VILNIAUS UNIVERSITETAS

Povilas Daniušis

POŽYMIŲ IŠSKYRIMAS OPTIMIZUOJANT PRIKLAUSOMUMO STRUKTŪRĄ

Daktaro disertacijos santrauka
Fiziniai mokslai, informatika (09P)

Vilnius, 2012

Disertacija rengta 2008 - 2011 metais Vilniaus universitete.

Mokslinis vadovas:

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

Disertacija ginama Vilniaus universiteto Informatikos mokslo krypties taryboje:

Pirmininkas

prof. dr. Rimantas Vaicekauskas (Vilniaus universitetas, fiziniai mokslai, informatika 09P),

Nariai:

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

prof. habil. dr. Algimantas Jonas Bikelis (Vytauto Didžiojo universitetas, fiziniai mokslai, matematika - 01P).

prof. dr. Romas Baronas (Vilniaus universitetas, fiziniai mokslai, informatika - 09P),

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

Oponentai:

prof. habil. dr. Antanas Žilinskas (Vilniaus Universitetas, fiziniai mokslai, informatika - 09P),

prof. dr. Dalius Navakauskas (Vilniaus Gedimino technikos universitetas, technologijos mokslai, informatikos inžinerija - 07T).

Disertacija bus ginama viešame Informatikos mokslo krypties tarybos posėdyje 2012 m. rugsėjo mėn. 18 d. 14 val.

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

Disertacijos santrauka išsiuntinėta 2012 m. rugpjūčio mėn. 17 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje.

VILNIUS UNIVERSITY

Povilas Daniušis

FEATURE EXTRACTION VIA DEPENDENCE STRUCTURE OPTIMIZATION

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

Vilnius, 2012

The dissertation work was carried out at Vilnius University from 2008 to 2011.

Scientific supervisor:

assoc. prof., dr. Pranas Vaitkus (Vilnius University, physical sciences, mathematics-01P).

The defence council:

Chairman

prof. dr. Rimantas Vaicekauskas (Vilnius University, physical sciences, informatics - 09P),

Members:

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

prof. habil. dr. Algimantas Jonas Bikelis (Vytautas Magnus University, physical sciences, mathematics - 01P).

prof. dr. Romas Baronas (Vilnius University, physical sciences, informatics - 09P),

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

Opponents:

prof. habil. dr. Antanas Žilinskas (Vilnius University, physical sciences, informatics - 09P),

prof. dr. Dalius Navakauskas (Vilnius Gediminas Technical University, technological sciences, informatics engineering - 07T).

The dissertation will be defended at the public meeting of the council on the 18st of September, 2012 at 14:00.
Adress: VU MIF Distance Learning Center, Šaltinių 1A, LT-03225 Vilnius.

The summary of the dissertation was distributed on the 17st of August, 2012.
The dissertation is available at the library of Vilnius University.

Turinys

1 Įvadas

Kompiuterių mokslo šaka, nagrinėjanti empirinių duomenų panaudojimą konstruojant automatines sprendimų priėmimo sistemas vadinama sistemų mokymu[1]. Tarkime, turime mokymo imtį, susidedančią iš įėjimo stebėjimų x_i, $i = 1, 2, ..., m$. Priklausomai nuo to, ar įėjimo stebėjimai susieti su priklausomu kintamuoju y_i, sistemų mokymo metodus galima suskirstyti į tris kategorijas: prižiūrimo mokymo[2] metodai nusakyti įėjimo stebėjimams, asocijuotiems su priklausomu kintamuoju, dalinai prižiūrimo mokymo[3] metoduose išnaudojami tiek asocijuoti su priklausomu kintamuoju įėjimo stebėjimai, tiek ir tie, kuriems priklausomas kintamasis nebuvo stebėtas, ir galiausiai neprižiūrimo mokymo[4] metodai remiasi tik įeities stebėjimais. Atitinkamos imtys taip pat vadinamos prižiūrimojo, dalinai prižiūrimo bei neprižiūrimo mokymo imtimis.

Daugelyje aktualių sričių duomenys x_i ar y_i yra didelio matavimo vektoriai ar struktūrizuoti objektai (pvz. grafai, tekstai ir kt.). Tokių sričių pavyzdžiais galėtų būti vaizdų analizė, teksto apdorojimas, bioinformatika ir kt. Minėtais atvejais požymių išskyrimas[5] yra pirmasis duomenų analizės žingsnis, dažnai iš esmės įtakojantis tolimesnio tyrimo sėkmę. Požymių išskyrimą suprantame kaip transformaciją, atvaidzuojančią stebėjimo kimtamąjį x_i į mažo matavimo vektorių, kuriame sukoncentruota esminė taikytoją dominanti informacija. Požymių išskyrimo metodo sąvoka padengia konceptualiai plačią metodų aibę, tad natūralu kad optimalus metodas dažnai priklauso nuo konkrečios situacijos: bendri algoritmai, tinkantys įvairiems duomenims, dažnai yra kur kas mažiau efektyvūs nei optimizuoti konkrečiam uždaviniui, ar pritaikyti prie duomenų specifikos. Iš kitos pusės, bendri algoritmai nereikalauja gilintis į duomenų specifiką.

1.1 Tyrimų objektas

Šio disertacinio darbo tyrimo objektas yra prižiūrimo bei dalinai prižiūrimo mokymo požymių išskyrimo metodai, optimizuojantys priklausomumo struktūrą. Nagrinėjamos dvi tokios struktūros:

[1] angl. machine learning.
[2] angl. supervised learning.
[3] angl. semi-supervised learning.
[4] angl. unsupervised learning.
[5] angl. feature extraction.

- Maksimizuojamas priklausomumas tarp požymių bei priklausomojo kintamojo,
- Papildomai siekiama, jog požymių tarpusavio priklausomumas būtų minimalus.

1.2 Darbo tikslas ir uždaviniai

Darbo tikslas yra pasiūlyti ir ištirti universalius prižiūrimo ir dalinai prižiūrimo mokymo požymių išskyrimo algoritmus, besiremiančius priklausomumo struktūros optimizavimu.

Universalumą šiuo atveju suprantame, jog algoritmas turi būti korektiškai apibrėžtas struktūrizuotiems duomenimis, veikti tiek tiesiniais, tiek ir netiesiniais atvejais, bei nenaudoti informacijos apie uždavinį, kuris bus sprendžiamas naudojant algoritmo pateiktus požymius.

Darbo tikslui pasiekti keliami šie uždaviniai:

1 uždavinys (U1): Sukonstruoti naujus prižiūrimo mokymo požymių išskyrimo algoritmus, besiremiančius priklausomumo struktūros optimizavimu;

2 uždavinys (U2): Pritaikyti pasiūlytus algoritmus dalinai prižiūrimo mokymo imtims bei ištirti eksperimentiškai;

3 uždavinys (U3): Eksperimentiškai ištirti kaip priklausomumo matas įtakoja požymių kokybę;

4 uždavinys (U4): Eksperimentiškai ištirti skirtingų priklausomumo struktūrų efektyvumą;

5 uždavinys (U5): Eksperimentiškai ištirti pasiūlytų algoritmų efektyvumą, kai duomenys yra netiesinai ar struktūrizuoti bei palyginti juos su egzistuojančiais požymių išskyrimo algoritmais.

