VILNIAUS UNIVERSITETAS

Paulius Danėnas

ATRAMINIŲ VEKTORIŲ MAŠINOMIS GRINDŽIAMI KLASIFIKAVIMO METODAI INTELEKTUALIOJE SPRENDIMŲ PARAMOS SISTEMOJE KREDITO RIZIKOS VERTINIMUI

Daktaro disertacijos santrauka
Fiziniai mokslai, informatika (09P)

Vilnius, 2013

Disertacija rengta 2008-2012 Vilniaus universitete

Mokslinis vadovas:

prof. dr. Gintautas Garšva (Vilniaus universitetas, fiziniai mokslai, informatika - 09P)

Konsultantas:

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

Disertacija ginama Vilniaus universiteto Informatikos mokslo krypties taryboje:

Pirmininkas:
prof. dr. Romas BARONAS (Vilniaus universitetas, fiziniai mokslai, informatika – 09P)
Nariai:
prof. habil. dr. Gintautas DZEMYDA (Vilniaus universitetas, fiziniai mokslai, informatika — 09P)
prof. habil. dr. Genadijus KULVIETIS (Vilniaus Gedimino technikos universitetas, fiziniai mokslai, informatika — 09P)
prof. dr. Eduardas BAREIŠA (Kauno technologijos universitetas, fiziniai mokslai, informatika – 09P)
prof. dr. Rimantas BUTLERIS (Kauno technologijos universitetas, fiziniai mokslai, informatika – 09P)

Oponentai:
prof. habil. dr. Feliksas IVANAUSKAS (Vilniaus universitetas, fiziniai mokslai, informatika – 09P)
prof. habil. dr. Aleksandras Vytautas RUTKAUSKAS (Vilniaus Gedimino technikos universitetas, socialiniai mokslai, ekonomika – 04S)

Disertacija bus ginama viešame Informatikos mokslo krypties tarybos posėdyje 2013 m. birželio mėn. 17 d. 10 val. Kauno humanitarinio fakulteto 10 auditorijoje.

Adresas: Muitinės g. 12, Kaunas LT-44280, Lietuva.

Disertacijos santrauka išsiuntinėta 2013 m. gegužės mėn.17 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje.

VILNIUS UNIVERSITY

Paulius Danėnas

SUPPORT VECTOR MACHINES BASED CLASSIFIERS IN INTELLIGENT DECISION SUPPORT SYSTEM FOR CREDIT RISK EVALUATION

Summary of PhD thesis
Physical sciences, Informatics (09P)

Vilnius, 2013

Thesis was prepared during 2008-2012 in Vilnius University

Thesis supervisor:

prof. dr. Gintautas Garšva (Vilnius University, Physical Sciences, Informatics - 09P)

Consultant:

prof. habil. dr. Rimvydas Simutis (Kaunas University of Technology, Physical Sciences, Informatics - 09P)

Thesis is defended in Vilnius University at the Vilnius University Board of Informatics science:

Chair:
prof. dr. Romas BARONAS (Vilnius University, Physical Sciences, Informatics – 09P)
Mambers:
prof. habil. dr. Gintautas DZEMYDA (Vilnius University, Physical Sciences, Informatics — 09P)
prof. habil. dr. Genadijus KULVIETIS (Vilnius Gediminas Technical University, Physical Sciences, Informatics — 09P)
prof. dr. Eduardas BAREIŠA (Kaunas University of Technology, Physical Sciences, Informatics – 09P)
prof. dr. Rimantas BUTLERIS (Kaunas University of Technology, Physical Sciences, Informatics – 09P)

Opponents:
prof. habil. dr. Feliksas IVANAUSKAS (Vilnius University, Physical Sciences, Informatics – 09P)
prof. habil. dr. Aleksandras Vytautas RUTKAUSKAS (Vilnius Gediminas Technical University, Social Sciences, Economics – 04S)

Thesis will be defended at June 17, 2013 in Kaunas Faculty of Humanities, at 10th auditorium.

Address: Muitinės St. 12, Kaunas LT-44280, Lithuania.

Thesis summary has been sent on 17 May, 2013

Thesis can be viewed at Library of Vilnius University

ĮVADAS

2008 m. ir 2010 m. kilusios finansinės krizės parodė, kad aktualus tikslesnių metodų bei jų kūrimui ir taikymui skirtų įrankių kredito rizikos vertinimui kūrimas. Kredito rizikoje pagrindinis uždavinys yra tikimybės, kad skolininkas negebės įvykdyti įsipareigojimų numatytu laiku, įvertinimas, kas tuo pačiu leidžia sumažinti visų pinigų praradimo tikimybę. Tokių įsipareigojimų sumažinimas yra kritinis rizikos valdymui ir optimaliam kapitalo suformavimui finansinėse institucijose, atsižvelgiant į tai, kad Basel II susitarime apibrėžiamuose reguliavimo standartuose kapitalo valdymui skiriamas ypatingas dėmesys. Šis susitarimas reglamentuoja bei pateikia gaires ir kredito vertinimo modelio kūrimui bei naudojimui, su tuo susijusių duomenų saugojimu bei prieiga prie jų, su kredito rizika siejamų rizikų įtakos vertinimui. Taigi teisingų, efektyvių, reglamentuotų bei tikslių kredito rizikos vertinimo modelių kūrimas yra sudėtingas uždavinys, esminis kiekvienai finansinei institucijai, kuris sprendžiamas įvairaus detalumo lygmenyse, įskaitant finansinio instrumento tipą, modeliavimo metodus (statistiniai, ekonometriniai, matematiniai, intelektiniai ir t.t.). Vienas iš plačiausiai naudojamų metodų yra klasifikavimas, einamuoju metu plačiai tiriamas bei aprašomas įvairiose publikacijose. Atraminių vektorių mašinos (angl. *Support Vector Machines*, sutr. SVM) einamuoju metu yra vienas iš plačiausiai kuriamų, tiriamų bei taikomų klasifikavimo metodų problemų įvairiose probleminėse srityse sprendimui, pasiūlytas Vapnik (Cortes, Vapnik, 1995; Vapnik, 1998) ir toliau vystomas bei aprašytas Scholkopf ir kt., Cristianini ir kt., Baesens, van Gestel ir kt., Mangasarian ir kt., Huang ir kt., Lai, Yu ir kt., Chang ir kt., Steinwart ir kt., Joachims darbuose. Jis naudojamas klasifikavimo problemų įvairiose srityse, tokiose, kaip bioinformatika ir skaičiuojamoji biologija (Yang, 2004; Ben Hur et al., 2008), dokumentų klasifikavimas (Joachims, 1998; Khan et al., 2010), ir kt. sprendimui. Nuolat siūlomi įvairūs būdai, apjungiantys SVM su kitais metodais, ar modifikuojantys standartinį SVM algoritmą, siekiant gauti geresnius rezultatus. Balthazar (Balthazar, 2006) teigia, kad SVM pagrįstas modelis naudojamas ir praktikoje, kreditų reitingus kuriančios ir teikiančios Standard & Poor‘s kompanijos naudojamuose algoritmuose. Šio metodo sudėtingumas, analitiniai, skaičiuojamieji bei kūrimo aspektai smarkiai apsunkina šio metodo taikymą realių uždavinių sprendimui, taip sudarydami „klasifikavimo metodo ekspertams“ įvaizdį. Šiame darbe siekiama ištirti

šiuo metodu grindžiamus modelius, jų kūrimą bei taikymą, identifikuoti pagrindinius metodo sunkumus, pasiūlyti naujus sprendimus, kurių pagrindu galėtų būti kuriami nauji metodai ar modeliai tiek kredito rizikos sričiai, tiek ir kitoms sritims.

Tyrimo problema

Kaip jau pastebėta įvade, ši tema plačiai tiriama ir svarbi finansinėms institucijoms, nors SVM pagrįstų algoritmų kūrimas yra aktualus visam dirbtinio intelekto, skaičiavimo ir modeliavimo mokslui. Šio tyrimo rezultatai gali būti pritaikyti praktiškai, įskaitant ir sukurtų metodų integravimą intelektualią sprendimų paramos sistemą, skirtą tiek moksliniams, tiek ir verslo poreikiams. Augantis atvirai prieinamų bei tarpusavyje susietų duomenų (angl. *Linked Data*) kiekis sudaro prielaidas jų išnaudojimui, papildant ar patikslinant jau esamus metodus bei į juos integruojant naujas žinias, apjungiant jau egzistuojančias ekspertines žinias ar turimą patirtį, išreikštą lingvistiniais terminais ar skaitinėmis reikšmėmis, su turimais duomenimis. Šiame kontekste tampa svarbi ir integracija su semantinio žiniatinklio standartais, ypač finansinės srities. Taigi kita svarbi problema yra tyrimų, orientuotų į intelektualių sprendimų paramos sistemų kūrimą, apimantį finansinių standartų ir struktūrų, tokių kaip XBRL (*Extensible Business Reporting Language*), trūkumas. Šie standartai palaikomi ir Basel II reglamentavimo, kadangi jie užtikrina aiškų ir struktūrizuotą finansinės informacijos pateikimą, taigi jų integracija svarbi visame šiuolaikiniame finansinių sprendimų paramos procese. Sprendimo paramos sistemos karkaso projektavimas bei kūrimas, apimantis ir metodinius dalykus, taip pat yra aktualus programinės įrangos inžinerijos mokslui, kadangi jis pateikia modernų požiūrį į sudėtingų išskirstytų daugiakomponenčių sprendimų paramos sistemų inžineriją, kuris ateityje gali būti toliau vystomas bei tobulinamas.

Tyrimo objektas

Šis darbas skirtas intelektinių kredito rizikos vertinimo bei bankrotų prognozavimo metodų, grindžiamų atraminių vektorių mašinų (angl. *Support Vector Machines*, sutr. SVM) klasifikavimo metodais, tyrimui. Antrinis šio tyrimo objektas yra karkasas intelektualių sprendimų paramos sistemų kūrimui, apimantis siūlomus bei panašius metodus, finansinius standartus, projektavimo ir kūrimo metodiką, galimą

realizavimo scenarijų.

Tyrimo tikslas ir uždaviniai

Tyrimo tikslas yra pasiūlyti metodą(-us) atraminių vektorių mašinų klasifikavimo metodu grindžiamo modelio kredito rizikos vertinimui kūrimui, naudojantį turimus finansinius duomenis bei turimus įverčius, išreiškiamus klasėmis (pavyzdžiui, ekspertinius vertinimus). Kitas tikslas yra sukurti karkasą intelektualių kredito rizikos srities sprendimų paramos sistemų kūrimui, integruojantį finansinius standartus, komponentus, būdingus tokio pobūdžio sistemoms, siūlomu klasifikavimo metodu grindžiamus sprendimus, pasiūlyti tokių sistemų projektavimo ir kūrimo metodiką bei galimą realizavimo scenarijų.

Šiam tikslui pasiekti sprendžiami tokie uždaviniai:

1. Išnagrinėti statistinius, ekonometrinius nei dirbtinio intelekto metodus, skirtus kredito rizikos vertinimui, pagrindinius einamuoju momentu sukurtus bei ankstesnius sprendimus kredito rizikos sričiai, grindžiamus šiais metodais, nustatyti jų pagrindinius privalumus.
2. Sukurti intelektiniais metodais grindžiamą klasifikavimo metodą ar metodus tiriamai problemai.
3. Atlikti sukurtų klasifikavimo metodų eksperimentinį įvertinimą, ištirti bei įvertinti gautus rezultatus.
4. Ištirti šiuo metu sukurtų finansinių sprendimų paramos sistemų struktūras, reglamentavimą bei reikalavimus tokių sistemų kūrimui, finansinius standartus, jų tikslus bei taikymo sritis, nustatyti ir, esant poreikiui, pasiūlyti jų integravimo bei taikymo intelektualioje sprendimų paramos sistemoje kredito rizikos vertinimui integravimo bei taikymo modelius.
5. Suprojektuoti ir sukurti intelektualių sprendimų paramos sistemų kredito rizikos vertinimui karkasą, apimantį siūlomus metodus, pagrindinius tokio pobūdžio sistemų komponentus, finansinių standartų integraciją, galimus projektavimo, kūrimo bei realizavimo scenarijus.
6. Realizuoti sprendimų paramos sistemą, naudojant sukurtą karkasą.

Tyrimų metodika ir įrankiai

Tyrime naudoti tokie metodai: bendrasis pažinimas (tyrimo uždavinių ir tikslų formulavimas, informacijos rinkimas bei analizė; apibendrinimas, išvadų formulavimas), bendrieji mokslinio tyrimo metodai, tokie, kaip indukcija, dedukcija, palyginimas (metodų, charakteristikų, panašumų, skirtumų), duomenų analizė ir modeliavimas; struktūrizavimas, grupavimas, apibendrinimas, abstrakcija bei pateikimas.

Tyrime pateikiamiems algoritmams ir metodams realizuoti naudoti atvirojo kodo automatinio mokymosi paketai WEKA, SVM paketai LibSVM ir LIBLINEAR. Pastarieji, taip pat RapidMiner bei kitos SVM algoritmų realizacijos, buvo naudojami lyginamojoje analizėje. Techninio skaičiavimo ir modeliavimo sistema MATLAB buvo naudojama PSO-LinSVM algoritmo pradiniam kūrimui, modeliavimui ir testavimui. Pasiūlytų algoritmų bei metodų tyrimuose naudoti SEC EDGAR duomenų bazės, apimančios 9365 JAV kompanijų iš 9 sektorių 1999-2008 m. finansinius (ketvirtinių bei metinių balansų ir finansinių ataskaitų) rodiklius, poaibiai, UCLA LoPucki bankrotų duomenų bazė, apimanti 911 realių JAV bankrotų duomenis (iš jų 253 kompanijos tiesiogiai susietos su tyrimuose naudojama EDGAR duomenų baze) bei Australijos ir Vokietijos UCI saugykloje pateikiami kreditų duomenų rinkiniai (atitinkamai 690 ir 1000 įrašų). UML ir BPMN notacijos buvo naudojamos metodų ir karkaso projektavimo iliustravimui; projektavimo ir kūrimo metodologijos kūrimui naudota šio darbo autoriaus pasiūlyta UML kalba ir domenu grindžiamo projektavimo (angl. *Domain Driven Design*) autoriaus Evans rekomendacijomis grindžiama notacija. Grafiniam modeliavimui įrankiai naudoti MagicDraw ir Microsoft Visio.

Disertacijos ginamieji teiginiai

1. Hibridinis spiečiaus optimizavimu ir tiesiniais atramos vektorių metodais grindžiamas klasifikavimo metodas, pasižymintis automatiniu tiesinio klasifikavimo metodo parinkimu iš aibės metodų su tokiais pačiais parametrais, kartu parenkant ir jo parametrus, gali būti efektyvus tiek su mažos, tiek ir su didesnės apimties duomenų rinkiniais.

2. Sukurtas klasifikavimo metodas, apimantis savybių parinkimą, SVM grindžiamą klasifikavimą bei „slenkančio lango“ testavimo principą, gali būti naudojamas kredito rizikos vertinimo srities klasifikavimo modelio kūrimui bei

testavimui.

3. XBRL finansinio standarto integravimas į finansinės srities sprendimų paramą gali išplėsti modelio kūrimo procesą papildomais duomenimis, įgalinti automatinį standartizuotų ir struktūrizuotų finansinių duomenų importavimą bei modelio kūrimą bei atnaujinimą realiu laiku.

