediction models need to be adaptive to the changes. In data mining the phenomenon of change in data over time is known as *concept drift*. Changes in underlying data might occur due to changing personal interests, changes in population, adversary activities or they can be attributed to a complex nature of the environment.

This thesis focuses on adaptive supervised learning techniques, where adaptivity to changes in data over time is achieved by selective training set formation. Our research design follows the three main drift types, starting from sudden change, via gradual drift to reoccurring concepts. We develop methodological contributions to concept drift phenomenon in data mining tasks as well as four algorithms for training set formation under different application contexts and expected change types.

There was no explicit distinction between the change point and the start of the training window in supervised learning under concept drift. The historical data was dropped as soon as a sudden change was detected. In Chapter 3 we made an explicit theoretical distinction between the sudden change point and the training window. We demonstrated that the impact of taking the difference into account to the classification accuracy is increasing along with the more complex data. Based on the theoretical distinction we developed a training window resizing algorithm WR* and demonstrated an improvement in classification accuracy as compared to the existing algorithms for variable window size, which do not make this distinction. Theoretical distinction between the training window and the change point when determining a variable window size allows to improve generalization performance under sudden concept drift.

So far either temporal instance selection (training windows) or instance selection in feature space was used for learning under concept drift. In Chapter 4 we developed a new distance measure unifying the distances in time and feature space for training set selection. We argued that both criteria are relevant under gradual concept drift and demonstrated this on real datasets from six domains. Using the new distance measure we developed a family of training set selection algorithms FISH. The three algorithms FISH1, FISH2 and FISH3 differ in determining the training set size and the proportion of time and space in the developed distance measure. In FISH2 only set size is learnable

online, while in FISH3 both set size and the proportion of time and space are learnable online. The extensive numerical experiments using four alternative base classifiers and two alternative distance in space measures on six real datasets demonstrated statistically significant improvement in the classification accuracy as compared to the two existing adaptive algorithms, which use only time and only space criterion. Integration of similarity in time and feature space when selecting training set allows to improve generalization performance as compared to using only time or only space criterion under gradual concept drift.

Using a real problem of food sales prediction, for which recurring concepts are relevant, we developed and experimentally validated a contextual method CAPA for training set formation in Chapter 5. We demonstrated that identifying and learning to recognize the types of historical behavior allows to form a training set in a way that this historical information contributes to the present prediction accuracy. CAPA forms a training set interactively, based on the type of historical behavior, which is determined employing structural features. We showed that online reassignment of the categories increases the prediction accuracy. The experiments demonstrated 5% improvement in the testing prediction accuracy as compared to the baseline prediction, which is relevant for the field applications. Contextual training set formation, while connecting the types of historical sales with the training set formation strategies and learning to recognize the types online using structural features, allows to improve generalization performance in food sales prediction task, where reoccurring concepts are expected.

We developed a mass flow estimation method OMFP for an industrial boiler in Chapter 6. OMFP takes into account concept drifts using a tailored training window strategy. The developed adaptive method for online estimation of the mass flow for an industrial boiler, which operates using a changing mix of fuel and changing input styles, allows to achieve more accurate estimates than using no adaptivity to changes and this way allows to improve the control system of the boiler.

The thesis contributes to understanding concept drift problem in general and training set selection under concept drift in particular.

Literatūra

[1] C. Aggarwal. Towards systematic design of distance functions for data mining applications. In *KDD '03: Proc. of the 9th ACM SIGKDD int. conf. on Knowledge discovery and data mining*, pages 9–18. ACM, 2003.

[2] A. Bifet and R. Gavalda. Learning from time-changing data with adaptive windowing. In *Proc. of SIAM int. conf. on Data Mining (SDM'07)*, pages 443–448. SIAM, 2007.

[3] V. Chandola, A. Banerjee, and V. Kumar. Anomaly detection - a survey. *ACM Computing Surveys*, 41(3):article no. 15, 2009.

[4] J. Demsar. Statistical comparisons of classifiers over multiple data sets. *Journal of Machine Learning Research*, 7:1–30, 2006.

[5] R. Duda, P. Hart, and D. Stork. *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[6] K. Fukunaga and R. R. Hayes. Estimation of classifier performance. *IEEE Transactions on Pattern Analysis Machine Intelligence*, 11(10):1087–1101, 1989.

[7] J. Gama, P. Medas, G. Castillo, and P. Rodrigues. Learning with drift detection. In *Advances In Artificial Intelligence, Proc. of the 17th Brazilian symposium on Artificial Intelligence (SBIA 2004)*, volume 3171 of *LNAI*, pages 286–295. Springer, 2004.

[8] H. Hotelling. The generalization of student's ratio. *Annals of Mathematical Statistics*, 2(3):360378, 1931.

[9] R. J. Hyndman and A. B. Koehler. Another look at measures of forecast accuracy. *Int. Journal of Forecasting*, 22(4):679–688, 2006.

[10] A. Jain, M. Murty, and P. Flynn. Data clustering: a review. *ACM Computing Surveys*, 31(3):264–323, 1999.

[11] R. Klinkenberg and I. Renz. Adaptive information filtering: Learning drifting concepts. In *Proc. of AAAI-98/ICML-98 workshop Learning for Text Categorization*, pages 33–40, 1998.

[12] L. I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004. ISBN 0471210781.

[13] S. Raudys. *Statistical and neural classifiers: an integrated approach to design*. Springer-Verlag, 2001.

[14] A. Tsymbal, M. Pechenizkiy, P. Cunningham, and S. Puuronen. Dynamic integration of classifiers for handling concept drift. *Information Fusion*, 9(1):56–68, 2008.