1.3 Tyrimų metodika

Disertacijoje nagrinėjami U1-U5 uždaviniai sprendžiami taikant teorinę bei eksperimentinę metodiką. Matematinis aparatas reikalingas siūlomiems algoritmams formuluoti remiasi tiesine algebra, funkcine analize bei daugiamate statistika. Iš kitos pusės, teoriškai

požymių išskyrimo algoritmų efektyvumą pamatuoti sunku, kadangi realiuose uždaviniuose stebėjimai dažniausiai turi nežinomas tikimybines charakteristikas. Todėl norint palyginti algoritmus taikoma empirinė tyrimų metodika - klasifikuojami požymiai, sugeneruoti iš įvairių įprastinio bei daugiažymio klasifikavimo[6] duomenų sekų, vertinamas klasifikavimo tikslumas, bei kiti efektyvumo matai. Siekiant gauti statistiškai patikimus rezultatus naudojami statistiniai testai. Eksperimentų rezultatai leidžia daryti ne tik kiekybines, bet ir kokybines išvadas: identifkuoti uždavinius ar situacijas, kuriems spręsti viena ar kita modifikacija yra efektyvesnė.

1.4 Svarbiausi rezultatai

Pagrindiniai disertacijos rezultatai yra šie:

- Pasiūlyti du priklausomumo struktūros optimizavimu besiremiantys požymių išskyrimo algoritmai: HBFE ir HSCA;
- HBFE ir HSCA išplėsti dalinai prižiūrimo mokymo imtims;
- Pasiūlyti algoritmai ištirti ir palyginti su alternatyviais algoritmais empiriškai, naudojant viešai prieinamas duomenų bazes.

1.4.1 Mokslinis naujumas

Požymių išskyrime priklausomumas tradiciškai nusakomas informacijos teorijos matais, kurie praktiniam taikymui dažnai yra nepatogūs. Nors HSIC ir anksčiau buvo taikomas sprendžiant kai kuriuos požymių išrinkimo bei išskyrimo uždavinius, žinių atsakyti į U1-U5 uždaviniuose iškeltus klausimus nepakako.

Disertacije pasiūlytas papildomas požymių tarpusavio priklausomumo HSIC įvertinio minimizavimas, nepaslinktojo HSIC įvertinio panaudojimas požymių išskyrimui bei Laplaso reguliarizacijos taikymas pasiūlytiems algoritmams adaptuoti dalinai prižiūrimo mokymo imčių atvejams autoriaus žiniomis yra nauji.

[6] angl. multi-label classification.

1.4.2 Aktualumas bei praktinis reikšmingumas

Daugelis svarbių sistemų mokymo uždavinių reikalauja neprisirišimo prie klasikinės vektorinės duomenų struktūros. Tiek HBFE tiek ir HSCA remiasi branduolių metodais [7], kurie yra pritaikomi bet kokios struktūros duomenims (pvz. grafams, tenzoriams ir pan.). HBFE ir HSCA tikslo funkcijos remiasi priklausomumo matais, todėl svarbus pasiūlytų algoritmų privalumas yra neprisirišimas prie sistemų mokymo uždavinio, kuris bus sprendžiamas naudojat išskirtus požymius. Tai ypač aktualu jei yra svarbu gauti kompaktišką pradinių duomenų reprezentaciją, kai iki galo nėra žinoma kas su jais bus daroma ateityje. Disertacijoje aprašyti eksperimentų rezultatai rodo, jog pasiūlytieji algoritmai efektyvūs su įvairių veiklos sričių duomenimis.

1.5 Aprobavimas

Doktoranto rezultatai disertacijos tema publikuoti 2 moksliniuose straipsniuose periodiniuose recenzuojamuose leidiniuose, iš jų vienas straipsnis įtrauktas į ISI proceedings sąrašą. Straipsnių sąrašas Vilniaus universiteto institucijos vardu pateikiamas 3 skyriuje. Autorius dalyvavo ir pristatė rezultatus tarptautinėse mokslinėse konferencijose: International Conference on Number Theory dedicated to the 60th birthday of Professor Antanas Laurinčikas (2008), IDEAL (International Conference on Intelligent Data Engineering and Automated Learning) (2009).

1.6 Disertacijos struktūra

Disertacinis darbas susideda iš šešių skyrių. Pirmame disertacijos skyriuje skaitytojas supažindinamas su darbo problematika, uždaviniais bei svarbiausiais rezultatais. Antrąjame skyriuje pateikiamos svarbiausios matematinės bei algoritminės žinios, naudojamos šiame disertaciniame darbe. Trečiojo skyriaus paskirtis - supažindinti skaitytoją su egzistuojančiais požymių išskyrimo algoritmais bei iliustruoti antrojo skyriaus medžiagą. Ketvirtąjame skyriuje sukoncentruoti nauji šio darbo rezultatai: HBFE bei HSCA algoritmai bei modifikacijos daliniai prižiūrimo mokymo imtims. Penktąjame skyriuje aprašomi eksperimentai, praktiškai iliustruojantys pasiūlytų algoritmų efektyvumą

[7] angl. kernel methods.

sprendžiant įprastinio bei daugiažymio klasifikavimo uždavinius. Disertacija baigiama šeštuoju skyriumi, kuriame pateikiamos darbo išvados.

1.7 Ginamieji teiginiai

1. Pasiūlyti du nauji požymių išskyrimo algoritmai (HBFE ir HSCA), besiremiantys priklausomumo struktūros optimizavimu.

2. Eksperimentai su binarinio klasifikavimo imtimis rodo, jog pasiūlyti algoritmai kai kuriais atvejais yra efektyvesni nei tiesinė diskriminantinė analizė ar pagrindinių komponenčių analizė.

3. HSIC įvertinio paslinktumas yra esminis veiksnys, įtakojantis pasiūlytų algoritmų efektyvumą. Eksperimentai su daugiažymio klasifikavimo imtimis rodo, jog HBFE su nepaslinktuoju įvertiniu yra efektyvesnis.

4. Eksperimentai su binarinio klasifikavimo imtimis rodo, jog HSCA taikomas požymių tarpusavio HSIC-priklausomumo minimizavimas gerina jų kokybę.

5. Eksperimentai su dalinai prižiūrimo mokymo imtimis rodo, jog kai kuriais atvejais Laplaso reguliarizacija gerina HSCA bei HBFE požymių efektyvumą.

2 Rezultatai

Tarkime $(\mathcal{X}, \mathcal{F}_X, \mathcal{P}_X)$ ir $(\mathcal{Y}, \mathcal{F}_Y, \mathcal{P}_Y)$ yra dvi tikimybinės erdvės, o X ir Y - jose nusakyti atsitiktiniai vektoriai. Turėdami X ir Y kovariacijos matricą Σ galime nagrinėti tiesines X ir Y sąsajas, bet jeigu šių dydžių ryšys yra netiesinis, kovariacijos jo aprašyti negali. Hilberto-Schmidto priklausomumo matas yra Σ Frobenijaus normos $||\Sigma||_F^2 = \sum_{ij} \Sigma_{ij}^2$ apibendrinimas reprodukuoto branduolio Hilberto erdvėse (RKHS[8]).

Apibrėžimas 1. *Simetrinė funkcija $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ vadinama teigiamai apibrėžtu branduoliu[9], jeigu $\forall x_1, x_2, ..., x_n \in \mathcal{X}$, $\forall \boldsymbol{c} \in \mathbb{R}^n$, teisinga nelygybė $\boldsymbol{c}^T \boldsymbol{K} \boldsymbol{c} \geq 0$, kur $\boldsymbol{K}$ ($\boldsymbol{K}_{ij} = k(x_i, x_j)$) yra Gramo (branduolių) matrica.*

Įrodoma [1], jog bet kuris teigiamai apibrėžtas branduolys k vienareikšmiškai nusako specialios struktūros Hilberto erdvę $\mathcal{H}$ ir apibrėžia skaliarinę sandugą joje:

$$k(x_i, x_j) =< \phi(x_i), \phi(x_j) >_{\mathcal{H}},$$

kur $\phi \in \mathcal{H}$. Tokia erdvė vadinama reprodukuotojo branduolio Hilberto erdve.