4. Struktūra, apjungianti siūlomus metodus, finansinius standartus, projektavimo ir kūrimo metodiką, galimą realizavimo scenarijų, grindžiamą daugiaplatformiškumu bei nepriklausomumu nuo duomenų šaltinio, yra svarbus įrankis modernių intelektualių sprendimo paramos sistemų kredito rizikos vertinimui kūrimui.

Darbo mokslinis reikšmingumas

Pasiūlytas naujas hibridinis klasifikavimo metodas PSO-LinSVM, naudojantis spiečiaus optimizavimu grindžiamą procedūrą automatiniam tiesinio SVM klasifikatoriaus parinkimui. Skirtingai nuo anksčiau pasiūlytų panašių euristiniu optimizavimu ir SVM grindžiamų metodų, pasiūlytas algoritmas parenka ne branduolio funkcijos parametrus, tačiau patį tiesinį SVM klasifikavimo metodą kartu su sudėtingumo ir poslinkio parametrais iš aibės tiesinių SVM klasifikavimo metodų su šiais parametrais. Patikrintas siūlomo metodo efektyvumas klasifikuojant įvairaus dydžio duomenis. Šis metodas gali būti naudojamas sprendžiant klasifikavimo problemas įvairiose srityse (finansuose, bioinformatikoje ir kt.). Darbe pasiūlytas kredito rizikos vertinimo metodas, grindžiamas išoriniais ekspertiniais, paremtais diskriminantiniais modeliais, vertinimais, turintis savybių atrinkimą, klasifikavimą bei slenkančio lango principu pagrįstą testavimą. Skirtingai nuo anksčiau pasiūlytų metodų, pasiūlytas metodas leidžia atlikti testavimą, naudojant ne duomenų dalį, o vieno ar daugiau sekančių periodų duomenis, tai padeda patikrinti modelio veiksnumą keliuose sekančiuose perioduose; taip pat pažymėtinas ir kontekstas, kuriame šis metodas buvo tiriamas – didelis naudotų duomenų kiekis bei jų matiškumas, integracija su išoriniais šaltiniais bei standartais, kas panašaus pobūdžio tyrimams iki šiol nebuvo būdinga, tačiau tampa aktualu vis labiau augant prieinamų informacinių šaltinių kiekiui ir juose pateikiamų duomenų apimčiai. Pasiūlytas komponentinis karkasas bei jo pagrindu kuriamos sistemos projektavimo ir kūrimo metodika aktualūs programinės įrangos inžinerijos mokslui, kadangi jie aprašo išskirstytų komponentinių skaitinio intelekto

metodais grindžiamų sprendimų paramos sistemų struktūrą, pagrindinius komponentus bei kūrimo procesus, naudojant tiriamą sistemą kaip atvejį studijai, pagal komponentinės programinės įrangos inžinerijos principus bei modernias metodikas, kas leidžia palengvinti jų kūrimą.

Darbo praktiniai rezultatai

Sukurti FS-SVM ir FS-SVM^{SWTest} metodai gali naudojami susieti išorinius reitingų duomenis ar ekspertinius įvertinimus su jau esančiais finansiniais, operaciniais ar rinkos duomenimis, naudojamais kredito rizikos vertinime, siekiant identifikuoti tarp jų esančias vidines priklausomybes bei esminius faktorius, taip pat nustatyti esminius duomenų taškus (atraminius vektorius), formuojančius ar esančius šalia skiriamosios plokštumos. Sukurti metodai taip leidžia realizuoti ekspertinę patirtį kaip modelį, kuris gali būti perkeltas ir naudojamas aplinkoje nepriklausomai nuo eksperto buvimo. Toks modeliavimo principas naudingas ir įvertinant atvejus, kurių nebuvo įmanoma įvertinti ekspertiniu būdu dėl trūkstamų duomenų ar atsirandančių neapibrėžtumų.

Sukurtas klasifikavimo metodas, grindžiamas tiesiniais SVM ir spiečiaus intelektu, gali būti naudojamas spręsti klasifikavimo problemas įvairiose srityse, kuriose gali būti išskirti esminiai požymiai ar faktoriai, leidžiantys formuluoti tokias problemas. Sukurtas karkasas ir nuo platformos ir duomenų šaltinių nepriklausomas jo realizavimo scenarijus gali būti naudojamas kurti modernias išskirstytas, sudarytas iš didelio komponentų skaičiaus bei apimančias keletą kontekstų kredito rizikos vertinimo sprendimų paramos sistemas su integruotais finansiniais standartais. Tokių sistemų projektavimo ir kūrimo metodika, pateikiama kaip šio karkaso dalis, gali būti adaptuota bei taikoma ir kitų (finansų, inžinerijos ir t.t.) sričių sprendimų paramos sistemų kūrimui.

DARBO SANTRAUKA

Atraminių vektorių mašinos

Atraminių vektorių mašinos (angl. *Support Vector Machines*, sutr. SVM) yra statistinio mokymosi teorija, sukurta Vapnik ir Chervonenkis, pagrįstas metodas; teoriniai šio metodo aspektai plačiau aprašyti (Vapnik, 1998). Pirmą sykį šis metodas aprašytas (Cortes, Vapnik, 1995); vėliau jis buvo plačiai naudojamas įvairių objektų atpažinimo, klasifikavimo ir regresijos uždavinių sprendimui įvairiose srityse, tokiose, kaip bioinformatika, finansai, dokumentų klasifikavimas, vaizdų atpažinimas ir kt.

Esminis šio metodo ypatumas netiesinis n-matės įeigos erdvės atvaizdavimas kitoje (galimai didesnių matavimų) erdvėje, leidžiančiai suformuoti tiesinę skiriamąją hiperplokštumą. Tuo pačiu metu minimizuojama empirinė klasifikavimo klaida bei maksimizuojama geometrinė riba; dėl šių priežasčių SVM dažnai dar vadinamas maksimalios ribos klasifikatoriumi (angl. *maximum margin classifier*). Formaliai šį uždavinį galima suformuluoti taip (Vapnik, 1998): jei duotas empirinių duomenų vektorius $(\mathrm{x}_1, \mathrm{y}_1), (\mathrm{x}_2, \mathrm{y}_2), \ldots, (\mathrm{x}_\mathrm{n}, \mathrm{y}_\mathrm{n}) \in \chi \times \{\pm 1\}$, $x_i \in R^N$, $y_i \in \{-1,1\}$, y – žymių (angl. labels) vektorius, ieškoma sprendimo funkcijos $f_{\mathbf{w},b}(x_i) = y_i$, *i = 1..n*. Panašumas tarp χ ir y gali būti formaliai aprašytas kaip branduolio funkcija $\mathrm{k} : \chi \times \chi \to \mathrm{R}, (\mathrm{x}, \mathrm{x}') \to \mathrm{k}(\mathrm{x}, \mathrm{x}')$, kurios rezultatas yra realus skaičius, aprašantis panašumą tarp *x* ir *x'*. Taigi ieškoma skiriamosios hiperplokštumos, minimizuojančios ribinę klaidą; ši hiperplokštuma aprašoma kaip atraminių vektorių (taškų, kuriems ir tik kuriems Lagranžo funkcija nelygi 0) aibė. Branduolio funkcija naudojama, siekiant iš netiesiškai atskiriamų taškų suformuoti kitą vektorinę erdvę, kurioje tiesinis atskyrimas įmanomas. Populiariausios branduolio funkcijos, naudojamos praktikoje, yra tiesinė, polinominė, radialinės bazės (RBF), sigmoidinė (Chang, Lin, 2001), nors gali būti naudojamos ir kitos, specialiai probleminei sričiai sukurtos funkcijos.

Klasikinis Vapnik aprašytas SVM algoritmas (dar žinomas kaip C-SVC) apibrėžiamas kaip kvadratinio optimizavimo uždavinys (Chang, Lin, 2001):

$$\min_{w,b,\zeta} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}\zeta_i$$

$$s.t.\ \ y_i(w^T\phi(x_i)+b) \geq 1-\zeta_i, \zeta_i \geq 0, i = 1,..,N$$

kur C apibrėžia kompromisą tarp maksimalios ribos ir minimalios klasifikavimo klaidos.

Sprendimas apibrėžiamas kaip funkcija $\langle \phi(\mathbf{x}) \cdot \mathbf{w} \rangle + b = 0$. Vėliau, remiantis šiuo algoritmu sukurta daug SVM algoritmų ir modifikacijų: ν–SVC (Scholkopf et al., 2000), nuoseklus minimalus optimizavimas (angl. *Sequential Minimal Optimization*, sutr. SMO) (Platt, 1999), mažiausių kvadratų SVM (Suykens, Vandevalle, 1999), SVM$^{\text{Light}}$ (Joachims, 2001), BSVM (Hsu, Lin, 2002), esminių vektorių mašinos (angl. *Core Vector Machines*, sutr. CVM) bei jų mažiausių kvadratų versija CVM-LS (Tsang et al., 2005), sferinių vektorių mašinos (angl. *Ball Vector Machines*, sutr. BVM) (Tsang et al., 2007), transduktyvūs SVM (Keerthi, 2005), potencialiniai SVM (Hochreiter, Obermayer, 2006) ir kt. Šių metodų detalesnis aprašymas pateikiamas darbe. Dar viena svarbi klasifikatorių klasė yra tiesiniai SVM klasifikatoriai, aprašyti (Fan et al, 2008). Pastebėtina, kad jų formuluotės skiriasi nuo klasikinio tiesinio SVM, be to, jie orientuoti į darbą su didesniais duomenų kiekiais ir pasižymi. Šie klasifikatoriai vėliau naudojami darbe.

Pastebėtina, kad originalus SVM algoritmas skirtas binariniam klasifikavimui, t.y., $y_i \in \{-1,1\}$. Siekiant jį naudoti keleto klasių ($y \in [1,2,..,N]$) klasifikavimo problemų sprendimui, naudojami *one-vs-all* (OVA), *one-vs-one* (OVO), Crammer-Singer metodai (Crammer, Singer, 2000; Debnath et al., 2004); jų formuluotės pateikiamos darbe. Darbe remiamasi „vienas prieš visus“ (*one-vs-all*) strategija.

Apžvelgiant programines šių metodų realizacijas, pastebėtina, kad apžvelgiami metodai realizuoti skirtingomis programavimo kalbomis, be to, trūksta brandžių programinių paketų, apimančių daugumą SVM algoritmų bei realizacijų Tai ženkliai apsunkina lyginamąjį šių metodų tyrimą bei jų tarpusavio apjungimą

Atlikus atraminių vektorių metodo analizę, išskirti tokie jo privalumai:

- Uždavinio konvertavimas į kvadratinio programavimo uždavinį leidžia išvengti lokalių minimumų ir gauti globaliai optimalų sprendimą;
- Parametrų erdvės kontrolė, naudojant optimalų ribinį parametrą;
- Geri klasifikavimo rezultatai, lyginant su panašiais metodais;
- Persimokymo, prisitaikymo prie duomenų, architektūros parinkimo bei testavimo problemų išvengimas;
- Plati tyrimų bazė, daug algoritmų bei modifikacijų;
- Išlygiagretinimo galimybė (Chang et al.. 2007).

Kaip pagrindiniai šio metodo trūkumai išskiriami tinkamų parametrų parinkimo

problema, lėtas testavimas, realizacinis sudėtingumas, poreikis skaičiavimų ištekliams, orientacija į binarinį klasifikavimą, kas apsunkina darbą su daugiau nei 2 klasių duomenimis, bei brandžių programinės įrangos paketų, orientuotų į šį metodą, trūkumas.

Kredito rizikos vertinimo ankstesni tyrimai

Finansinės rizikos vertinimas yra viena svarbiausių problemų finansų valdyme tiek analitikui, tiek investuotojui, tiek ir finansinei institucijai, kadangi yra siekiama identifikuoti pagrindinius parametrus, apibrėžiančius kiekybinį rizikos, prisiimamos investuojant ar skolinant pinigus, kiekį – rizikos kiekį, išreiškiamą galimo praradimo apimtimi, bei kokybę, apibūdinama praradimo tikimybe. Kiekis gali būti apribojamas apribojant prisiimamų rizikų apimtį, tuo tarpu kokybės vertinimui naudojamos įvairios priemonės, įskaitant ir skolininkų reitingų sudarymą. Kredito rizika ypatingai svarbi bankams, kadangi ji apibūdina paskolų portfelio valdymo, kuris paprastai sudaro didžiausią banko aktyvų dalį, kokybę; nuo efektyvaus kredito rizikos valdymo tiesiogiai priklauso banko akcininkų pelnas bei banko finansinių praradimų kiekis. Pati kredito rizika gali būti apibūdinama kaip praradimai, patiriami banko, kai viena iš pusių negali įvykdyti savo įsipareigojimų (Kancerevyčius, 2004). Ši rizika gali būti patiriama įvairiuose lygmenyse: instrumentų (obligacijos, opcionai, ateities sandoriai, palūkanos), skolininkų (individualūs asmenys, kompanijos, vyriausybės) ir t.t. Kredito rizikos vertinimas dar žinomas kaip kredito analizė; žmonės, atliekantys šį procesą, apibūdinami kaip kredito analitikai.

Šios problemos vertinimui naudojama įvairių metodų, tokių, kaip 5-C (*Character*, *Capital*, *Capacity*, *Collateral*, *Cycle (or Economic) Conditions*) taisyklė (Saunders, Allen, 2002; Anderson, 2007), vidinis reitingais grindžiamas modelis (Saunders, Allen, 2002), kredito reitingai. Pagrindiniai rodikliai, apibrėžiantys riziką kiekybiškai, yra nemokumo tikimybė (angl. *Probability of Default*, sutr. PD), kredito pozicija įsipareigojimų neįvykdymo atveju (angl. *Exposure at Default*, sutr. EAD), atgavimo rodiklis (angl. *Recovery Rate*), nuostolis nemokumo atveju (angl. *Loss Given Default*, sutr. LGD), brandumas (angl. *Maturity*) (Balthazar, 2006; van Gestel, Baesens, 2009). Kita plačiai naudojama priemonė yra skolininkų reitingavimas, suteikiant jiems vidinius ar išorinius reitingus. Vidiniai reitingai suteikiami pačių kreditorių, atsižvelgiant į ankstesnę paskolų istoriją bei finansinę skolininkų būklę. Išoriniai reitingai suteikiami

pripažintų kredito rizikos vertinimo kompanijų, tokių, kaip *Fitch ratings, Moody's, Standard & Poor's, Dunn & Bradstreet,* ir patvirtina subjektų reputaciją, finansinius pajėgumus bei galimybę laiku įvykdyti nominalios vertės bei palūkanų mokėjimus paskolos suteikimo atveju. Taip pat taikomos statistiniu bei matematiniu modeliavimu grindžiami struktūriniai, sumažinto formos, VaR (*Value at Risk*) bei mirtingumo rodiklio modeliai (Allen, 2002; Elizalde, 2006a; Elizalde, 2006b). Struktūriniais modeliais grindžiami plačiai paplitę komerciniai KMW/Moody's (Merton OPM), KMV's Credit Manager ir Moody's RiscCalc kredito rizikos modeliai (Allen, 2002). Šie modeliai darbe plačiau nenagrinėjami, tačiau pateikiamas jų palyginimas su automatinio mokymosi bei statistiniais klasifikavimo metodais, sudarytas pagal (Balthazar, 2006).