Šis faktas leidžia bet kurį tiesinį algoritmą, priklausantį tik nuo skaliarinių sandaugų išplėsti į RKHS. Tokiu būdu ne tik netiesiniams, bet ir struktūrizuotiems duomenims (pvz. kai stebėjimus sudaro grafai, tekstai ar paveikslėliai) pakanka nusakyti teigiamai apibrėžtą branduolį ir automatiškai gaunami žinomų ir gerai išnagrinėtų algoritmų variantai, tinkantys nagrinėjamai situacijai.

HSIC gaunamas apibendrinant Frobenijaus normą tiesiniams operatoriams [7].

Apibrėžimas 2. *Tegul $\mathcal{F}$ ir $\mathcal{G}$ yra RKHS, kuriose galima rasti ortogonalias bazes $(u_i)_{i\geq 1}$ and $(v_j)_{j\geq 1}$. Tarkime $C : \mathcal{G} \to \mathcal{F}$ yra tiesinis operatorius. Tada, jeigu suma konverguoja, C Hilberto-Schmidto norma nusakoma lygybe $||C||_{HS}^2 := \sum_{i,j} < Cv_i, u_j >_{\mathcal{F}}^2$.*

Tarkime, k ir l yra $\mathcal{F}$ ir $\mathcal{G}$ nusakantys branduoliai, $\phi \in \mathcal{F}$, $\psi \in \mathcal{G}$ - atitinkami skaliarines sandaugas realizuojantys atvaizdžiai, o $C_{xy} = \mathbb{E}_{xy}(\phi(x) - \mathbb{E}_x\phi(x)) \otimes (\psi(y) - \mathbb{E}_y\psi(y))$ -

[8] reproducing kernel Hilbert spaces, RKHS.
[9] toliau vadinsime tiesiog branduoliu.

kovariacinis operatorius[10].

$$HSIC(\mathcal{F}, \mathcal{G}, P_{xy}) := ||C_{xy}||^2_{HS} = \sum_{i,j} < C_{xy} v_i, u_j >^2_{\mathcal{F}} . \quad (1)$$

Tarkime, turime prižiūrimojo mokymo imtį $T = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^m$, kur $\mathbf{x}_i \in \mathbb{R}^{D_x}$ yra įvesties kintamojo stebėjimai, o $\mathbf{y} \in \mathbb{R}^{D_y}$ - priklausojo kintamojo stebėjimai. Pažymėkime duomenų matricas $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_m]$, $\mathbf{Y} = [\mathbf{y}_1, \mathbf{y}_2, ..., \mathbf{y}_m]$, o išskirtų požymių matricą $\mathbf{F} = [\mathbf{f}_1, \mathbf{f}_2, ..., \mathbf{f}_m]$. Pažymėkime $\mathbf{K}$ ir $\mathbf{L}$ atitinkamas Gramo matricas.

Pasiūlyti du HSIC įvertiniai[11] ([7], [10]):

$$HSIC_0(\mathbf{X}, \mathbf{Y}) := (m-1)^{-2} Tr(\mathbf{KHLH}), \quad (2)$$

ir

$$HSIC_1(\mathbf{X}, \mathbf{Y}) := \frac{1}{m(m-3)} \left(Tr\widetilde{\mathbf{K}}\widetilde{\mathbf{L}} + \frac{\mathbf{1}^T\widetilde{\mathbf{K}}\mathbf{1}\mathbf{1}^T\widetilde{\mathbf{L}}\mathbf{1}}{(m-1)(m-2)} - \frac{2}{m-2}\mathbf{1}^T\widetilde{\mathbf{K}}\widetilde{\mathbf{L}}\mathbf{1}) \right), \quad (3)$$

kur $\widetilde{\mathbf{K}}$ ir $\widetilde{\mathbf{L}}$ yra atitinkamos branduolių matricos su nulinėmis diagonalėmis, $\mathbf{A} = \mathbf{1}\mathbf{1}^T$, $\mathbf{H} = \mathbf{I} - m^{-1}\mathbf{A}$. Įrodoma, jog įvertinys (2) yra paslinktasis, o (3) - nepaslinktasis. Taip apibrėžtas HSIC gali detektuoti bet kokią statistinę priklausomybę (t.y., jei tik branduoliai nepašalina informacijos apie aukštesniųjų eilių priklausomybę, HSIC tampa 0 tada ir tik tada kai X ir Y yra nepriklausomi).

2.1 Požymių išskyrimas maksimizuojant HSIC įvertinius

HBFE ieškoma projekcijų matrica $\widetilde{\mathbf{P}}$, kuri maksimizuoja $\widehat{HSIC}(\mathbf{P}^T\mathbf{X}, \mathbf{Y})$. Įvesties stebėjimams fiksuokime branduolį k, o išvesties stebėjimams - l. Tarkime $\mathbf{K}$ ir $\mathbf{L}$ yra atitinkamos Gramo matricos. Disertacijoje parodome, jog priklausomai nuo HSIC įvertinio paslinktumo $\widetilde{\mathbf{P}}$ sudaryta iš Lent. 1 nusakyto tikrinių vektorių uždavinio sprendinių,

[10] $\otimes$ žymime tenzorinę sandaugą.

[11] Toliau, jei paslinktumas nebus esminis aspektas, paprastumo dėlei įvertinius žymėsime $\widehat{HSIC}$.

	Paslinktasis	Nepaslinktasis
Tiesinis	$\mathbf{XHLHX}^T\mathbf{p} = \lambda\mathbf{p}$	$\mathbf{X}(\widetilde{\mathbf{L}} + \frac{\mathbf{A}\widetilde{\mathbf{L}}\mathbf{A}-\mathbf{1}^T\widetilde{\mathbf{L}}\mathbf{1}\mathbf{I}}{(m-1)(m-2)} - \frac{\widetilde{\mathbf{L}}\mathbf{A}+\mathbf{A}\widetilde{\mathbf{L}}-2diag(\widetilde{\mathbf{L}}\mathbf{A})}{m-2})\mathbf{X}^T\mathbf{p} = \lambda\mathbf{p}$
Netiesinis	$\mathbf{KHLHKp}^T = \lambda\mathbf{Kp}$	$\mathbf{K}(\widetilde{\mathbf{L}} + \frac{\mathbf{A}\widetilde{\mathbf{L}}\mathbf{A}-\mathbf{1}^T\widetilde{\mathbf{L}}\mathbf{1}\mathbf{I}}{(m-1)(m-2)} - \frac{\widetilde{\mathbf{L}}\mathbf{A}+\mathbf{A}\widetilde{\mathbf{L}}-2diag(\widetilde{\mathbf{L}}\mathbf{A})}{m-2})\mathbf{Kp} = \lambda\mathbf{Kp}$

1 lentelė: HBFE sprendiniai

atitinkančių didžiausias tikrines reikšmes. Tiesiniu atveju naujo stebėjimo $\mathbf{x}$ požymiai $\mathbf{f} := \mathbf{f(x)}$ gaunami tiesiškai projektuojant $\mathbf{f} := \widetilde{\mathbf{P}}^T\mathbf{x}$, o bendruoju - $\mathbf{f} := \widetilde{\mathbf{P}}^T\mathbf{k}_\mathbf{x}$, kur $\mathbf{k}_\mathbf{x} := [k(\mathbf{x}, \mathbf{x}_1), ..., k(\mathbf{x}, \mathbf{x}_m)]^T$.

Disertacijoje taip pat pasiūlytas alternatyvus netiesiškumų modeliavimas taikant dirbtinius neuroninius tinklus (NeuroHBFE ir NeuroHSCA metodai). NeuroHBFE ieškomi daugiasluoksnio perceptrono parametrai, maksimizuojantys HSIC įvertinį tarp tinklo atsako ir priklausomo kintamojo. NeuroHSCA papildomai minimizuoja požymių tarpusavo priklausomumą.

2.2 Požymių išskyrimas papildomai minimizuojant požymių tarpusavio priklausomumą

Tarkime turime prižiūrimojo mokymo imtį $T = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^m$, kur $\mathbf{x}_i \in \mathbb{R}^{D_x}$ yra įvesties stebėjimai, o $\mathbf{y} \in \mathbb{R}^{D_y}$ - priklausomi stebėjimai. Be to, tarkime jog jau turime išskyrę vieną požymį f^1. Intuityviai aišku, kad net jei naujojo požymio f^2 ir priklausomojo kintamojo priklausomybė didelė, iš jo nedaug naudos jei f^2 labai priklauso nuo f^1. Todėl HSCA papildomai minimizuojamas HSIC-priklausomumas tarp naujojo ir jau išskirtųjų požymių. t-ojoje algoritmo iteracijoje maksimizuojamas Relėjaus santykis:

$$\eta_t(\mathbf{p}) = \frac{\widehat{HSIC}(\mathbf{p}^T\mathbf{X}, \mathbf{Y})}{\widehat{HSIC}(\mathbf{p}^T\mathbf{X}, \mathbf{P}_t^T\mathbf{X})}, \tag{4}$$

kur $\mathbf{P}_t = [\mathbf{p}_1, ..., \mathbf{p}_{t-1}]$. Tarkime $\mu(\mathbf{A}, \mathbf{B})$ yra apibendrintojo tikrinių reikšmių uždavinio $\mathbf{Ap} = \lambda\mathbf{Bp}$ vedantysis tikrinis vektorius. Disertacijoje parodoma, jog tiesinio branduolio atveju tai galima pasiekti vykdant Algoritmą 1. Bendruoju atveju, taikomas Algoritmas 2.

Algoritmas 1 $HSCA(\mathbf{X}, \mathbf{Y}, d_x)$

Įvestis: Duomenų matricos $\mathbf{X}, \mathbf{Y}$, dimensija d_x.
Išvestis: Projekcijos matrica $\mathbf{P}$.

1. Užpildyti $m \times m$ branduolių matricą $\mathbf{L} : \mathbf{L}_{i,j} = l(\mathbf{y}_i, \mathbf{y}_j)$, priskirti $\mathbf{M} = \mathbf{I}$, ir $\mathbf{P} = \emptyset$.
2. **for** $t \in 1, .., d_x$ **do**
3. Papildyti $\mathbf{P} := [\mathbf{P}, \mu(\mathbf{XHLHX}^T, \mathbf{M})]$
4. Užpildyti $m \times m$ jau išskirtų požymių branduolių matricą $\mathbf{L}_f(i,j) = l_f(\mathbf{P}^T\mathbf{x}_i, \mathbf{P}^T\mathbf{x}_j)$.
5. $\mathbf{M} := \mathbf{XHL}_f\mathbf{HX}^T$.
6. **end**
7. Grąžinti projekcijos matricą $\mathbf{P}$. Įėjimo $\mathbf{x}$ požymiai nusakomi $\mathbf{P}^T\mathbf{x}$.

Algoritmas 2 $KHSCA(\mathbf{X}, \mathbf{Y}, d_x)$

Įvestis: Duomenų matricos $\mathbf{X}, \mathbf{Y}$, dimensija d_x.
Išvestis: Projekcijos matrica $\mathbf{P}$.

1. Užpildyti $m \times m$ branduolių matricas $\mathbf{K} : K_{i,j} = k(\mathbf{x}_i, \mathbf{x}_j)$, $\mathbf{L} : \mathbf{L}_{i,j} = l(\mathbf{y}_i, \mathbf{y}_j)$, set $\mathbf{M} = \mathbf{K}$, ir $\mathbf{P} = \emptyset$.
2. **for** $t \in 1, .., d_x$ **do**
3. Papildyti $\mathbf{P} := [\mathbf{P}, \mu(\mathbf{KHLHK}, \mathbf{M})]$
4. Sukonstruoti $m \times m$ jau išskirtų požymių branduolių matricą $\mathbf{L}^f_{i,j} = l_f(\mathbf{P}^T\mathbf{x}_i, \mathbf{P}^T\mathbf{x}_j)$.
5. $\mathbf{M} := \mathbf{KHL}^f\mathbf{HK}$.
6. **end**
7. Grąžinti projekcijos matricą $\mathbf{P}$. Įėjimo $\mathbf{x}$ požymiai nusakomi $\mathbf{P}^T[k(\mathbf{x}, \mathbf{x}_1), k(\mathbf{x}, \mathbf{x}_2), ..., k(\mathbf{x}, \mathbf{x}_m)]^T$.

2.3 Dalinai prižiūrimo mokymo imčių atvejis

Praktikoje sutinkama nemažai atvejų, kai priklausomo kintamojo stebėjimo kaštai smarkiai viršija įeities kintamojo stebėjimo kaštus. Tokių atvejų pavyzdžiais galėtų būti įvairūs medicinos, inžinerijos, finansų uždaviniai. Tarkime, turime dalinai prižiūrimo mokymo imtį susidedančią iš $T = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^m$ ir papildomos neprižiūrimojo mokymo imties $\mathbf{x}_{m+1}, ..., \mathbf{x}_{m+M}$, kur M paprastai viršija m keletą ar daugiau kartų. Disertacijoje HBFE ir HSCA algoritmai pritaikomi dalinai prižiūrimo mokymo imtims į tikslo funkciją įkomponuojant Laplaso reguliarizavimo narį $\Omega(\mathbf{F}) = \frac{1}{2}\sum_{i,j=1}^m ||\frac{\mathbf{f}_i}{\sqrt{\mathbf{D}_{i,i}}} - \frac{\mathbf{f}_j}{\sqrt{\mathbf{D}_{j,j}}}||^2 \mathbf{W}_{i,j}$ [3], "baudžiantį" jei arti esančias įėjimų poras atitinka nutolę požymiai. Pažymėkime $\widetilde{\mathbf{X}}$ išplėstąją stebėjimų matricą, susidedančią iš visų $\mathbf{x}_i$, $i \in \{1, 2, ..., m + M\}$.

2.3.1 HBFE atvejis

Dalinai prižiūrimo mokymo imties atveju maksimizuojame tikslo funkciją

$$\widehat{HSIC}(\mathbf{P}^T\mathbf{X}, \mathbf{Y}) - \beta\Omega(\mathbf{P}^T\widetilde{\mathbf{X}}), s.t. \mathbf{P}^T\mathbf{P} = \mathbf{I}, \tag{5}$$

kur $0 \geq \beta \geq 1$ Laplaso reguliarizacijos daugiklis.