Kredito vertinimo balais (angl. *scoring*) metu yra skaičiuojamas kredito įvertinimas (angl. credit score), naudojant apie skolininką turimą informaciją. Šiam įverčiui gauti naudojami įvairūs metodai, aprašomi ir pateikiamame darbe. Pirmieji tyrimai siejami su statistiniais diskriminantinės analizės (Altman, 1968; Springate, 1978; Altman, 2000), probit analizės (Zmijewski, 1984), logistinės regresijos (Ohlson, 1980), pavojaus (angl. *hazard*) analizės (Shumway, 2001) metodais. Taip pat tirti bei taikyti ir hibridiniai skaitinio intelekto metodai, tokie, kaip neuroniniai tinklai (Wong, Selvi, 1998; Vellido et al., 1999; Wong et al., 2000), taip pat ir saviorganizuojančių neuroninių tinklų tipas (Kaski et al., 2001; van den Berg, 2006; Merkevičius et al., 2007), neraiškiųjų tinklų (Piramuthu, 1999; Mahotra, Mahotra, 2002), neraiškiųjų SVM (Hao et al., 2007; Hao et al., 2008; Chong, 2008), šiurkščios aibės (Lv, Peng, 2008; Zhou et al., 2008) bei jų ir SVM hibridiniai metodai (Wang et al., 2007; Wang et al., 2008; Ping, Yongheng, 2011). Šie metodai trumpai taip pat apžvelgiami pačiame darbe bei SVM grindžiamų tyrimų kredito rizikoje analizėje (Danenas, Garšva, 2009).

Kita aktuali problema, kuriant tokio tipo modelius, yra finansinių duomenų panaudojimas bei rodiklių atrinkimas. Remiantis šaltiniais, identifikuota 20 grupių kintamųjų, kurie gali būti naudojami kurti kredito rizikos vertinimo modelius klasifikavimui; pastebėtina, kad kai kurių grupių rodikliai aktualūs tik tam tikrų reitingų kūrimui (pavyzdžiui, politinių rodiklių grupė labiau atspindi vyriausybių ir valstybių veiklą. Atsižvelgiant į duomenų prieinamumą bei naudojamų duomenų pobūdį, darbe apsiribojama kompanijų vertinimu pagal jų finansinius rodiklius.

Spiečiaus optimizavimas ir jo pagrindu sukurtas algoritmas

Spiečiaus intelekto metodai grindžiami įvairių būtybių kolektyvine socialine elgsena bei jos sinchroninių veiksmų modeliais. Spiečiaus optimizavimas (angl. *Particle swarm optimization*, sutr. PSO), pasiūlytas Kennedy (Kennedy, 1995), yra efektyvus metodas sudėtingų optimizavimo problemų su daug lokalių minimumų sprendimui, grindžiamas paukščių būrio elgesiu, ieškant maisto atsitiktinai tame pačiame plote. Šiuo atveju kiekvienas sprendimas (dalelė) atitinka paukštį, jo santykinis atstumas nuo ieškomo objekto, apibrėžiamas tinkamumo (angl. *fitness*) funkcija. Visos dalelės turi vieną funkcijos vertę, apibrėžiamą optimizuojama funkcija, bei judrumo (angl. *velocity*) parametru, apibrėžiančiu judėjimo kryptį bei atstumą. Visos dalelės iteratyviai atlieka paiešką sprendimų erdvėje, sekdamos tuo metu optimalią tinkamumo funkcijos vertę turinčią dalelę. Kiekvienos iteracijos metu kiekvienos dalelės tinkamumas keičiamas pagal du ekstremumus: dalelės rastą optimalų sprendimą ir viso spiečiaus rastą geriausią sprendimą. Pagrindiniai šio metodo parametrai yra spiečiaus dydis, iteracijų skaičius, judrumą įtakojantys koeficientai c_1 (kognityvinis, dar įvardijamas kaip nostalgijos) ir c_2 (socialinis, dar apibūdinamas kaip pavydo), taip pat inercijos svorio faktorius (Kennedy et al., 1995; Engelbrecht, 2007). Nors (Kennedy et al., 1995; Engelbrecht, 2007) pateikia rekomendacijas šių judrumo koeficientų nustatymui, tačiau jis priklauso nuo sprendžiamo uždavinio, todėl tyrimuose šie parametrai pasirenkami pasirinktinai.

PSO-LinSVM. Atlikta tiesinių SVM algoritmų analizė leido identifikuoti jų bendruAtsižvelgiant į tiesinių SVM bendrus parametrus, sukurtas hibridinis klasifikavimo metodas, naudojantis spiečiaus optimizavimo tiesinio SVM klasifikavimo ir jo parametrų metodo parinkimui. Šiuo atveju dalelė $P = <p_1;p_2;p_3>$ apibrėžiama kaip:

p_1 *– sveikaskaitė (angl. Integer) reikšmė, atitinkanti klasifikavimo algoritmą*

p_2 *– realioji reikšmė, C parametras*

p_3 *–realioji reikšmė, poslinkis (angl. bias) parametras*

Tinkamumo funkcija apibrėžiama kaip teisingų teigiamų reikšmių santykio rodiklio (angl. *True Positive Rate*, sutr. TPR) rodiklių sumos maksimizavimas, siekiant subalansuoti klasifikavimo rezultatus nesubalansuotiems duomenų rinkiniams:

$$f_{fitness} = \sum_{i=1}^{N_C} TPR_i = \sum_{i=1}^{N_C} \frac{TP_i}{FN_i + TP_i},$$

kur N_C yra klasių skaičius, TP_i – teisingai identifikuotų „teigiamų“ reikšmių skaičius *i*-

tajai klasei, FN_i – neteisingai identifikuotų „neigiamų“ reikšmių skaičius *i*-tajai klasei. TPR reikšmės gaunamos, naudojant kryžminio validavimo metodą. Siūlomame algoritme taip pat atsižvelgiama ir į judrumo ribotumo (angl. *velocity clamping*) problemą (Engelbrecht, 2007). Kadangi optimalus sprendimas galimas tik idealaus klasifikavimo atveju, apsiribojama geriausio sprendimo paieška, todėl algoritmas vykdomas iki tol, kol nebelieka gerėjimo jo klasifikavimo našume. Pastebėtina, kad vidinis algoritmų žymėjimas nėra svarbus tikslumui, kadangi komponentė p_1 inicializuojama visoje paieškos erdvėje, naudojant atsitiktines reikšmes iš intervalo, aprašančio šį kodavimą, o pats optimizavimas atliekamas pagal SVM klasifikavimo modelio šiame taške našumą. Jis gali būti priklausomas nuo atsitiktinių skaičių generatoriaus, naudojamo algoritmo realizacijoje ar sistemoje. Pastebėtina, kad pačiam kodavimui yra keliamos tam tikros sąlygos – žymėjimui turi būti naudojamos sveitaskaitės neneigiamos reikšmės, einančios viena po kitos (t.y., $cl_{min} \leq P_{i1} \leq cl_{max,}\ S(i) = i+h$ kiekvienai p_1 reikšmei P_{i1}).Veikimo našumą gali įtakoti ir dalelių skaičius, naudojamas PSO optimizavimo procedūroje – kuo didesnis dalelių skaičius naudojamas, tuo geriau padengiama galimų sprendinių erdvė, tačiau tuo pačiu proporcingai auga ir skaičiavimui reikalingų išteklių poreikis.

Tolimesniuose tyrimuose naudojamas toks tiesinių SVM algoritmų kodavimas:

0 -- L2-regularized logistic regression (primal)

1 -- L2-regularized L2-loss support vector classification (dual)

2 -- L2-regularized L2-loss support vector classification (primal)

3 -- L2-regularized L1-loss support vector classification (dual)

4 -- L1-regularized L2-loss support vector classification

5 -- L1-regularized logistic regression

6 -- L2-regularized logistic regression (dual)

Pagrindiniai PSO-LinSVM algoritmo parametrai yra spiečiaus dydis *n*, kognityvinis koeficientas c_1, socialinis koeficientas c_2, maksimalus iteracijų, po kurių algoritmas baigia darbą, skaičius bei iteracijų, po kurių algoritmas nutraukia darbą, negerėjant rezultatams skaičius, C ir nuokrypio parametrų intervalai bei klasifikatorių sąrašas (čia apibrėžiamas intervalu, atsižvelgiant į jų vidinį kodavimą).

PSO-LinSVM(n, c1, c2, rangeC, rangeBias, terminate_iterations, max_iterations)
$k \leftarrow 3$ (number of dimensions in particle, representing different SVM classifiers as described above)

perf ← *[]*
$cl \leftarrow \{i \mid cl_{\min} \le i \le cl_{\max}, cl_{\min} \in Z, i \in Z, cl_{\max} \in Z\}$
global_fitness ← *0*
term_iterations ← *0*
$t \leftarrow 0$ *number of iterations*
$P \leftarrow Init(n)$ *Initialize a 3-dimensional swarm*
for $\forall p_x \in P$
$p_{x1} \leftarrow cl_{min} + round(rand(0,1) * (cl_{max} - cl_{min}))$
$p_{x2} \leftarrow cl_{min} + rand(0,1) * (C_{max} - C_{min})$
$p_{x3} \leftarrow b_{min} + rand(0,1) * (b_{max} - b_{min})$
$y_p \leftarrow p;$
repeat
if *no_iterations* = *max_iterations* return SVM(y_p);
for $\forall p_x \in P$ *set the personal best position*
$f(x_p) \leftarrow evalSVM(p_{x1}, p_{x2}, p_{x3})$
if $f(x_p) < \hat{y}(t)$ *set the global best position*
$y_p \leftarrow x_p$;
$term\_iterations \leftarrow 1$ *no need to terminate, continue searching*
else
$term\_iterations \leftarrow term\_iterations + 1$
if $f(y_p) < f(\hat{y}) \quad \hat{y} = y_p$
for $\forall p_x \in P$
for j=1:k
$Vmax \leftarrow \delta_j \times (R_{max,j} - R_{min,j})$ *Maximum allowed velocity*
if (j = 1)
$Vmax \leftarrow round(Vmax)$;
$v_{pj}(t+1) = v_{pj}(t) + round(c_1 \times rand(0,1) \times (y_{pj}(t) - x_{pj}(t)) + c_2 \times rand(0,1) \times (\hat{y}_j(t) - x_{pj}(t)))$
else
$v_{pj}(t+1) \leftarrow v_{pj}(t) + c_1 \times rand(0,1) \times (y_{pj}(t) - x_{pj}(t)) + c_2 \times rand(0,1) \times (\hat{y}_j(t) - x_{pj}(t))$
$v_{pj}(t+1) \leftarrow (v_{pj}(t+1) < Vmax\ ?\ v_{pj}(t+1) : Vmax)$
$x_p(t+1) \leftarrow x_p(t) + v_p(t+1)$
$y_p(t+1) \leftarrow \begin{cases} y_p(t), if\ f(x_p(t+1)) \le f(y_p(y)) \\ y_p(t+1), if\ f(x_p(t+1)) > f(y_p(y)) \end{cases}$
if $x_{p1}(t+1) > cl_{mas}$
$x_{p1}(t+1) \leftarrow cl_{min}$;
if $x_{p2}(t+1) < C_{min}$)
$x_{p2}(t+1) \leftarrow C_{min}$;
$\hat{y}(t) \leftarrow \min(f(y_0(t)), ..., f(y_n(t)))$
$t \leftarrow t+1$
until (*term_iterations* < *terminate_iterations*);
return SVM(y_p)
Rezultatas: Optimalus tiesinis SVM klasifikatorius SVM(y_p)

Algoritmas 1. PSO-LinSVM algoritmas

Judrumo ribotumui pagerinti skaičiuojamos maksimalaus leidžiamo judrumo $V_{max,j}$ reikšmės kiekvienai dimensijai. Tai atliekama įvertinant paieškos erdvės δ dalį; siūlomame algoritme šios reikšmės skaičiuojamos kaip

$$\delta_j = \frac{R_{\max,j}}{|R_{\min,j}| + |R_{\max,j}|} * 0.8$$

kur $R_{min,j}$ yra j-osios dalelės dimensijos minimali reikšmė, $R_{max,j}$ – jos maksimumas.

Eksperimentiniai šio algoritmo tyrimai atlikti su laisvai prieinamais Vokietijos ir Australijos kreditų duomenų rinkiniais MATLAB aplinkoje, naudojant *Optimization* pakete esantį modeliuojamojo atkaitinimo (angl. *Simulated Annealing*) algoritmą, Chen MATLAB aplinkoje sukurtą PSO realizaciją bei lyginant sukurtą klasifikavimo metodą su LibSVM ir LS-SVM paketų našumu. Tikslumas vertintas, naudojant du skirtingus metodus tinkamumo funkcijos formavimui: tikslumą bei TPR rodiklių kiekvienai klasei sumą. Naudotas 5 žingsnių kryžminis validavimas bei 7:3 santykis duomenų padalijimui į mokymo ir testavimo rinkinius (t. y., 70% duomenų naudojama mokymui). Taip pat naudotas panašus į PSO-LinSVM evoliucinis tiesinių SVM klasifikavimo metodas (Danenas, Garsva, 2012a; Danenas, Garsva, 2012b), kurio esminis skirtumas susijęs su realizacija – skirtingai nuo siūlomo algoritmo, tai vien realiąsias reikšmes naudojanti versija.

1 lentelė. PSO-LinSVM klasifikavimo rezultatai Vokietijos kreditų duomenų rinkiniui

	Tinkamumas pagal tikslumą					Tinkamumas pagal TPR sumą				
	Klas. metodo kodas	C parametras	Klaidų rodiklis	TPR_1	TPR_2	Klas. metodo kodas	C parametras	Klaidų rodiklis	TPR_1	TPR_2
LIBLINEAR + DS	0	46	0.214	0.897	0.527	0	46	0.214	0.897	0.527
PSO-LinSVM	3	14.808	**0.187**	0,894	0,634	3	14.808	**0.187**	0,894	0,634
Spiečiaus optimizacija										
LIBLINEAR	5	99,274	0,197	0,894	0,602	7	96,112	0,233	0,797	0,699
$LibSVM^{RBF}$	-	0,014	0,197	0,903	0,591	-	0,016	0,217	0,874	0,581
$LibSVM^{Sigmoid}$	-	2,885	0,247	0,889	0,581	-	11,406	0,380	0,720	0,613
$LS\text{-}SVM^{Poly}$	-	8,659	0,330	0,763	0,462	-	2,859	0,490	0,536	0,452
$LS\text{-}SVM^{RBF}$	-	4,674	0,217	0,889	0,548	-	3,944	0,240	0,870	0,516
Modeliuojamasis atkaitinimas										
LIBLINEAR	6	76,788	0,203	0,884	0,602	5	85,577	0,200	0,874	0,634
$LibSVM^{RBF}$	-	0,013	0,207	0,889	0,581	-	0,012	0,200	0,889	0,602
$LibSVM^{Sigmoid}$	-	19,520	0,297	0,966	0,140	-	10,932	0,273	0,908	0,387
$LS\text{-}SVM^{Poly}$	-	9,969	0,357	0,720	0,473	-	0,138	0,363	0,705	0,484
$LS\text{-}SVM^{RBF}$	-	2,198	0,240	0,894	0,462	-	5,752	0,230	0,870	0,548

1 lentelėje pateikiami PSO-LinSVM metodo rezultatai, gauti atliktus klasifikavimą su Vokietijos kreditų duomenų rinkiniu. Kiek netikėtai optimizuotas tiesinis SVM klasifikavimo metodas parodė geriausius rezultatus tikslumo bei atskirų klasių atpažinimo atžvilgiu; panašius rezultatus parodė tik netiesinis LibSVM klasifikatorius, naudojantis RBF branduolio funkciją. Tai, kad šie klasifikatoriai parodė geresnį našumą, nei tiesioginė parametrų paieška (LIBLINEAR + DS), įrodo, kad euristikos naudojimas hibridiniuose metoduose panašių problem sprendimui reikalingas. Be to, PSO-LinSVM parodė geresnius rezultatus tiek optimizuojant pagal tikslumą, tiek

pagal TPR rodiklių sumą (šiuo atveju tai tas pats klasifikatorius).