2.3.2 HSCA atvejis

Dalinai prižiūrimo mokymo imties atveju kiekvienoje HSCA iteracijoje maksimizuojame

$$\eta_t(\mathbf{p}) = \frac{(1-\beta)\widehat{HSIC}(\mathbf{p}^T\mathbf{X}, \mathbf{Y}) - \beta\Omega(\mathbf{p}^T\widetilde{\mathbf{X}})}{\widehat{HSIC}(\mathbf{p}^T\mathbf{x}, \mathbf{P}_t^T\mathbf{X})}, \tag{6}$$

kur $0 \geq \beta \geq 1$ Laplaso reguliarizacijos daugiklis.

2.4 Eksperimentai

Eksperimentinėje disertacijos dalyje nagrinėjami tokio tipo požymiai: nemodifikuoti duomenys (žym. *Full*), HBFE su nepaslinktu ir paslinktu HSIC įvertiniu (žym. $HBFE_1$ ir

$HBFE_0$), HSCA su nepaslinktu ir paslinktu HSIC įvertiniu (žym. $HSCA_1$ ir $HSCA_0$), pagrindinės komponentės (žym. PCA) ir diskriminantinė analizė (žym. LDA). Taikoma toliau aprašyta schema.

1. Duomenų imtis atsitiktinai suskaidoma į apmokymo bei testavimo imtis, kurios normalizuojamos atimant vidurkį bei dalijant iš vidutinio kvadratinio nuokrypio.
2. Naudojant apmokymo imtį kiekvienam algoritmui įvertinama projekcijos matrica. Dalinai prižiūrimo mokymo scenarijaus atveju laikoma jog tik pirmiems 30% apmokymo imties įėjimo stebėjimams priklausomas kintamasis yra žinomas.
3. Naudojant projekcijos matricos įvertį randamos apmokymo bei testavimo imties stebėjimų projekcijos (požymiai).
4. Naudojant k artimiausių kaimynų klasifikatorių ($k \in \{1, 3, 5\}$) testavimo imties požymiai klasifikuojami, klasifikatoriaus efektyvumo matu laikant jo tikslumą.

Ši schema taikoma T kartų, kiekvieną kartą požymių dimensija optimizuojama pagal klasifikavimo tikslumą taikant 3 suskaidymų kryžminį validavimą. Siekiant įvertinti rezultatų statistinį patikimumą taikomas Wilcoxon'o ženklų kriterijus. Jeigu kuris nors iš nagrinėtų požymių tipų buvo efektyvesnis už visus likusius su p-reikšme neviršijančia 0.05, jis išskiriamsis pabrauktu šriftu. Papildomai palygintas HBFE ir HSCA su paslinktu ir nepaslinktu HSIC įvertiniu rezultatų statistinis reikšmingumas, efektyvesnįjį įvertinį atitinkantį rezultatą pažymint paryškintu šriftu. Duomenų imtys pažymimos simboliu $\diamond$, jei efektyviausias iš pasiūlytųjų algoritmų (t.y. $HBFE_0$, $HBFE_1$, $HSCA_0$, $HSCA_1$) buvo tikslesnis už efektyviausią iš likusiųjų algoritmų, ir simboliu $\star$ - jei atsitiko priešingai. Analogiškai, jei HSCA buvo efektyvesnis už HBFE, imtis pažymima $\bullet$, o priešingi atvejai pažymimi $\circ$.

2.4.1 Binarinio klasifikavimo duomenys

Taikant aukščiau aprašytą schemą su $T = 50$ nagrinėjamos binarinio klasifikavimo duomenų imtys (žr. Lent. 2.4.1).

1. Palyginami tiesiniai algoritmų variantai (žr. Lent. 3).

2. Palyginami netiesiniai algoritmų variantai su Gauso branduoliu $k(\mathbf{x}, \mathbf{x}') = exp(-||\mathbf{x}_i - \mathbf{x}_j||^2/2\sigma^2)$, kur parametras σ parenkamas taikant 3 suskaidymų kryžminį validavimą (žr. Lent. 4).

3. Palyginami tiesiniai dalinai prižiūrimo mokymo algoritmų variantai, laikant kad 30% mokymo duomenų sužymėti (žr. Lent. 5). Laplaso reguliarizacijai naudotas Gauso branduolys, parametrą σ parenkant su 3 suskaidymų kryžminiu validavimu.

2 lentelė: Klasifikavimo duomenų aprašas.
m - imties dydis, D_x - dimensija, p^+, and p^- - klasių pasiskirstymas.

Data set	m	D_x	p^+	p^-
Ames	2495	377	0.3523	0.6477
Australian	690	14	0.55507	0.44493
Breast cancer	683	10	0.34993	0.65007
Covertype	581	54	0.45783	0.54217
Derm	358	34	0.31006	0.68994
German	1000	24	0.7	0.3
Heart	270	13	0.55556	0.44444
Ionosphere	351	34	0.64103	0.35897
Sonar	208	60	0.46635	0.53365
Spambase	4601	57	0.39404	0.60596
Specft	80	44	0.5	0.5
Wdbc	569	30	0.62742	0.37258

Iš Lent. 3, Lent. 4 ir Lent. 5 matyti, jog požymių tarpusavio priklausomumą minimizuojantis HSCA efektyvesnis nei HBFE.

Eksperimentai su Gauso branduoliu (žr. Lent. 4) rodo, jog pasiūlytieji HBFE ir HSCA algoritmai geba išnaudoti duomenyse glūdinčius netiesiškumus, bet kadangi netiesiškumų gali ir nebūti, ar Gauso branduolys jiems modeliuoti gali būti neefektyvus, matome, jog tik keletu atvejų tikslumas ženkliai pagerėjo. Be to Gauso branduolio atveju tenka įvertinti daugiau parametrų - tai leidžia stipriau pasireikšti permokymo efektams [12].

Lent. 5 pateikti rezultatai dalinai prižiūrimojo tipo imtims. Statistiškai reikšmingi atvejai, kai Laplaso reguliarizacija pakeitė požymių efektyvumą lyginant su atveju kai ji nebuvo taikoma, pažymėti viršuje procentais nurodant tikslumo pokytį. Rezultatai rodo, jog HBFE ir HSCA algoritmams Laplaso reguliarizacija tendencingai buvo efektyvi, bet

[12] angl. overfitting.

kartu pasitaikė ir atvejų kai ji paveikė požymių kokybę neigiamai. Taip galėjo nutikti dėl to, kad Laplaso reguliarizacijos prielaida konkrečiu atveju nebuvo teisinga.

2.4.2 Daugiažymiai duomenys

Daugiažymio klasifikavimo uždaviniuose stebėjimas susiejamas su keletu klasių. Traktuoti kiekvieną šių klasių kombinaciją kaip atskirą klasę bei taikyti įprastinius klasifikavimo metodus tampa neįmanoma dėl kombinatorinio sprogimo. Daugiažymio klasifikavimo klasifikatoriaus efektyvumui nagrinėti taikysime matus: Hammingo atstumas, vieneto klaida, padengimas, rangavimo klaida, vidutinis tikslumas [12]. Klasifikavimas atliekamas su modifikuotu k artimiausių kaimynų klasifikatoriumi [12], naudojant tą pačią eksperimentų schemą kaip ir anksčiau (suskaidymų skaičius $T = 25$).

HSIC patogus tuo, jog kaip branduoliais besiremiantis matas yra taikytinas bet kokios struktūros duomenims. Eksperimentų rezultatai Lent. 6[13] rodo, jog HBFE yra efektyvesnis su nepaslinktu HSIC įvertiniu. [13] gauti rezultatai, jog HBFE su paslinktu HSIC įvertiniu efektyvumu lenkia MRSI [11] ir LPP [8] algoritmus. Eksperimentai su netiesiniais branduoliais ir HSCA praleisti dėl didelės skaičiavimų apimties.

2.4.3 Struktūrizuoti duomenys

Taikydami tą pačią atsitiktinių suskaidymų schemą su $T = 50$ nagrinėsime UCI [2] *promoters* klasifikavimo duomenis, sudarytus iš genetinių sekų atkarpų, kurias reikia priskirti vienai iš dviejų vienodai galimų klasių. Natūralu įtarti jog esminė informacija glūdi vidinėje genetinės sekos struktūroje. Šią prielaidą patvirtina Lent. 7, kurioje lyginami tiesinio, Gauso ir eilučių[14] [9] branduolių efektyvumas. Rezultatai rodo, jog eilučių branduolys, atsižvelgiantis vidinę įėjimo vektoriaus struktūrą, leido gerokai pagerinti klasifikavimo tikslumą lyginant su Gauso ir tiesiniu branduoliais.

[13] $\downarrow$ žymi, jog mažesnė mato reikšmė reiškia geresnį klasifikavimą, o $\uparrow$ atitinka priešingą atvejį.
[14] angl. string.

3 lentelė: Klasifikavimo tikslumas tiesiniu atveju.

Duomenys	*Full*	$HBFE_1$	$HBFE_0$	$HSCA_1$	$HSCA_0$	*PCA*	*LDA*
1-AK klasifikatorius							
Ames • ◇	0.7753	0.7589	**0.7765**	0.7826	**0.8012**	0.7786	0.7714
Australian	0.7933	0.7987	0.8045	0.8093	0.8095	0.7868	0.8114
Breastcancer	0.9558	0.9553	0.9543	0.9553	0.9595	0.9566	0.9562
Covertype ◇	0.6868	**0.6956**	0.6732	0.7086	0.7014	0.6748	0.6756
Derm	0.9906	0.9973	0.9973	0.9973	0.9971	0.9949	0.9971
German	0.6698	**0.6841**	0.6730	0.6833	0.6910	0.6700	0.6851
Heart	0.7618	0.7600	0.7677	0.7612	0.7627	0.7520	0.7698
Ionosphere • ◇	0.8421	0.8555	**0.8683**	0.8686	0.8773	0.8581	0.8171
Sonar	0.8146	**0.7819**	0.7538	0.7427	**0.8046**	0.8146	0.6938
Spambase •	0.8975	0.8993	0.8979	0.9116	0.9056	0.9015	0.8680
Specft	0.6770	0.6690	0.7030	0.6730	0.7020	0.6630	0.5370
Wdbc •	0.9503	0.9355	**0.9455**	0.9476	**0.9570**	0.9506	0.9528
3-AK klasifikatorius							
Ames • ◇	0.7872	0.7753	**0.7852**	0.7897	**0.8158**	0.7878	0.7853
Australian •	0.8328	0.8329	0.8314	0.8431	0.8423	0.8227	0.8393
Breastcancer	0.9618	0.9642	0.9656	0.9661	0.9674	0.9650	0.9660
Covertype	0.6796	0.6720	**0.6956**	0.7046	0.7063	0.6672	0.7032
Derm	0.9926	0.9973	0.9973	0.9973	0.9973	0.9937	0.9980
German	0.7036	0.7131	0.7078	0.7181	0.7208	0.7031	0.7169
Heart	0.8003	0.7828	0.7959	0.7905	0.7988	0.7876	0.7938
Ionosphere • ◇	0.8313	0.8542	**0.8656**	0.8715	0.8763	0.8615	0.8384
Sonar •	0.7908	0.7450	0.7373	0.7542	**0.7996**	0.7962	0.6923
Spambase • ◇	0.8993	0.9087	0.9051	0.9158	0.9128	0.9065	0.8930
Specft ◇	0.6810	0.7350	0.7550	0.6780	0.7130	0.6780	0.5370
Wdbc •	0.9597	0.9482	0.9528	0.9539	**0.9628**	0.9577	0.9575
5-AK klasifikatorius							
Ames • ◇	0.7916	**0.7882**	0.7819	0.7961	**0.8199**	0.7900	0.7904
Australian	0.8445	0.8421	0.8463	0.8499	0.8500	0.8322	0.8529
Breastcancer	0.9641	0.9666	0.9646	0.9668	0.9663	0.9653	0.9665
Covertype	0.6661	0.6843	0.7007	**0.7157**	0.7029	0.6561	0.7113
Derm	0.9911	0.9971	0.9973	0.9971	0.9971	0.9951	0.9984
German •	0.7113	0.7262	0.7209	0.7288	0.7343	0.7127	0.7302
Heart	0.8133	0.8033	0.8068	0.8145	0.8136	0.8062	0.8065
Ionosphere • ◇	0.8279	0.8571	0.8610	0.8622	**0.8763**	0.8539	0.8398
Sonar •	0.7512	0.7108	**0.7369**	0.7488	**0.7812**	0.7631	0.6912
Spambase • ◇	0.8970	0.9057	0.9086	**0.9186**	0.9128	0.9083	0.9003
Specft ◇	0.6680	0.7340	0.7470	0.6800	**0.7240**	0.6860	0.5370
Wdbc • ◇	0.9593	0.9485	0.9528	0.9523	**0.9675**	0.9577	0.9570

4 lentelė: Klasifikavimo tikslumas taikant Gauso branduolį.

Duomenys	*Full*	$HBFE_1$	$HBFE_0$	$HSCA_1$	$HSCA_0$	*PCA*	*LDA*
1-AK klasifikatorius							
Australian •	0.7924	0.7900	0.7927	0.7878	**0.8185**	0.7807	0.8110
Breastcancer	0.9508	0.9505	0.9458	0.9472	0.9466	0.9522	0.9487
Covertype • ⋆	0.6932	0.6480	**0.6738**	0.6855	0.6860	0.6777	0.7091
Derm	0.9872	0.9985	0.9989	0.9993	0.9989	0.9935	0.9984
German • ◇	0.6717	0.6668	0.6784	**0.6956**	0.6797	0.6737	0.6802
Heart	0.7638	0.7689	0.7679	0.7575	0.7669	0.7588	0.7366
Ionosphere	0.8521	0.8895	**0.9128**	0.9025	0.9158	0.9083	0.8927
Sonar •	0.8358	0.6955	**0.7942**	0.7204	**0.8344**	0.8387	0.8258
Spambase • ⋆	0.8570	0.7994	0.7903	0.8119	0.8129	0.8471	0.8779
Specft	0.6778	0.7283	0.7317	0.7650	0.7400	0.6370	0.7370
Wdbc ⋆	0.9510	0.9148	**0.9425**	0.9320	0.9424	0.9510	0.9599
3-AK klasifikatorius							
Australian • ◇	0.8362	0.8344	0.8311	0.8233	**0.8491**	0.8193	0.8275
Breastcancer	0.9607	0.9595	0.9548	0.9554	0.9587	0.9617	0.9558
Covertype • ⋆	0.6874	0.6522	0.6731	**0.6989**	0.6752	0.6784	0.7241
Derm	0.9882	0.9993	0.9993	0.9996	0.9996	0.9960	0.9981
German • ◇	0.7006	0.6987	**0.7111**	0.7269	0.7120	0.6913	0.6880
Heart	0.8082	0.8104	0.8257	0.8133	0.8188	0.8012	0.7436
Ionosphere ◇	0.8314	0.8994	**0.9261**	0.9204	0.9230	0.9013	0.8971
Sonar •	0.8100	0.7256	**0.7814**	0.7312	**0.8093**	0.8186	0.8244
Spambase ⋆	0.8630	0.8189	0.8124	0.8276	0.8266	0.8497	0.8883
Specft • ◇	0.6722	0.7450	0.7400	**0.7867**	0.7367	0.6648	0.7370
Wdbc ⋆	0.9596	0.9200	**0.9437**	0.9401	0.9465	0.9560	0.9620
5-AK klasifikatorius							
Australian	0.8514	0.8460	0.8375	0.8399	**0.8567**	0.8406	0.8351
Breastcancer	0.9643	0.9632	0.9597	0.9543	0.9586	0.9646	0.9565
Covertype • ⋆	0.6748	0.6703	0.6699	**0.7048**	0.6768	0.6659	0.7315
Derm	0.9850	0.9993	0.9993	0.9996	0.9996	0.9966	0.9981
German ◇	0.7123	0.7159	0.7251	0.7360	0.7285	0.7004	0.6948
Heart	0.8169	0.8321	0.8296	0.8212	0.8281	0.8111	0.7403
Ionosphere	0.8216	0.9078	0.9135	0.9204	0.9223	0.9098	0.9003
Sonar ⋆	0.7878	0.7263	**0.7821**	0.7369	**0.7928**	0.7935	0.8215
Spambase • ⋆	0.8637	0.8223	0.8141	0.8374	0.8355	0.8528	0.8953
Specft • ◇	0.6833	0.7400	0.7500	**0.7900**	0.7400	0.6926	0.7370
Wdbc ⋆	0.9589	0.9242	**0.9488**	0.9380	0.9481	0.9573	0.9682