2 lentelėje pateikiami PSO-LinSVM, gauti atliktus klasifikavimą su Australijos kreditų duomenų rinkiniu. Šiuo atveju pastebėta, kad tiesioginis parametrų parinkimas davė geresnius rezultatus, nei SVM su euristiniu parametrų parinkimu naudojimas. PSO-LinSVM rezultatai vėl buvo geresni nei kitų euristiniu parametrų parinkimu grindžiamų SVM klasifikavimo metodų.

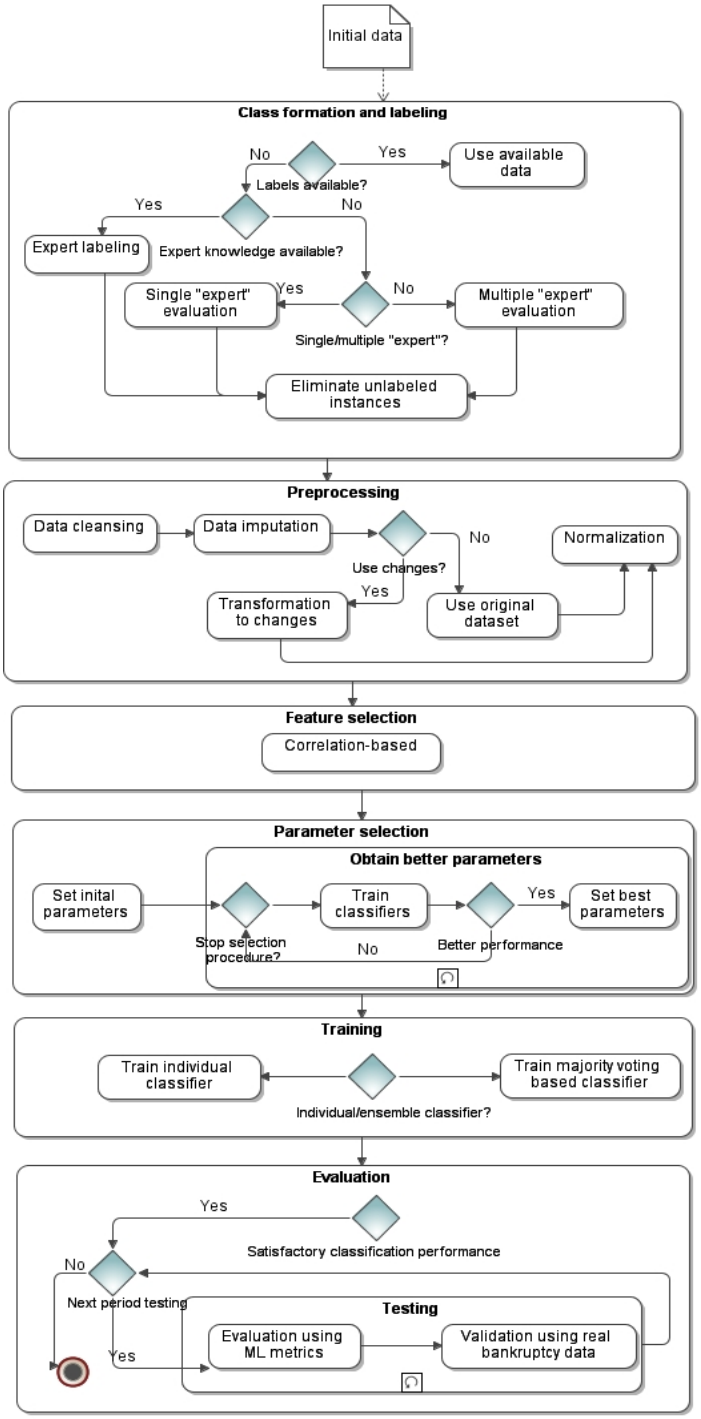
2 lentelė. PSO-LinSVM klasifikavimo rezultatai Australijos kreditų duomenų rinkiniui

	Tinkamumas pagal tikslumą					Tinkamumas pagal TPR sumą				
	Klas. metodo kodas	C parametras	Klaidų rodiklis	TPR_1	TPR_2	Klas. metodo kodas	C parametras	Klaidų rodiklis	TPR_1	TPR_2
LIBLINEAR+DS	5	6	**0.122**	0.864	0.896	5	6	**0.122**	0.864	0.896
PSO-LinSVM	1	33,606	**0.126**	0,853	0,901	1	33,606	**0.126**	0,853	0,901
Spiečiaus optimizacija										
LIBLINEAR	6	15,401	0,164	0,905	0,747	6	64,458	0,169	0,862	0,791
$LibSVM^{RBF}$	-	0,020	0,184	0,914	0,692	-	0	0,150	0,785	0,934
$LibSVM^{Sigmoid}$	-	20	0,169	0,905	0,736	-	9,159	0,159	0,922	0,769
$LS\text{-}SVM^{Poly}$	-	3,850	0,430	0,655	0,462	-	3,201	0,430	0,690	0,418
$LS\text{-}SVM^{RBF}$	-	9,972	0,164	0,879	0,780	-	10,327	0,164	0,879	0,780
Modeliuojamasis atkaitinimas										
LIBLINEAR	7	0,005	0,159	0,905	0,758	2	45,864	0,155	0,871	0,813
$LibSVM^{RBF}$	-	0,057	0,213	0,888	0,659	-	0	0,150	0,785	0,934
$LibSVM^{Sigmoid}$	-	0,119	0,179	0,897	0,725	-	16,945	0,159	0,871	0,890
$LS\text{-}SVM^{Poly}$	-	0,010	0,164	0,914	0,736	-	0,4	0,193	0,897	0,692
$LS\text{-}SVM^{RBF}$	-	2,656	0,159	0,879	0,791	-	4,363	0,145	0,897	0,802

Pastebėta, kad ne vienu atveju tik TPR suma grindžiamos tinkamumo funkcijos naudojimas leido pasiekti mažesnį klaidų rodiklio įvertį. Atsižvelgiant į tai, kad tiesiniai SVM klasifikavimo metodai parodė geresnius rezultatus nei netiesiniai, tolimesniame tyrime duomenų klasifikavimui nuspręsta remtis siūlomu klasifikavimo metodu.

Tiriamai sričiai tirti siūlomas metodas ir jo eksperimentinis tyrimas

Siekiant atlikti tiriamos srities modeliavimą, sukurtas modelis, apimantis tokius žingsnius, kaip klasių suformavimas pagal ekspertinius įvertinimus, nesant faktinių duomenų, duomenų transformavimas į pokyčius, jei siekiama tirti pokyčių dinamiką, apmokymo bei testavimo duomenų rinkinių suformavimas, reikšmingų atributų atrinkimas ir kt.



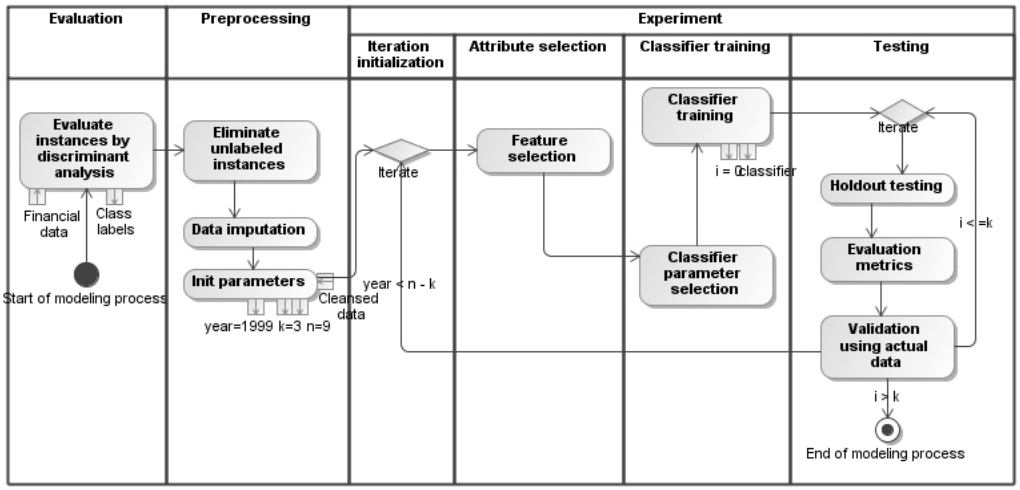
1 pav. FS-SVM grindžiamo klasifikavimo metodo kūrimo apibendrinta schema

Siekiant problemą tirti kaip klasifikavimo problemą, ekspertiniai įvertinimai buvo modeliuojami, naudojant kredito rizikos analizėje paplitusius diskriminantinius Altman, Springate, Zmijewski, Shumway modelius, t.y., kiekvienas finansinių duomenų vektorius buvo įvertinamas, naudojant šiuos modelius, ir gautos reikšmės konvertuotos į klases. Reikšmingų atributų atrinkimo žingsnis leidžia tiek sumažinti duomenų daugiamatiškumą, tiek ir suformuoti naują finansinių atributų aibę, kurios pagrindu

generuojamas naujas klasifikatorius.

FS-SVM modelis ir jo realizacija, grindžiama diskriminantinę analizę naudojančiu ekspertiniu vertinimu, FS-SVM^{DA} išsamiau aprašytas disertacijos 2.2 skyriuje; šio modelio tyrimo rezultatai pateikiami (Danenas, Garsva, 2010; Buzius et al., 2010; Danenas et al, 2011). Šio modelio schema pateikiama 1 pav. 2.4 skyriuje pateikiama jo modifikacija, praplečiant slenkamojo lango testavimu (Danenas, Garsva, 2012a; Danenas, Garsva, 2012b; Danenas, Garsva, 2012c). Šiuo atveju kiekvienam $m \in [1, n-k]$, kur n yra bendras periodų skaičius, k – periodų skaičius, naudojamas prognozavimui, atliekami tokie žingsniai:

a. Atributų atrinkimo procedūros pritaikymas, siekiant atrinkti reikšmingiausius atributus bei sumažinti duomenų daugiamatiškumą;
b. Klasifikatoriaus parametrų parinkimas pasirenkant parametrus ar naudojant euristines procedūras jų parinkimui. Taip pat čia gali būti naudojamas siūlomas PSO-LinSVM algoritmas, aprašytas anksčiau.



2 pav. Eksperimento, pagrįsto FS-SVM^{DA} ir slenkančio lango principu, sekos diagrama

c. Klasifikatorius apmokomas, naudojant pirmų m periodų duomenis.
d. Klasifikatorius testuojamas su p periodų duomenimis, $p \in [m+1, m+k]; \; p \in \mathrm{N}$

.

2 pav. pateikiama proceso modeliavimo, naudojant šį metodą, diagrama. Kiekvienos iteracijos rezultatas yra suformuotas klasifikatorius (SVM atveju tai atraminių vektorių aibė) bei atrinktų atributų aibė.

Disertacijoje pateikiami tyrimai atlikti, naudojant šiuos metodus ir įvairius SVM klasifikavimo algoritmus bei jų realizacijas (LibSVM, LS-SVM, LIBLINEAR, SVM^{Light},

CVM, CVM-LS ir kt.). Kai kurie tyrimo rezultatai pateikiami 2 lentelėje (disertacijoje pateikiami platesni ir įvairesni tyrimai). Eksperimentas atliktas naudojant EDGAR duomenų bazės, apimančios JAV kompanijų finansinius duomenis iš balansų bei pajamų ataskaitų, gamybos sektoriaus 1999-2008 m. periodo duomenis. Pradinis duomenų rinkinys apėmė 51 santykinį finansinį rodiklį; savybių atrinkimo procedūra ši skaičių sumažino. Pagrindinės duomenų rinkinio charakteristikos (duomenų kiekis, atrinktų rodiklių skaičius) pateikiami 3 lentelėje.

3 lentelė. Eksperimento duomenų rinkinio charakteristikos

Year	**Entries labeled as**		**Total entries**	**No of selected attributes**
	Risky (R)	**Not risky (NR)**		
1999	1312	537	1849	12
2000	1869	589	2458	15
2001	1753	672	2425	15
2002	1709	777	2486	13
2003	1770	723	2493	14
2004	1920	637	2557	13
2005	1964	660	2624	14
2006	1636	429	2065	14
2007	1545	393	1938	14
2008	483	109	592	14
Total	**15961**	**5527**	**21487**	

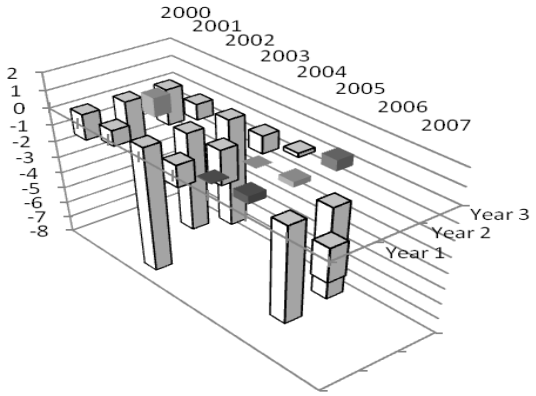
3 lentelėje pateikiami tyrimo rezultatai, gauti naudojant PSO-LinSVM grindžiamą klasifikavimo metodą su slenkančio lango principu. Klasių sudarymui naudotasi Zmijewski modeliu. Gauti klasifikavimo rezultatai geri, svyruojantys nuo 86% iki 95%; be to, aukštos TP (angl. *True Positive*) ir F-Measure rodiklių reikšmės rodo ir gerą atskyrimą tarp pačių klasių. Pastebėtina, kad nebuvo vyraujančio SVM klasifikatoriaus – tai rodo, kad automatinis klasifikatoriaus parinkimo procedūra padengė didelę dalį galimų sprendimų erdvės.

4 lentelė. Tyrimo rezultatai, naudojant PSO-LinSVM grindžiamą klasifikatorių

Training period			*2000*	*2001*	*2002*	*2003*	*2004*	*2005*	*2006*	*2007*
Classifier			L2-RLR (primal)	L2-SVM (dual)	L2-SVM (dual)	L2-RLR	L2-SVM (primal)	L2-SVM (dual)	L2-SVM (dual)	L2-SVM (dual)
C parameter			46,5068	9,4532	20,0452	76,0741	1,0000	32,1152	40,2581	20,4178
Bias parameter			-3,5519	9,5337	3,5257	-0,6641	5,2068	6,7547	2,2369	1,3727
	Accuracy		**95,218**	**95,46**	**87,655**	**94,253**	**91,679**	**93,52**	**86,12**	**90,372**
Year 1	**TP**	**R**	0,984	0,977	0,979	0,976	0,967	0,968	0,857	0,990
		NR	0,869	0,905	0,626	0,841	0,769	0,812	0,878	0,523
	F-Measure	**R**	0,967	0,967	0,918	0,962	0,945	0,959	0,908	0,944
		NR	0,91	0,926	0,747	0,879	0,824	0,839	0,719	0,667
Year 2	**Accuracy**		**93,853**	**95,311**	**89,367**	**94,743**	**92,94**	**91,744**	**86,486**	-
	TP	**R**	0,98	0,972	0,984	0,985	0,969	0,955	0,865	-
		NR	0,847	0,906	0,62	0,836	0,777	0,771	0,862	-
	F-	**R**	0,956	0,967	0,933	0,965	0,956	0,949	0,913	-

	Measure	**NR**	0,896	0,918	0,744	0,89	0,821	0,791	0,701	-
Year 3	**Accuracy**		**93,908**	**95,387**	**90,053**	**96,373**	**91,073**	**92,736**	-	-
	TP	**R**	0,967	0,976	0,993	0,99	0,954	0,969	-	-
		NR	0,87	0,889	0,629	0,863	0,74	0,743	-	-
	F-Measure	**R**	0,957	0,969	0,937	0,977	0,945	0,956	-	-
		NR	0,893	0,906	0,762	0,908	0,771	0,790	-	-

Siekiant palyginti rezultatus, kitas eksperimentas buvo atliktas su tuo pačiu klasifikavimo metodu, tačiau iteratyviai naudojant vartotojo pasirenkamus klasifikatorius. Testavimo rezultatai pasirinkti pagal geriausią kiekvieno sugeneruoto klasifikatoriaus testavimo rezultatų vidurkį. 3 pav. iliustruoja rezultatų skirtumus - jei PSO-LinSVM grindžiamo klasifikatoriaus rezultatai prastesni nei parinkti tiesinio SVM rezultatai su geriausiais rezultatais (t.y., skirtumas tarp atitinkamų testavimo rezultatų tikslumo atžvilgiu mažiau nei 0), jis atvaizduojamas kaip permatomas stulpelis, kitu atveju – kaip užpildytas.



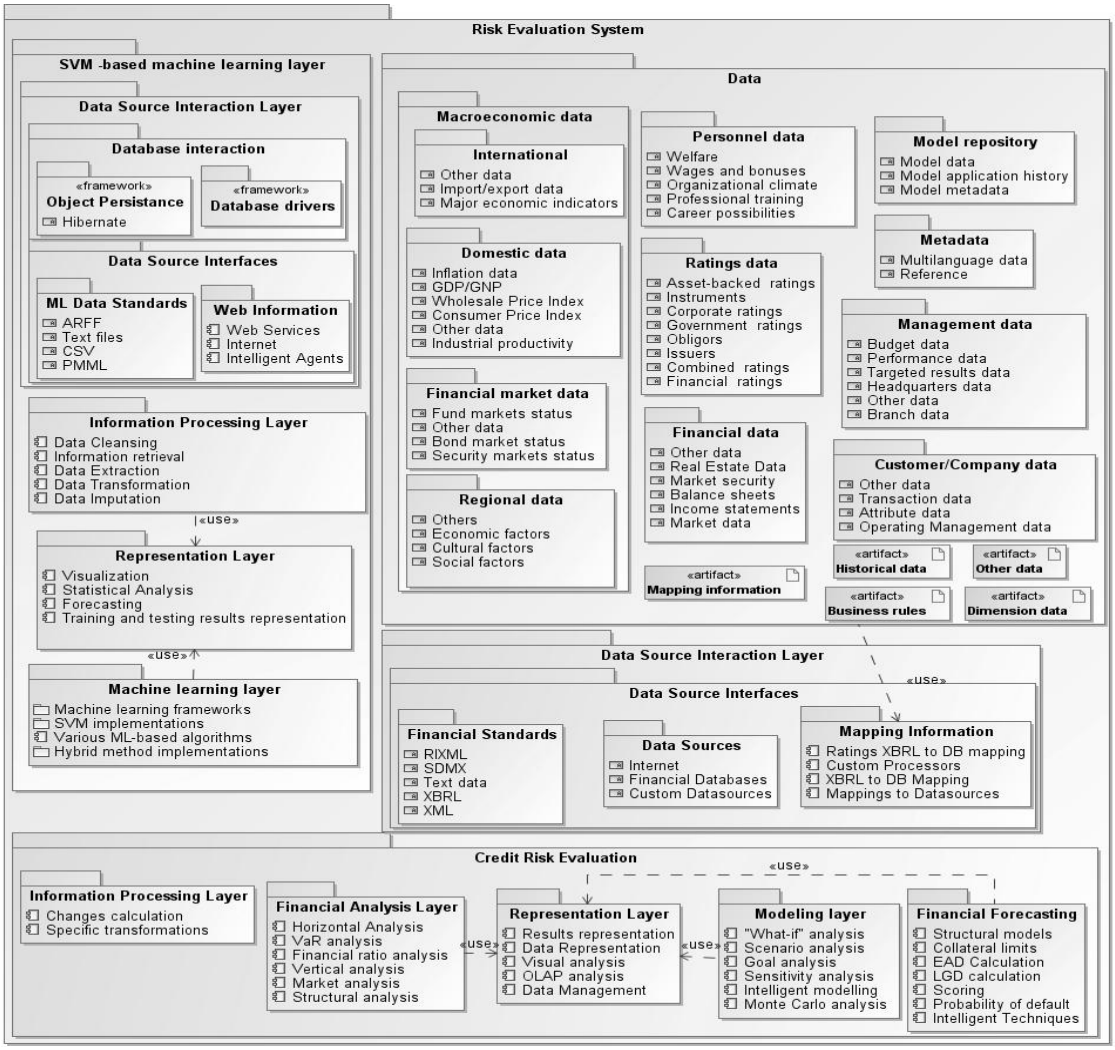
3 pav. Eksperimentų rezultatų skirtumų vizualizacija

Gauti rezultatai rodo, kad skirtumai buvo netolygūs – kai kuriais atvejais PSO-LinSVM grindžiamas klasifikatorius pasiekė ženkliai prastesnius rezultatus (>5% blogesnius nei parinkti vartotojo klasifikatoriai), tačiau kitais atvejais tie skirtumai nebuvo tokie ženklūs arba PSO-LinSVM grindžiamo klasifikatoriaus naudai.

Intelektuali sprendimų paramos sistema, grindžiama finansiniais standartais, SVM bei siūlomu modeliu

Sprendimų paramos sistemos (sutr. SPS) yra viena pagrindinių priemonių, skirtų sprendimo proceso palaikymui įvairiuose lygmenyse: individo, organizacijos, vyriausybinio. Jų kūrimas bei efektyvus panaudojimas aktualus jau daugiau kaip 40 metų (Power, 2008). Jos gali būti apibrėžtos kaip „duomenų modeliavimo bei ataskaitų pateikimo sistema, sukurta specifinių verslo klausimų ar problemų sprendimų bei atsakymų paieškai į iškylančius klausimus“ (Raynor, 1999). Hamilton apibrėžia SPS

kaip kompiuterizuotą sistemą, teikiančią problemų sprendimo bei komunikacijos galimybes pusiau struktūrizuotų/nestruktūrizuotų uždavinių sprendimui" (Hamilton, 2004). Šie apibrėžimai rodo, kad SPS skirtos dėl didelio duomenų ar susijusios informacijos, sudėtingo sprendimo išvedimo proceso ar dažnų situacijos pokyčių specifikuojamų užduočių bei sprendimų. Tokių sistemų kūrimas dažnai yra sudėtingas uždavinys, kadangi jis gali apimti ne vieną kontekstą, probleminę sritį, programavimo ar kūrimo paradigmą skirtingų komponentų kūrimui. Šiame darbe pateikiama XBRL finansiniu standartu grindžiama sprendimų paramos sistema kartu su jos realizavimo scenarijumi, taip pat pateikiama projektavimo ir kūrimo metodika tokiai sistemai bei aprašomas sukurtas jos prototipas.

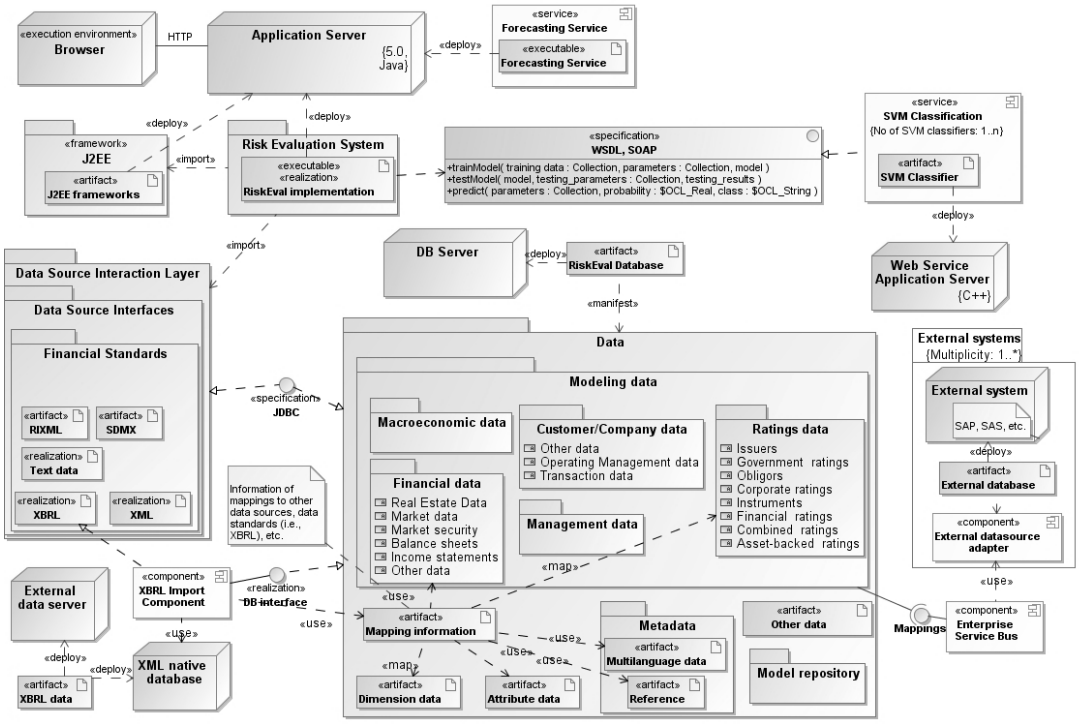


4 pav. Sukurto SPS karkaso sluoksninė diagrama

Sukurtas karkasas gali būti apibūdinamas kaip daugiasluoksnis komponentinis modelis, kurio kiekvienas komponentas aprašo tam tikro domeno ar jo aspekto

funkcionalumą. Tokia struktūra leidžia atskirti skaičiavimo bei domeno funkcionalumą, leidžia pakartotinai panaudoti bei adaptuoti sukurtą funkcionalumą kitose sistemose ar srityse. Be ankstesniuose darbuose pateiktų pagrindinių komponentų – modelių saugyklos, duomenų saugyklos, vartotojo sąsajos bei veiklos logikos – šis karkasas apima ir automatizuotą duomenų gavimą bei apdorojimą, ką įgalina modernūs XML standartai.

Siūlomas karkasas sudarytas iš trijų pagrindinių sluoksnių – SVM grindžiamo automatinio mokymosi sluoksnio (SVM-ML sluoksnis), apimančio automatinio mokymosi algoritmus bei metodus bei šio proceso palaikymui reikalingus procesus, tokius, kaip informacijos apdorojimas bei atvaizdavimas; duomenų sluoksnio, aprašančio modeliavimui reikalingus duomenis bei realizuotus modelius bei kredito rizikos vertinimo sluoksnis (CRE sluoksnis), kuriame realizuotas visas analitinė, modeliavimo, prognozavimo bei vertinimo logika, o taip pat ir vizualizacijos funkcionalumas, siejamas su kredito rizikos statistiniu bei matematiniu vertinimu. Toks aspektų atskyrimas leidžia sukurtus ir/ar integruotus modeliavimo algoritmus pakartotinai panaudoti kitose sistemose.



5 pav. Siūloma pagal sukurtą karkasą kuriamos sistemos realizavimo diagrama

Kiti šiame karkase aprašomi sluoksniai yra sąsajos su duomenų šaltiniais, atvaizdavimo ir informacijos apdorojimo sluoksniai, apibrėžiami abiejuose SVM-ML ir

CRE sluoksniuose ir aprašantys atitinkamus, realizuojančius specifinį tiems sluoksniams funkcionalumą, poaibius, taip pat finansinės analizės modeliavimo ir prognozavimo moduliai, apibrėžti kredito rizikos funkcionalumo sluoksniui.

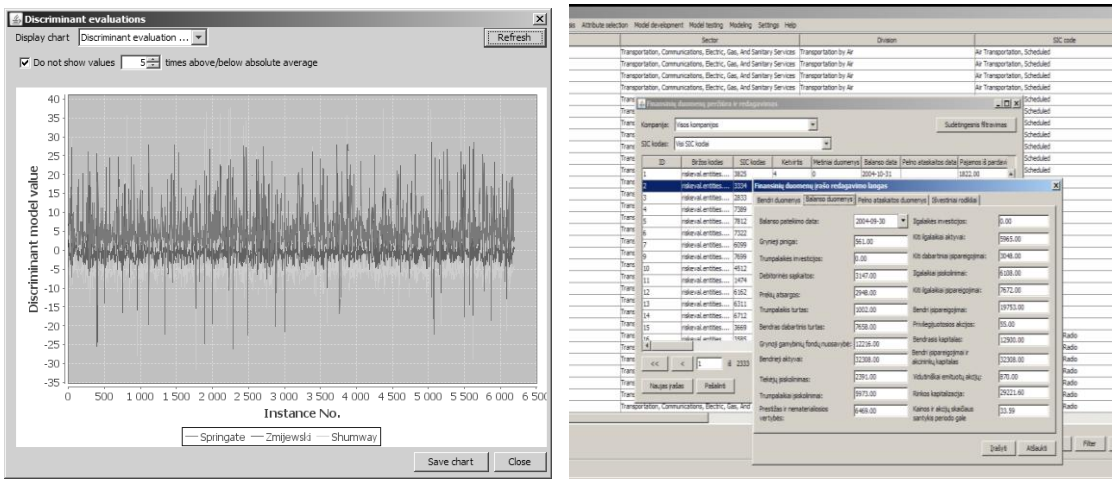
5 pav. pateikiama UML diegimo diagrama kaip siūlomas sukurtu karkasu grindžiamos sistemos realizavimo scenarijus, kuriame atspindimos vykdymo aplinkos bei infrastruktūra, taip pat technologijos, kuriomis ši sistema gali būti realizuota. Atsižvelgiant į SVM realizacijų heterogeniškumą, sistema suprojektuota kaip išskirstyta sistema, realizuojama, naudojant Interneto paslaugas. Realizavimui pasirinktos JAVA technologijos; kaip parodė susijusių technologijų analizė; šia kalba parašyti programiniai karkasai automatiniam mokymuisi ir XBRL funkcionalumui yra brandžiausi ir geriausiai išbaigti.

Išskiriami tokie šio realizacijos modelio ypatumai:

- Nepriklausomas nuo platformos (angl. *Cross-platform*), atsirandantis iš XBRL ir kitų XML grindžiamų standardų bei JAVA technologijų nepriklausomumo nuo platformos; C++ kalbai taip pat sukurta daugiaplatforminių programinių karkasų, tokių, kaip Qt (Digia, 2012);
- Nepriklausomas nuo duomenų šaltinio – objektų išlaikymo (angl. *object persistence*) technologijos naudojimas naudojant tarpinę objektiškai orientuotų užklausų technologiją leidžia realizuoti sistemą beveik nepriklausomai nuo duomenų šaltinio.