5 lentelė: Klasifikavimo tikslumas dalinai prižiūrimo mokymo atveju.

Duomenys	*Full*	$HBFE_1$	$HBFE_0$	$HSCA_1$	$HSCA_0$	*LDA*
1-AK klasifikatorius						
Australian	0.779	0.791	0.787	0.803	$0.801^{-1.43}$	0.810
Breastcancer	0.950	0.953	0.951	0.951	0.953	0.950
Covertype	0.629	$0.620^{0.52}$	0.641	0.645	0.646	$0.650^{-3.40}$
German	0.664	0.673	$0.660^{1.03}$	$0.685^{-0.45}$	0.679	0.680
Heart	0.764	0.755	0.759	0.756	0.765	0.749
Ionosphere	0.802	$0.772^{3.83}$	$0.840^{-1.79}$	$0.791^{2.21}$	0.823	0.774
Sonar	0.720	0.656	0.689	0.674	0.709	0.643
Wdbc	0.938	$0.916^{0.18}$	0.929	0.924	0.948	$0.919^{2.01}$
3-AK klasifikatorius						
Australian	0.819	0.831	0.823	0.839	$0.840^{-1.27}$	0.838
Breastcancer	0.958	0.962	0.962	0.961	0.961	0.958
Covertype	0.624	$0.626^{0.48}$	$0.650^{0.18}$	0.657	0.654	$0.668^{-2.81}$
German	0.689	$0.696^{0.31}$	$0.688^{0.99}$	0.713	0.707	0.707
Heart	0.791	0.787	0.791	$0.783^{0.83}$	$0.793^{-1.30}$	$0.758^{3.14}$
Ionosphere	0.768	$0.784^{1.80}$	$0.832^{-1.25}$	$0.801^{2.14}$	0.812	0.774
Sonar	0.674	0.671	0.685	0.682	0.704	0.643
Specft	0.620	0.660	$0.653^{1.61}$	0.636	0.640	$0.639^{-2.50}$
Wdbc	0.941	0.927	0.937	0.930	0.951	$0.922^{1.90}$
5-AK klasifikatorius						
Australian	0.835	0.841	$0.835^{0.81}$	0.848	$0.848^{-0.79}$	0.850
Breastcancer	0.959	0.964	0.964	0.964	0.963	0.962
Covertype	0.621	$0.627^{0.76}$	$0.650^{0.96}$	0.662	0.655	$0.668^{-2.19}$
German	0.705	$0.708^{0.64}$	$0.703^{0.95}$	$0.723^{0.62}$	0.718	0.716
Heart	0.801	0.802	$0.801^{0.24}$	$0.794^{1.06}$	$0.802^{-0.8}$	$0.769^{1.60}$
Ionosphere	0.735	$0.787^{2.14}$	$0.818^{-1.54}$	$0.802^{1.2}$	$0.804^{0.77}$	1.774
Sonar	0.660	$0.677^{0.78}$	$0.681^{0.51}$	0.690	0.706	0.643
Specft	0.589	0.665	$0.642^{2.3}$	$0.614^{1.68}$	0.635	$0.637^{-2.55}$
Wdbc	0.941	0.930	0.936	$0.931^{0.21}$	$0.948^{0.18}$	$0.923^{1.99}$

6 lentelė: Suvidurkinti rezultatai Yahoo duomenų imtims.

	$Full$	$HBFE_0$	$HBFE_1$	PCA
Hammingo atstumas ↓	0.0431	<u>0.0426</u>	0.0428	0.0431
Vieneto klaida ↓	0.4716	0.4641	<u>0.4585</u>	0.4729
Padengimas ↓	4.2265	4.2137	<u>4.1604</u>	4.2443
Rangavimo klaida ↓	0.1055	0.1052	<u>0.1030</u>	0.1060
Vidutinis tikslumas ↑	0.6211	0.6260	<u>0.6322</u>	0.6200

7 lentelė: Eksperimentų su struktūrizuotais duomenimis rezultatai

Branduolys	Full	$HSIC_1$	$HSIC_0$	$HSCA_1$	$HSCA_0$	PCA	LDA
1-AK klasifikatorius							
Tiesinis		0.7484	0.7553	0.6755	**0.7264**	0.7138	0.6031
Gauso	0.7264	0.7711	0.7686	0.7774	0.7862	0.7195	0.7736
Eilučių		0.9145	0.9126	0.9145	0.9126	0.8226	0.9176
3-AK klasifikatorius							
Tiesinis		0.7635	0.7566	0.7340	**0.7623**	0.7623	0.6031
Gauso ◇	0.7635	0.7899	0.7792	0.7843	0.7836	0.7371	0.7736
Eilučių		0.9075	0.9082	0.9075	0.9082	0.7421	0.9176
5-AK klasifikatorius							
Tiesinis ◇		0.7799	0.7642	0.7491	0.7616	0.7415	0.6031
Gauso ◇	0.7428	0.7868	0.7818	0.7887	0.7799	0.7365	0.7736
Eilučių		0.9170	0.9182	0.9170	0.9182	0.7459	0.9176

2.5 Išvados

Disertacijoje pasiūlyti HBFE ir HSCA požymių išskyrimo algoritmai, optimizuojantys priklausomumo struktūrą, nusakytą HSIC mato įvertiniais. HBFE maksimizuoja priklausomumą tarp požymių ir priklausomojo kintamojo, o HSCA papildomai minimizuoja požymių tarpusavio priklausomumą. Pateikiamos šių algoritmų formuluotės prižiūrimo bei dalinai prižiūrimo mokymo imtims, pastaruoju atveju taikant Laplaso reguliarizaciją.