Darbe aprašoma ir SPS projektavimui adaptuota projektavimo bei kūrimo metodika, grindžiama domenu grindžiamu projektavimu (angl *Domain Driven Design*, sutr. DDD) (Evans, 2003) ir savybėmis grindžiamu kūrimu (angl. *Feature Driven Development*, sutr. FDD)(Palmer, Felsing, 2002). Domenu grindžiamas projektavimas gerai tinkamas kelis domenus, kontekstus ir/ar aspektus apimančių sudėtingų sistemų kūrimui, todėl jis pasirinktas kaip pagrindas SPS, grindžiamos siūlomu karkasu, projektavimui. Į savybes orientuotas kūrimas yra viena iš judriojo kūrimo metodikų, pilnai suderinama su iteratyviu ir augančiu modeliais grindžiamu programinės įrangos kūrimo procesu; tai gerai tinka sistemų, savyje integruojančiu automatinio mokymosi bei statistinius metodus, komponentų, klasių ar modulių kūrimui, kadangi kiekviena savybė kuriama iteratyviai ir dažnai. Disertacijoje ir (Danenas, Garsva, 2012) šaltinyje pateikiamas išsamesnis šios metodikos bei jos sąsajos su DDD tyrimas siūlomo karkaso

atžvilgiu.



6 pav. Sukurto SPS prototipo grafinės sąsajos iliutracijos

Einamuoju momentu sukurtas sistemos prototipas, realizuojantis dalį duomenų apdorojimo ir analizės galimybių, naudojant aprašytus karkasą, projektavimo metodiką bei realizavimo scenarijų. Šis prototipas naudoja SEC EDGAR duomenų bazę, apimančią 9365 JAV kompanijų duomenis. Realizavimui naudota PostgreSQL DBVS; didžioji dalis duomenų apdorojimo ir sukurtų modelių funkcionalumo realizuota, naudojant Weka karkasą. Kartu pateikiamos sukurtų metodų eksperimentinės realizacijos Java kalba, kurios gali būti integruotos į siūlomą prototipą. 6 pav. pateikiamos sukurto prototipo vartotojo sąsajos iliustracijos.

Disertacijos išvados

1. Ankstesnių skaitinio intelekto metodų ir jų tyrimų kredito rizikos srityje analizė leido suformuluoti tokias išvadas:

a. Einamuoju momentu SVM metodai dominuoja tarp statistinių ir dirbtinio intelekto metodų tyrimų kredito rizikos srityje ir įprastai parodo panašius ar geresnius rezultatus nei panašūs statistiniai ar skaitinio intelekto metodai, todėl yra perspektyvūs ir tirtini.

b. Skirtingų autorių, tiriančių skaitinio intelekto metodus kredito rizikos srityje, gautus rezultatus palyginti yra sudėtinga, kadangi jie priklauso nuo eksperimente naudojamų duomenų, eksperimente naudojamų metodų, pačių skaitinio intelekto metodų realizacijų, eksperimento konfigūracijos, išteklių, kurie buvo prieinami eksperimento metu.

c. Hibridiniai SVM grindžiami modeliai įprastai parodo panašius ar geresnius

rezultatus nei panašūs statistiniai, ekonometriniai ar atskiri SVM modeliai.

d. Atraminių vektorių mašinomis grindžiamas metodas turi kelis privalumus, lyginant su panašiais metodais, tokius, kaip sąlyginai paprasta architektūra, galimybė išvengti persimokymo ir per didelio prisitaikymo prie duomenų, patekimo į lokalius minimumus išvengimas, didelis sukurtų algoritmų ir jų realizacijų skaičius. Pagrindiniai jo trūkumai yra sudėtingas optimalių parametrų parinkimas, brandaus programinio paketo, apimančio visus ar bent daugumą SVM metodų, kas apsunkina jų tyrimą bei pritaikymą.

e. Finansinių atributų, kuriuos galima panaudoti modelio savybių vektoriaus formavimui, kiekis nėra apribotas (analizės metu išskirta 20 tokių rodiklių grupių); įvairūs tyrimai naudoja skirtingus šių atributų poaibius. Statistiškai reikšmingų rodiklių atrinkimui gali būti naudojamas požymių atrinkimo procedūros žingsnis.

2. Ankstesnių tyrimų kredito rizikos srityje ir sprendimų paramos sistemų analizė leido suformuluoti tokias išvadas:

a. Kredito rizikos vertinimo procesas yra daugiamatis procesas, kadangi kredito rizikos analizė gali būti taikoma skirtinguose lygmenyse: finansinis sektorius (grupė, šaka), finansinių duomenų periodiškumas (ketvirtis/metai), skolininko tipas (individualus asmuo, kompanija, vyriausybė), reitingo tipas, valiuta, globalumas (nacionalinis, tarptautinis), periodas (trumpas, vidutinis, ilgas). Tokiam vertinimui sukurtas su sukurtais metodais suderinamas daugiamatės analizės modelis rezultatų analizei.

b. Diskriminantinės analizės ir logistinės regresijos metodai dažniausiai naudojami bei taikomi realių klasifikavimo uždavinių sprendimui kredito rizikos srityje. Altman, Springate, Zmijewski, Ohlson modeliai nagrinėtoje literatūroje minimi kaip populiariausi ir dažniausiai naudojami.

c. Keletas autorių pateikia sprendimų paramos sistemų klasifikacijas, tačiau nei vienoje iš jų neatspindima SPS su integruotu duomenų importu ar integracija su standartuose pateikiamais metaduomenimis. Atsižvelgiant į tai, Holsapple pasiūlyta jungtinio į duomenų bazę bei uždavinių sprendimą orientuotos SPS modelio struktūra praplėsta automatiniu duomenų integravimu bei modelių atnaujinimu, suformuojant panašią į agentinės sistemos architektūrą, ir pateikiama kaip intelektualios SPS, kurios karkasas aprašomas šiame darbe, architektūros pasirinkimas.

d. Atlikus Basel bankinio reglamentavimo standarto atitinkamų dalių analizę, apibrėžti SPS pagrindiniai saugumo, saugomų duomenų, auditavimo prieigos bei saugojimo infrastruktūros reikalavimai, taip pate reikalavimai patiems reitingams.

e. Finansiniai standartai, tokie, kaip XBRL, RIXML tampa svarbūs finansiniam reglamentavimui, kadangi jie pateikia aiškią, išplečiamą, adaptyvią bei struktūrizuotą priemonę finansinei atskaitomybei, duomenų perdavimui ir atvaizdavimui. XBRL apibrėžia ir validavimo bei išvestinių rodiklių taisykles, kurios gali būti išnaudotos finansinių sprendimų SPS duomenų integralumo užtikrinimui ar naujų kintamųjų išvedimui modeliavimui. Ankstesniuose darbuose, kuriuose aprašomos sprendimų paramos sistemos kredito rizikos ar finansinės analizės sritims, šis aspektas nebuvo nagrinėjamas, todėl šiame darbe kaip siūlomos SPS karkaso dalis pateikiamas ir XBRL susiejimų su SPS modelis.

f. Sprendimų paramos sistemų klasifikacijų bei SPS kredito rizikos vertinimui analizė leido identifikuoti esminius šių sistemų komponentus – modelių saugykla, išvadų darymo (veiklos logikos) komponentas, žinių bazė ir duomenų saugykla.

3. Sukurtas spiečiaus optimizavimo metodu ir tiesinėmis atraminių vektorių mašinomis grindžiamas hibridinis PSO-LinSVM metodas, pasižymintis įvairiomis savybėmis:

a. Labiau adaptuotas darbui su įvairaus dydžio duomenų rinkiniais, nei panašūs netiesiniai SVM metodai.

b. Automatinis klasifikatoriaus parinkimas su sudėtingumo ir nuokrypio parametrais iš aibės SVM klasifikavimo metodų su tais pačiais parametrais.

c. Paprastesnis konfigūravimas nei genetinio algoritmo atveju.

4. Eksperimentinis PSO-LinSVM tyrimas parodė, kad:

a. Eksperimentai, atlikti su Vokietijos ir Australijos duomenų rinkiniais, parodė, kad PSO-LinSVM gali parodyti geresnius rezultatus, lyginant su panašiais optimizuotais tiesiniais ir netiesiniais klasifikavimo metodais.

b. Gauti geriausi rezultatai, atsižvelgiant į atvejų atskyrimo kokybę, išreikštą jautrumo ir specifiškumo suma.

c. Eksperimentiniai rezultatai parodė, kad PSO paieška PSO-LinSVM algoritme padengė didelę galimų sprendimų erdvę, dėl ko buvo gauti įvairūs

klasifikatoriai, lyginant su panašiu GA-LinSVM metodu, kur dominavo vienintelis klasifikavimo metodas.

5. Sukurti metodai FS-SVM^{DA} ir FS-SVM^{SWTest}, naudojant išorinius/ekspertinius vertinimus, klasifikavimo metodus ir slenkančio lango testavimą. Sukurti metodai pasižymi tokiomis savybėmis:

a. Reikšmingų rodiklių atrinkimas, naudojant požymių atrinkimo algoritmus.

b. Antrojo metodo atveju testavimas atliekamas keliems sekantiems periodams; tai leidžia užtikrinti, kad sukurtas modelis tinkamas ne tik sekančiam periodui, bet ir daug ilgesniam laikotarpiui.

6. Eksperimentinis FS-SVM^{DA} ir FS-SVM^{SWTest} su slenkančio lango testavimu tyrimas parodė, kad:

a. Gauti rezultatai skirtingi įvairiems sektoriams, taip pat ir priklauso nuo naudojamo duomenų rinkinio.

b. Eksperimentiniai FS-SVM^{DA} rezultatai, naudojant duomenis iš visų sektorių, parodė, kad vidutinis tikslumas buvo virš 80% tiesinių klasifikavimo metodų atveju, ir virš 86% C-SVC grindžiamų klasifikavimo metodų atveju (modeliavimui naudotas Altman modeliu grindžiamas įvertinimas). Šie klasifikavimo modeliai nebuvo efektyvūs, vertinant pokyčius.

c. SVM pagrindu sukurti klasifikavimo metodai SMO, esminių vektorių mašinos (angl. *Core Vector Machines*), sferinių vektorių mašinos (angl. *Ball Vector Machines*), mySVM gali būti gera alternatyva bei parodyti rezultatus, lygintinus ar geresnius nei įprastos tyrimuose naudojamos realizacijos (pavyzdžiui, LibSVM ar SVM^{Light}). Jų detalų tyrimą riboja realizacijų heterogeniškumas bei jų pateikiamos detalios apmokymo ir testavimo informacijos trūkumas.

d. FS-SVM^{SWTest}, naudojant PSO-LinSVM klasifikavimui, tyrimo rezultatai parodė, kad jo naudojimas leidžia pasiekti aukštus klasifikavimo rezultatus (virš 90%), tačiau tikslumas, imant atskirus atvejus, buvo skirtingas ir priklausė nuo sektoriaus bei naudojamo ekspertinio įvertinimo.

7. Suprojektuotas ir pasiūlytas UML diagramomis aprašomas intelektualios SPS, grindžiamos SVM ir XBRL ir suderinamos su siūlomais metodais, karkasas kūrimui:

a. Jį sudaro 5 sluoksniai, atitinkantys svarbiausius SPS komponentus:

duomenų, veiklos (srities) logikos, modelių (automatinio mokymosi) bei atvaizdavimo, taip pat sąsajos su duomenų šaltiniais kartu su atvaizdžių informacija.

b. Sukurto karkaso struktūra leidžia jį pakartotinai panaudoti (pilnus komponentus ar dalinai) panašių ar skirtingų problemų sprendimui, kadangi tokia struktūra aiškiai atskiria įvairius sistemos aspektus (probleminę dalį, funkcionalumą) kaip komponentus.

c. Pasiūlytas techninės realizacijos scenarijus, apibrėžiantis nepriklausomą nuo platformos ir duomenų šaltinio siūlomos sistemos kūrimą, kaip UML realizacijos diagrama.

d. Pasiūlyta metodologija intelektualios SPS kūrimui, naudojant šį karkasą, grindžiama domenu grindžiamu projektavimu ir savybėmis grindžiamu kūrimu, kartu su architektūriniais šio karkaso modeliais.

LITERATŪROS SĄRAŠAS

1. Altman E. I.. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. The Journal of Finance, Vol. 23(4), 1968, pp. 589–609.
2. Altman E. I. Predicting financial distress of companies: Revisiting the Z-score and Zeta models (2000), http://www.defaultrisk.com/pp_score_14.htm, last accessed 2012.08.08.S. J. Russell, P. Norvig. Artificial Intelligence: A Modern Approach. Prentice-Hall, 1995.
3. Anderson R. The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation, Oxford University Press Inc., New York, 2007.
4. Balthazar L. From Basel 1 to Basel 3: The Integration of State-of-the-Art Risk Modeling in Banking Regulation. Palgrave Macmillan, 2006.
5. Ben-Hur A., Ong C.S., Sonnenburg S., Schölkopf B., Rätsch G. (Support Vector Machines and Kernels for Computational Biology. PLoS Computational Biology Vol. 4(10), 2008.
6. Buzius G., Danenas P., Garsva G. Credit risk evaluation using SVM and Bayesian classifiers. Proceedings of the 15th Conference for Master and PhD students "Information Society and University studies", Kaunas, Lithuania, 2010, pp. 27-32.
7. Chang C.-C., Lin C.-J. LIBSVM : a library for support vector machines. ACM Transactions on Intelligent Systems and Technology, Vol. 2(3), 2011, Article 27.
8. Chang, E. Y., Zhu, K., Wang, H., Bai, H., Li, J., Qiu, Z., Cui, H. PSVM: Parallelizing Support Vector Machines on Distributed Computers. Advances in Neural Information Processing Systems, Vol. 20, 2007.
9. Chong W., Yingjian G., Dong W. Study on Capital Risk Assessment Model of Real Estate Enterprises Based on Support Vector Machines and Fuzzy Integral. Proceedings of Control and Decision Conference, 2008, pp. 2317-2320.
10. Cortes C., Vapnik V. Support-vector networks. Machine learning, Vol. 20, No. 3, 1995, pp. 273–297.
11. Crammer K., Singer Y. On the learnability and design of output codes for multiclass problems. Proc. of the 13th Annual Conference on Computational Learning Theory, Vol. 28, 2000, pp 35–46.
12. Danenas P., Garsva G. Credit risk evaluation using SVM-based classifier. Lecture notes in business information processing, Berlin, Springer, Vol. 57, Part 1, 2010, pp. 7-12.
13. Danenas P., Garsva G. Domain Driven Development and Feature Driven Development for Development of Decision Support Systems. Information and Software Technologies: Proceedings of 18th International Conference (ICIST 2012), Communications in Computer and Information Science, Vol. 319, Part 4, 2012, pp. 187-198.
14. Danenas P., Garsva G., Gudas S. Credit Risk Evaluation Model Development Using Support Vector Based Classifiers. Procedia Computer Science, Vol. 4, Elsevier, 2011, pp. 1699-1707.
15. Danenas P., Garsva G. Credit Risk Modeling of USA Manufacturing Companies Using Linear SVM and Sliding Window Testing Approach. Lecture Notes in Business Information Processing, Vol. 117, Part 8, 2012, pp. 249-259.
16. Danenas P., Garsva G. Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach. Procedia Computer Science, Vol. 9, 2012, pp. 1324 – 1333.
17. Danenas P., Garsva G. PSO-based Linear SVM Classifier Selection for Credit Risk Evaluation Modeling Process. Proceedings of 14th International Conference on Enterprise Information Systems (ICEIS 2012), Vol. 1, 2012, SciTePress.
18. Danenas P., Garsva G. Support Vector Machines and their Application In Credit Risk Evaluation Process. Transformations in Business & Economics, Vol. 8, No. 3 (18), 2009, pp. 46-58.
19. Danenas P., Garsva G. SVM and XBRL based decision support system for credit risk evaluation. Proceedings of the 17th International Conference on Information and Software Technologies (IT 2011), Technologija, Kaunas, Lithuania, 2011, pp. 190-198.
20. Debnath R., Takahide N., Takahashi H. A decision based one-against-one method for multi-class support vector machine. Pattern Analysis and Applications, Vol. 7(2), 2004, pp. 164-175.
21. Elizalde A. Credit risk models II: structural models. Working Papers wp2006_0606, CEMFI, 2006.

22. Elizalde A. Credit Risk Models III: Reconciliation Reduced-Structural Models. Working Papers wp2006_0607, CEMFI, 2006.

23. Engelbrecht A. Computational intelligence: an introduction, 2nd ed., Wiley & Sons Inc., 2007.

24. Evans, E. Domain-Driven Design: Tackling Complexity in the Heart of Software. Addison Wesley, 2003.

25. Fan R., Chang K., Hsieh C., Wang X., Lin C. LIBLINEAR: A library for large linear classification. The Journal of Machine Learning Research, Vol. 9. 2008, pp.1871–1874.

26. Garsva G., Danenas P. XBRL Integration Into Intelligent System For Credit Risk Evaluation. Transformations in Business & Economics, Vol. 10, No. 2 (23), 2011, pp. 88-103.

27. Hamilton A. IT62 Decision Support Systems. Lectures. University of Stirling 2004.

28. Hao Y., Chi Z., Yan D. Fuzzy Support Vector Machine Based on Vague Sets for Credit Assessment. FSKD '07 Proceedings of the Fourth International Conference on Fuzzy Systems and Knowledge Discovery, Vol. 1, 2007, pp. 603–607.

29. Hao P.-Y, Lin M.-Sh., Tsai L.-B. A New Support Vector Machine with Fuzzy Hyper-Plane and Its Application to Evaluate Credit Risk. 2008 Eighth International Conference on Intelligent Systems Design and Applications, Vol. 3, 2008, pp.83-88

30. Hsu C., Lin C. A simple decomposition method for support vector machines. Machine Learning, Vol. 46, 2002, pp. 291–314.

31. Yang Z. R. Biological applications of support vector machines. Brief Bioinformatics, Vol. 5(4), 2004, pp. 328-338.

32. Joachims Th. SVMlight - Support Vector Machine, http://www.cs.cornell.edu/people/tj/svm_light, last accessed 2012.07.31.

33. Joachims, T. Text categorization with support vector machines: learning with many relevant features. Proceedings of ECML-98, 10th European Conference on Machine Learning, 1998, pp. 137–142.

34. Kancerevyčius G. Finansai ir investicijos (in Lithuanian). Kaunas, Smaltija, 2004.

35. Kaski S., Sinkkonen J., Peltonen J. Bankruptcy Analysis with Self-Organizing Maps in Learning Metrics. IEEE Transactions on Neural Networks, 2001.

36. Kennedy J. The Particle Swarm: Social Adaptation of Knowledge. Proceedings of the IEEE International Conference on Evolutionary Computation, 1997, pp. 303–308.

37. Khan A., Baharudin B., Lee L.H., Khan Kh. A Review of Machine Learning Algorithms for Text-Documents Classification. Journal of Advances in Information Technology, Vol 1, No 1, pp. 4-20, 2010, doi:10.4304/jait.1.1.4-20.

38. Lv G.-L., Peng L. Commercial Banks' Credit Risk Assessment Based on Rough Sets and SVM. Proceedings of 4th International Conference on Wireless Communications, Networking and Mobile Computing, 2008, pp. 1-4.

39. Mahotra R., Mahotra D.K. Differentiating between good credits and bad credits using neuro-fuzzy systems. European Journal of Operational Research, Vol. 136, 2002, pp. 190-211.

40. Merkevicius E., Garsva G., Simutis R. Neuro-discriminate Model for the Forecasting of Changes of Companies Financial Standings on the Basis of Self-organizing Maps. Lecture Notes In Computer Science, Vol. 4488, 2007, pp. 439-446.

41. Ohlson J. Financial Ratios and the Probabilistic Prediction of Bankruptcy. Journal of Accounting Research Vol. 18(1), 1980, pp. 109-131.

42. Palmer S. R., Felsing, J. M. A Practical Guide to Feature-Driven Development. Prentice Hall, 2002.

43. Ping Y., Yongheng L. Neighborhood rough set and SVM based hybrid credit scoring classifier. Expert Systems with Applications, Vol. 38, No. 9, 2011, pp. 11300–11304.

44. Piramuthu S. Financial credit-risk evaluation with neural and neurofuzzy systems. European Journal of Operational Research, Vol. 112, 1999, pp. 310-321.

45. Platt J. Sequential minimal optimization: A fast algorithm for training support vector machines. Advances in Kernel Methods-Support Vector Learning, 1999, pp. 185 – 208.

46. Power D. J. Decision Support Systems: A Historical Overview. In: Burstein F., Holsapple C. W. (eds). Handbook on Decision Support Systems 1: Basic Themes. Springer-Verlag Berlin Heidelberg, 2008.

47. Raynor W. J. The international dictionary of artificial intelligence. The Glenlake Publishing Company, Ltd., 1999.

48. Saunders A., Allen L. Credit risk measurement: new approaches to value at risk and other paradigms. John Wiley & Sons, Inc., 2002.

49. Schölkopf B., Smola A., Williamson R., Bartlett P. New support vector algorithms. Neural Computation, Vol. 23, No. 1, 2000, pp. 60–73.

50. Shumway T. Forecasting bankruptcy more accurately: A simple hazard model. Journal of Business, Vol. 74(1), 2001, pp. 101–124.

51. Sindhwani V., Keerthi S. S. Newton methods for fast solution of semi-supervised linear SVMs. Large Scale Kernel Machines (eds. L. Bottou, O. Chapelle, D. DeCoste, J. Weston), MIT Press, Cambridge, MA, 2007.

52. Springate G. L. V. Predicting the Possibility of Failure in a Canadian Firm, Unpublished M.B.A. Research Project, Simon Fraser University, 1978.

53. Suykens J., Vandewalle J. Least Squares Support Vector Machine Classifiers. Neural Processing Letters, Vol. 9, No. 3, 1999, pp. 293–300.

54. Tsang I. W., Kocsor A., Kwok J. T. Simpler core vector machines with enclosing balls. Proceedings of the Twenty-Fourth International Conference on Machine Learning, Corvallis, USA, 2007, pp.911-918.

55. Tsang W., Kwok J. T., Cheung P. M. Core vector machines: Fast SVM training on very large data sets. Journal of Machine Learning Research, Vol. 6, 2005, pp.363-392.

56. UCLA-LoPucki Bankruptcy Research Database, http://lopucki.law.ucla.edu/index.htm

57. van den Berg J. Credit Rating Prediction with Self-Organizing Maps. Expert Systems with Applications, Volume 30, Issue 3, April 2006, Pages 479-487.

58. van Gestel T., Baesens B. Credit Risk Management: Basic Concepts. Oxford University Press, USA, 2009..

59. Vapnik V. N. Statistical learning theory. New York: Wiley, 1998.

60. Vellido A., Lisboa P. J. G., Vaughan B. Neural networks in business: a survey of applications (1992 – 1998), Expert Systems with Applications, Vol. 17, 1999, pp. 51-70.

61. Wang B., Liu Y., Hao Y., Liu Sh. Defaults Assessment of Mortgage Loan with Rough Set and SVM. International Conference on Computational Intelligence and Security (CIS 2007), 2007, pp.981-985.

62. Wang B., Wang D., Liu Sh., Hao Y. Research of Housing Loan Credit Evaluation Based SVM. 2008 Fourth International Conference on Natural Computation, Vol. 2, 2008, pp.144-147.

63. Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka, last accessed 2012.07.31.

64. Wong B. K., Lai V. S., Lam J. A bibliography of neural network business applications research: 1994-1998, Computers & Operations Research, Vol. 27, 2000, pp. 1045-1076.

65. Wong B., Selvi Y. Neural network applications in finance: A review and analysis of literature (1990 - 1996), Information & Management, Vol. 34, 1998, pp. 129-139.

66. Zhou J. , Tian J . Credit risk assessment based on rough set theory and fuzzy support vector machine. Advances in Intelligent Systems Research, ISKE-2007 Proceedings. Atlantis Press, 2007.

67. Zmijewski M. Methodological Issues Related to the Estimation of Financial Distress Prediction Models. Journal of Accounting Research Vol.22, 1984, pp. 59-82.

INTRODUCTION

Financial crisis of 2008 and 2010 identified the need of more precise techniques for credit risk evaluation, together with tools necessary to develop them. The main objective of credit risk is computing possibility that debtor will fail to meet his obligations before by agreed terms, which helps to reduce the probability to lose invested money. Minimization of such debts is critical for managing risk and optimal capital allocation in financial institutions as Basel II capital accord defines new regulatory standards which have to be met. Thus proper, efficient and effective credit risk evaluation tools for credit risk, such as highly discriminative credit scoring models, are obligatory for every financial institution. This problem is solved in multiple dimensions, including debtor type (individual, organization, government), financial instrument type (loan issue, financial derivatives), modelling techniques (parametric, non-parametric, VaR, probability default and etc.), length and others. Classification technique, with emphasis to associate an obligor with one of risk classes or identify whether it is tend to bankruptcy, is one of the most popular techniques widely applied and discussed in various papers, with various modern and complex techniques including statistical, econometric and artificial intelligence based techniques. Support Vector Machines (abbr. as SVM) at the moment of writing is one of most widely developed, researched and applied techniques in this field, proposed by Vapnik and further developed or discussed in books and papers by Scholkopf et al., Cristianini et al., Baesens, van Gestel et al., Mangasarian et al., Huang et al., Lai, Yu et al., Chang et al., Steinwart et al. and etc. It is used to solve various classification problems in different domains, including bioinformatics and computational biology, document classification and etc. Various approaches which combine SVM with other techniques or apply inner modifications for initial SVM algorithm in order to obtain faster, more accurate and efficient solutions are permanently proposed; they are also reviewed further in this work. Balthazar refers that SVM-based model is used by Standard & Poors rating company. Yet complexity of this technique, various analytical, computational and development issues makes this task more sophisticated; this is one of the reasons which make SVM “a classification technique for experts”.

Research problem

As it is mentioned in the introduction, this topic is widely researched and important for financial institutions, although development of SVM-based algorithms is important for the whole computational science. The results of this research can be applied in practice, including integration of developed techniques into intelligent decision support system for both scientific and business purposes. Increasing amount of available open and linked data (including financial) offers new possibilities to develop new models or improve existing by integrating new knowledge within them, combining available expert knowledge and experience with this data. Integration with various Semantic Web technology based standards, especially the ones from financial domain, becomes important in this context. Thus another important aspect in such research is lack of research aimed at intelligent financial decision support system development, including both integration of financial standards and frameworks such as Extensible Business Reporting Language (XBRL), and development methodologies for similar systems. These standards are also supported by Basel II regulatory standard therefore their application can be viewed as necessity for similar future decision support. Therefore, design and research of resulting DSS framework is also relevant for software engineering science, offering new viewpoints for engineering of complex modern decision support systems, which can be further developed and enhanced.

The object of research

This work analyzes intelligent credit risk evaluation techniques based on Support Vector Machines. Therefore, the main object of this work is hybrid Support Vector Machines based classification techniques for credit risk evaluation and bankruptcy prediction. A framework comprising such and similar techniques, financial standards, design and development methodology for intelligent systems based on these techniques, possible implementation scenario is also defined as secondary research object.

The goal and objectives of the research

The primary objective of the research is to propose an approach to develop Support Vector Machines classification based classifier for credit risk evaluation which combines existing financial data and external evaluations (e.g., expert evaluations) available. Secondary objective is to propose and develop a framework for intelligent

decision support systems for financial domain which integrates financial standards, a solution based on proposed classifier, design and development methodology, together with main components which are common for such type of DSS.

The objectives of the dissertation are as following:

1. Analyse statistical, econometric and artificial intelligence techniques, current developments and previous works in credit risk domain based on these techniques, identify their main advantages.
2. Develop hybrid intelligent classification method and/or approach, based on artificial intelligence techniques, for researched problem.
3. Analyse developed structures of decision support systems for researched field, financial standards and regulations, their purpose and fields of application, identify and propose possible ways of their integration and application in intelligent decision support system for credit risk domain.
4. To carry out experimental evaluation of developed techniques, analyse and evaluate obtained results.
5. Design and develop a framework for intelligent decision support system for credit risk evaluation which includes developed approaches, components, common for such systems, integration of financial standards, design, development and implementation scenarios.
6. Implement a decision support system using the designed framework.

Research methodology and tools

The following methods were used in the research: general cognition (formulation of research tasks and aims of research, collection and analysis of information; generalization; formulation of conclusions); general scientific research techniques such as induction, deduction, comparison (techniques, characteristics, similarities, differences); data analysis and modelling; structuring, grouping, generalization, abstraction and presentation.

Open source machine learning frameworks WEKA, SVM toolboxes LibSVM, LIBLINEAR were used to implement the algorithms and techniques presented in this research. These tools, together with RapidMiner and various SVM implementations, were also used in research for benchmarking implementations. Technical computing

system MATLAB was used for initial developing, modelling and testing PSO-LinSVM algorithm. Subsets of SEC EDGAR database, which comprises financial ratios from yearly and quarterly balance and income statements in 1999-2008 of 9365 USA based companies from 9 sectors, UCLA LoPucki bankruptcy database, which contains actual bankruptcy data of 911 USA bankruptcy companies (with 253 companies directly mapped to EDGAR database used in the research), and Australian and German credit datasets from UCI machine learning repository (with 690 and 1000 instances respectively) were used for experimental research of developed algorithms and techniques. UML and BPMN notations were used for framework and method design; diagrams for design and development methodology were prepared on custom notation based on UML and recommendations of Domain Driven Design author given in his book. Graphical modelling tools such as MagicDraw and Microsoft Visio were used to develop the diagrams.

The statements of the thesis

1. Particle Swarm Optimization and linear Support Vector Machines based classifier, which can automatically select optimal classifier together with its parameters from a set of classifiers with the same set of parameters, can perform efficiently with both small and large datasets.

2. The developed classification technique, comprising feature selection, SVM classification and sliding window testing principle, can be used to develop and test classification model for credit risk domain.

3. Integration of XBRL financial standard to decision support for financial domain can improve model development process with additional data variables, enable automated import of standardized and structured financial data, real-time model development and update.

4. A framework integrating proposed techniques, financial standards, design and development methodology, possible implementation scenario based on cross-platform and data source independency is an important tool to develop modern intelligent decision support systems for credit risk evaluation.