Eksperimentais parodyta, jog HSIC įvertinio paslinktumas iš esmės įtakoja gautų požymių efektyvumą sprendžiant įprastinio bei daugiažymio klasifikavimo uždavinius, dalinai prižiūrimo mokymo imčių atvejais Laplaso reguliarizacija taip pat leidžia padidinti HBFE ir HSCA efektyvumą. HSCA algoritmas pasirodė efektyvesnis už HBFE, nors jo įvykdymui reikia atlikti daugiau skaičiavimų. Duomenyse glūdintys netiesiškumai bei vidinėje struktūroje esanti informacija gali būti efektyviai išnaudojama pritaikius tinkamai parinką branduolį. Eksperimentai rodo, jog HBFE ir HSCA kai kurioms binarinio klasifikavimo duomenų imtims yra efektyvesni nei PCA ar LDA.

3 Doktoranto publikacijos disertacijos tema

Toliau pateikiami doktoranto moksliniai straipsniai disertacijos tema Vilniaus universiteto vardu. Pilnas straipsnių sąrašas pateikiamas disertacijoje.

1. Daniušis, P., and Vaitkus, Pr., (2009). Supervised Feature Extraction Using Hilbert-Schmidt Norms. Lecture Notes in Computer Science. Springer, Vol. 5788, pp. 25-33. ISSN: 0302-9743 [ISI proceedings].

2. Daniušis, P., and Vaitkus, Pr., (2009). A feature extraction algorithm based on the Hilbert-Schmidt independence criterion. Siauliai Mathematical Seminar, Vol. 4(12), pp. 35-42. ISSN: 1822-511X.

4 Trumpos žinios apie doktorantą

Povilas Daniušis 2001 m. baigė Druskininkų "Atgimimo" vidurinę mokyklą, 2005 m. Šiaulių universitete įgijo matematikos bakalauro laipsnį (BSc), 2007 m. Vilniaus universitete įgijo matematikos magistro laipsnį (MSc). Nuo 2007m. iki 2011m. Vilniaus universiteto doktorantas. Doktorantūros metu P. Daniušis 4 mėn. stažavosi Makso Planko biokibernetikos institute (Vokietija). Čia atliktais tyrimais paremtas [6] straipsnis gavo *Best Student Paper Award* premiją.

5 Santrauka

Daugelis praktiškai reikšmingų sistemų mokymo uždavinių reikalauja gebėti panaudoti didelio matavimo, struktūrizuotus, netiesinius duomenis. Vaizdų, teksto, socialinių bei verslo ryšių analizė, įvairūs bioinformatikos uždaviniai galėtų būti tokių uždavinių pavyzdžiais. Todėl požymių išskyrimas dažnai yra pirmasis žingsnis, kuriuo pradedama duomenų analizė ir nuo kurio priklauso galutinio rezultato sėkmė.

Šio disertacinio darbo tyrimo objektas yra požymių išskyrimo algoritmai, besiremiantys priklausomumo sąvoka. Darbe nagrinėjamas priklausomumas, nusakytas kovariacinio operatoriaus Hilberto-Šmidto normos (HSIC mato) branduoliniu įvertiniu. Pasiūlyti šiuo įvertiniu besiremiantys HBFE ir HSCA algoritmai leidžia dirbti su bet kokios struktūros duomenimis, bei yra formuluojami tikrinių vektorių terminais (tai leidžia optimizavimui naudoti standartinius paketus), bei taikytini ne tik prižiūrimo, bet ir dalinai prižiūrimo mokymo imtims. Pastaruoju atveju HBFE ir HSCA modifikacijos remiasi Laplaso reguliarizacija.

Eksperimentais su klasifikavimo bei daugiažymio klasifikavimo duomenimis parodyta, jog pasiūlyti algoritmai leidžia pagerinti klasifikavimo efektyvumą lyginant su PCA ar LDA.

6 Summary

In many important real world applications the initial representation of the data is inconvenient, or even prohibitive for further analysis. For example, in image analysis, text analysis and computational genetics high-dimensional, massive, structural, incomplete, and noisy data sets are common. Therefore, feature extraction, or revelation of informative features from the raw data is one of fundamental machine learning problems. Efficient feature extraction helps to understand data and the process that generates it, reduce costs for future measurements and data analysis. The representation of the structured data as a compact set of informative numeric features allows applying well studied machine learning techniques instead of developing new ones..

The dissertation focuses on supervised and semi-supervised feature extraction methods, which optimize the dependence structure of features. The dependence is measured using the kernel estimator of Hilbert-Schmidt norm of covariance operator (HSIC measure). Two dependence structures are investigated: in the first case we seek features which maximize the dependence on the dependent variable, and in the second one, we additionally minimize the mutual dependence of features. Linear and kernel formulations of HBFE and HSCA are provided. Using Laplacian regularization framework we construct semi-supervised variants of HBFE and HSCA.

Suggested algorithms were investigated experimentally using conventional and multi-label classification data sets. The extracted features were classified by k nearest neighbor classifier, and their quality is evaluated by classification performance measures. Experiments show that in certain cases our algorithms are more efficient comparing to PCA or LDA.

Literatūra

[1] Aronszajn, N., 1950. Theory of reproducing kernels. Transactions of the American Mathematical Society, Vol. 68, pp. 337-404.

[2] Asuncion, A. and Newman, D.J., 2007. UCI Machine Learning Repository [http://archive.ics.uci.edu/ml/]. University of California, School of Information and Computer Science.

[3] Belkin, M. and Niyogi, P., 2002. Laplacian eigenmaps and spectral techniques for embedding and clustering. Advances in Neural Information Processing System, pp. 585-591.

[4] Daniušis, P., Vaitkus, Pr., 2009. Supervised feature extraction using Hilbert-Schmidt norms. Lecture Notes In Computer Science, Vol. 5788, pp. 25-33.

[5] Daniušis, P., Vaitkus, Pr., 2009. A feature extraction algorithm based on the Hilbert-Schmidt independence criterion. Šiauliai Mathematical Seminar, Vol. 4(12), pp. 35-42.

[6] Daniušsis P., Janzing D., Mooij J., Zscheischler J., Steudel B., Zhang K., and Schölkopf B., 2010. Inferring deterministic causal relations. Proceedings of UAI2010.

[7] Gretton, A., Bousquet, O., Smola, A., Schölkopf B., 2005. Measuring statistical dependence with Hilbert-Schmidt norms. Proceedings of 16th International Conference on Algorithmic Learning Theory, pp. 63-77.

[8] He, X., and Niyogi, P., 2004. Locality preserving projections. In Advances in Neural Information Processing Systems, Cambridge, MA.

[9] Lodhi, H., Saunders, C., Shawe-Taylor, J., Cristianini, N., and Watkins C., 2002. Text classification using string kernels. Journal of Machine Learning Research, Vol. 2, pp. 419-444.

[10] Song, L., Smola, A., Gretton, A., Borgwardt, K., and Bedo, J., 2007. Supervised feature selection via dependence estimation. In Proc. Intl. Conf. Machine Learning, pp. 823-830.

[11] Yu, K.; Yu, S., and Tresp, V., 2005. Multi-label informed latent semantic indexing. In SIGIR, pp. 258-265.

[12] Zhang, M.L., and Zhou, Z.H., 2007. ML-kNN: A lazy learning approach to multi-label learning. Pattern Recognition, Vol. 40(7), pp. 2038-2048.

[13] Zhang, Y., Zhou, Zhi-Hua., 2008. Multi-label dimensionality reduction via dependence maximization. Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence, pp 1503-1505.