Scientific significance of this work

New hybrid classification technique PSO-LinSVM, which uses particle swarm

optimization based procedure for automatic selection of linear SVM classifier, is proposed in this work. Differently from techniques proposed previously, this algorithm selects linear SVM classifier together with its complexity and bias parameters from a set of linear SVM classifiers with these parameters. Its classification efficiency is tested with datasets of various sizes. Proposed technique can be used to solve classification problems in various domains such as finance, text analysis, bioinformatics and etc. This work also proposes credit risk evaluation technique, based on discriminant models or external evaluations, feature selection, classification and sliding window testing approach. Differently from previous techniques, this approach enables testing of developed model using data from one or more sequential periods which helps to evaluate its performance in several periods. Research context, such as large amount and dimensionality of used data, integration with external data sources and standards is also important as it is not typical for such research but becomes relevant as the number of available data sources and amounts of data tend to rise. Proposed framework for decision support system, together with design and development methodology, are important for software engineering science as they describe framework for development of distributed component and computational intelligence based systems, main components and processes, using researched DSS for credit risk evaluation as case study. This helps to enhance development of such systems using principles of component-based software engineering.

Practical results of proposed work

Proposed techniques might be used to develop models based on external evaluators, exploiting existing data to produce new models or improve existing. They can also be used: to map rating data (internal or external) to existing financial data, identify their inner dependencies and help to improve quality of financial analysis as well as identify key factors; to embed expert knowledge and experience into a particular data-driven model which can be moved to a new environment. Developed intelligent model may also be used to evaluate the instances which cannot be evaluated by the expert either because of the missing data or mathematical problems (e.g., division by zero), as well as improve quality of base evaluator.

The proposed PSO-LinSVM classification technique, based on Particle Swarm

Optimization and linear SVM, can be applied to solve classification problems in any domain. The developed platform and data source independent framework might be used to develop complex, with financial standards integrated, modern decision support system for credit risk evaluation. The described design and development methodology can be adapted and applied to develop other large scale DSS.

Publications

In international journals, which are included in Scientific Master Journal List (ISI):

1. Danenas P., Garsva G. Support Vector Machines and their Application in Credit Risk Evaluation Process. Transformations in Business & Economics (2009), Vol. 8, No. 3 (18), pp. 46-58, ISSN 1648-4460.
2. Garsva G., Danenas P. XBRL Integration Into Intelligent System For Credit Risk Evaluation. Transformations in Business & Economics (2011), Vol. 10, No. 2 (23), pp. 88-103, ISSN 1648-4460.

In proceedings of scientific conferences, indexed in Scientific Master Journal Proceeding List (ISI):

1. Danenas P., Garsva G. Credit risk evaluation using SVM-based classifier. Lecture notes in business information processing (2010), Berlin, Springer, Vol. 57, Part 1, pp. 7-12, ISBN 978-3-642-15401-0
2. Danenas P., Garsva G., Gudas S. Credit Risk Evaluation Model Development Using Support Vector Based Classifiers. Procedia Computer Science, Vol. 4 (2011), Elsevier, pp. 1699-1707, ISSN 1877-0509
3. Danenas P., Garsva G. SVM and XBRL Based Decision Support System for Credit Risk Evaluation. Proceedings of the 17th International Conference on Information and Software Technologies (IT) (2011), Technologija, Kaunas, Lithuania, pp. 190-198, ISSN 2029-0020.
4. Danenas P., Garsva G. Simutis R. Development of Discriminant Analysis and Majority-Voting Based Credit Risk Assessment Classifier. Proceedings of the 2011 International Conference on Artificial Intelligence (ICAI 2011), CSREA Press, Vol.1, pp. 204-209, ISBN: 1-60132-183-X, 1-60132-184-8 (1-60132-185-6).
5. Danenas P., Garsva G. Credit risk evaluation modeling using evolutionary

linear SVM classifiers and sliding window approach. Procedia Computer Science Vol. 9 (2012), Elsevier, pp. 1324 – 1333, ISSN: 1877-0509.

6. Danenas P., Garsva G. Credit Risk Modeling of USA Manufacturing Companies Using Linear SVM and Sliding Window Testing Approach. Lecture Notes in Business Information Processing, Vol. 117, Part 8 (2012), pp. 249-259, ISBN: 978-3-642-30359-3.

7. Danenas P., Garsva G. PSO-based Linear SVM Classifier Selection for Credit Risk Evaluation Modeling Process. Proceedings of 14th International Conference on Enterprise Information Systems (ICEIS 2012), Vol. 1 (2012), SciTePress, ISBN: 978-989-8565-10-5.

8. Danenas P., Garsva G. Domain Driven Development and Feature Driven Development for Development of Decision Support Systems. Information and Software Technologies: Proceedings of 18th International Conference (ICIST 2012), Communications in Computer and Information Science, Vol. 319, Part 4 (2012), pp. 187-198, Springer-Verlag Berlin Heidelberg, ISSN 1865-0929.

In proceedings of other conferences:

1. Buzius G., Danenas P., Garsva G. Credit risk evaluation using SVM and Bayesian classifiers. Proceedings of the 15th Conference for Master and PhD students "Information Society and University studies" (2010), Kaunas, Lithuania, pp. 27-32, ISSN 2029-4824.

2. Danenas P., Garsva G. A model for multidimensional analysis for credit risk evaluation based on intelligent techniques (in Lithuanian). Conference Proceedings of "Information Technology", 16th Conference for Master and PhD students (2011), Kaunas, Lithuania, pp. 49-52, ISSN 2029-249X

3. Galkus E., Danenas P., Garsva G. Application of ensemble classification methods in credit risk evaluation (in Lithuanian). Conference Proceedings of "Information Technology 2012", 17th Conference for Master and PhD students (2012), Kaunas, Lithuania, pp. 70-73, ISSN 2029-249X.

Thesis structure

The thesis contains the introduction (including list of the author's publications), 4 chapters, conclusions (6 chapters in total), list of references and 14 appendixes. The total

volume of the dissertation is 212 pages, including 34 tables, 38 pictures and 8 algorithms. The list of references contains 248 various sources, including books, scientific papers, technical reports, Internet sources.

The work consists of six main parts – introduction, analytical part, methodical part, experimental part, implementation part and conclusions. Each of these parts discusses particular aspects of research and decision support system which is discussed, designed and developed.

- *Introduction (first chapter)* presents research problem, object, aims and objectives, research findings and results, its scientific significance, describes possible practical application, presents information about the papers, in which the main results of the research were published.
- *Analytical part (second chapter)* discusses main concepts and definitions of artificial intelligence, problems that it can solve, main techniques from this field such as artificial neural networks, decision trees, evolutionary and swarm intelligence techniques and etc. with their previous research in credit risk domain as well as various statistical and intelligent feature selection techniques. Support Vector Machines, as technique discussed in this work, is described in more detail in separate section. Main concepts from credit risk domain such as various kinds of other risks, sample ratios, taxonomies of ratings, evaluation techniques are also shortly described. Finally, financial standards, their structure and adoption for financial reporting and evaluation, together with intelligent decision support systems, their taxonomies, main components and use cases for financial decision support are discussed.
- *Methodology part (third chapter)* describes developed techniques and tools for evaluation such as machine learning metrics. It gives generic framework for hybrid model development which generalizes previously made researches and proposes a generic structure for further classifier research. Implemented classification approaches (SVM with feature selection and its extension for sliding window testing based approach), together with PSO-LinSVM and GA-LinSVM classifiers, combining correspondingly Particle Swarm Optimization and Genetic Algorithm techniques with based linear SVM classifier selection from a family of similar classifiers together with its parameters are also described.

- *Experimental part (fourth chapter)* describes experimental research made with developed techniques, presents obtained results and their analysis.
- *Implementation part (fifth chapter)* describes functionality for intelligent decision support system, describes framework for its development with integration of financial standards, together with its design and development methodology, possible implementation scenario and development possibilities. It also describes developed prototype, its current capabilities and future development.
- *Conclusions part (sixth chapter)* presents final conclusions of the dissertation.

At the end of the dissertation there are references, the list of publications and appendixes.

THESIS RESULTS AND CONCLUSIONS

1. The overview of previous research in computational intelligence in credit risk domain has shown that:

a. SVM oriented research at the moment of writing tends to be dominating among research of various statistical and computational intelligence techniques in financial and credit risk domain and usually shows comparable or better results than similar statistical or computational intelligence techniques, thus their research is promising.

b. The results presented by different authors who applied computational intelligence techniques in credit risk domain are difficult to compare, as the results highly depend on data used in the experiment, the approach used in the experiment, implementation of techniques, the resources that were available during the experiment, experiment configuration.

c. Hybrid methods based on SVM tend to show comparable or better performance than similar statistical, econometrical or standalone SVM methods.

d. Support Vector Machines based technique has several advantages over similar techniques, such as comparably simple architecture, ability to avoid overtraining and overfitting. It also has numerous algorithms and implementations. Its main disadvantages are complicated choice of optimal parameters, lack of mature framework combining all or most of these techniques which slows down its research and adoption.

e. There is a large variety of financial attributes (20 groups of such ratios are described in Appendix H of the thesis) which can be used in feature vector formation;

however, all papers use different subsets of such ratios. Therefore, feature selection step can be applied to select most significant features.

2. The overview of previous research in credit risk domain and decision support systems has shown that:

a. Credit risk evaluation process is a multidimensional process, as credit risk analysis can be applied in several dimensions: financial sector (group, industrial code), financial data period (quarter, year), type of obligor (individual, company, government), type of rating, currency, globality (national, international), period (short, medium, long). Therefore, a multidimensional model for result analysis, consistent with the technique in this work, has been developed as a tool for such evaluation.

b. Multiple discriminant analysis and logistic regression techniques are mostly known and applied in real world statistical techniques, with Altman, Springate, Zmijewski, Ohlson models referred as the most popular.

c. Several authors give their classifications of decision support systems; however, none of them describes a DSS with automated data import functionality or integration with metadata that can be provided by some modern standards. Therefore, model of combined database and solver oriented DSS architecture described by Holsapple is extended with automated data integration and model update functionality layer, forming similar to agent-based DSS architecture, and is selected as a choice for architecture of intelligent DSS proposed in this work.

d. Overview of banking regulation standards (particularly Basel standard) identified the key requirements for such system in terms of security, data stored, supervisory access and storage facility, as well as requirements for ratings themselves.

e. Financial standards such as XBRL, RIXML are becoming an important part of financial regulation as they offer clear, extensible, flexible, adaptable and structured framework for financial reporting, data transfer and representation. XBRL enables definition of validation or derived rules which can be applied in decision support to ensure data integrity or derivation of new variables. These standards can be applied to solve data interchange and data quality problems often faced in banking institutions. None of previous work related to development of similar frameworks or decision support systems for credit risk evaluation described possible interfacing with XBRL, thus a mapping model compatible with classification problem researched in this work is also

proposed as part of the proposed framework.

f. The analysis of taxonomies for decision support systems, existing structures for their applications in credit risk domain helped to identify core components of these systems – model repository, inference engine (business logic), knowledge base and data storage facility.

3. A new classifier PSO-LinSVM, based on Particle Swarm Optimization and linear SVM, is proposed with following capabilities:

a. More suitable for large-scale learning than nonlinear SVM technique.

b. Automatic selection of SVM classifier from a family of similar classifiers with the same parameters.

c. Less complex configuration than using other evolutionary techniques, e.g., Genetic Algorithm approach.

d. An option to optimize for either accuracy or TP ratio performance which makes it usable with both balanced and unbalanced data.

4. Experimental research of PSO-LinSVM identified that:

a. According to experiments performed on German and Australian credit datasets, PSO-LinSVM is capable to show better performance compared to similar optimized linear and nonlinear SVM classifiers.

b. It also resulted in best quality of data separation in terms of the sum of sensitivity and specificity.

c. Experimental results showed that it covered a large space of possible solutions and resulted in larger variety of obtained classifiers, compared to similarly developed GA-LinSVM technique, where single classifier dominated.

5. Two approaches for development of classifier for credit risk evaluation using external evaluations – FS-SVM^{DA} and FS-SVM^{SWTest} - are presented and researched in this work. They combine feature selection, classification; FS-SVM^{SWTest} also uses sliding window testing. They have following properties:

a. Both of these approaches were tested on datasets of various sizes which make them suitable for both small and large scale learning.

b. Feature selection step automatically identifies significant ratios.

c. In case of the second technique, the testing is done for next several periods; this helps to ensure that trained classifier is consistent not only with the

following, but also but much larger period.

6. Experimental evaluation of FS-SVM^{DA} and FS-SVM^{SWTest} approach identified that:

a. The results varied on different sectors; therefore it highly depends on the dataset that is used in the research.

b. Experimental results of FS-SVM^{DA} using data from all sectors showed that average accuracy was above 80% for linear SVM based classifiers, and over 86% for C-SVC based classifiers (Altman based evaluator was used in experiments). However, it was not efficient in prediction of evaluator changes.

c. SVM-based classifiers SMO, Core Vector Machines, Ball Vector Machines, mySVM are a good alternative for larger scale learning and show performance comparative or better than standard implementations (e.g., LibSVM or SVM^{Light}). This implies that more attention should be given to these techniques in future research.

d. Results of FS-SVM^{SWTest} approach and PSO-LinSVM as base classifier showed that it is capable of performing classification with high accuracy (over 90%), although accuracy varied, depending on the sector and evaluator used.

e. Application of FS-SVM^{SWTest} approach to actual bankruptcy identification resulted in promising results as ir performed better than original evaluator. Although these results should be treated carefully at the moment, they give a premise to develop an approach for classifier selection with respect to original evaluations and identification performance based on proposed tecghnique.

7. A framework for intelligent DSS based on SVM and XBRL development, consistent with proposed techniques, is designed and described using UML diagrams:

a. It consists of 5 layers which represent most important aspects: data, business (domain) logic, models (machine learning) and representation, as well as data source interaction together with mapping information.

b. It enables reuse for other similar problems (whole components or parts of them) as such structure clearly separates various aspects as components.

c. A possible implementation scenario as UML implementation diagram is also presented; it proposes development of cross-platform and data source independent DSS.

d. A combined methodology based on Domain Driven Design and Feature Driven Development is described and proposed for development of DSS based on suggested framework together with architectural design models for this framework.

CURRICULUM VITAE

Personal data

Name: Paulius Danėnas
Date and place of birth: August 18, 1982, Kaunas
E-mail: paulius.danenas@vukhf.lt

Institution

Vilnius University, Kaunas Faculty of Humanities
Department of Informatics
Muitinės Str. 8, LT-44280, Kaunas, Lithuania

Education background

2008-2012	PhD studies at Department of Informatics, Kaunas Faculty of Humanities, Vilnius University, Kaunas, Lithuania
2006-2008	Master degree of Informatics, Department of Informatics, Kaunas Faculty of Humanities, Vilnius University, Kaunas, Lithuania
2006.10-2007.02	ERASMUS Exchange student at Free University of Bozen-Bolzano, Italy, Computer Science studies
2002-2004	Bachelor degree of Informatics, Department of Informatics, Kaunas Faculty of Humanities, Vilnius University, Kaunas, Lithuania

Academical and professional experience

The author is a certified Microsoft Certified Technology Specialist, Oracle Certified Java 6.0 Programmer and Zend PHP 5.3 Certified Engineer. Since 2008 he has also worked as developer and analyst. In 2011 he attended PhD School in Scientific GPU Computing at Lyngby, Denmark.

Awards

The author was awarded with a scholarship by the Research Council of Lithuania in 2012. His paper "Credit risk evaluation modeling using evolutionary linear SVM classifiers and sliding window approach" received Green Group Award of Computational Finance and Business Intelligence for best paper in this workshop in ICCS 2012 (International Conference on Computational Science) conference at Omaha, Nebraska, USA.