VILNIAUS UNIVERSITETO
KAUNO HUMANITARINIO FAKULTETO

INFORMATIKOS KATEDRA

Verslo informatikos studijų programa
Kodas 62409P104

GEDIMINAS BŪZIUS

MAGISTRO BAIGIAMASIS DARBAS

BAJESO METODO TAIKYMAS KREDITO RIZIKOS VALDYME

Kaunas 2010

VILNIAUS UNIVERSITETO
KAUNO HUMANITARINIO FAKULTETO
INFORMATIKOS KATEDRA

GEDIMINAS BŪZIUS

MAGISTRO BAIGIAMASIS DARBAS

BAJESO METODO TAIKYMAS KREDITO RIZIKOS VALDYME

Leidžiama ginti _____

Magistrantas _____

(parašas)

Darbo vadovas _____

(parašas)

dr. doc. Gintautas Garšva

Darbo įteikimo data _____

Registracijos Nr. _____

Kaunas 2010

TURINYS

ILIUSTRACIJOS

LENTELĖS

SUTRUMPINIMAI

AI/DI – angl. artificial intelligence/liet. dirbtinis intelektas

Branduolys (angl. Kernel) – funkcija, pagal duotus du taškus grąžinanti jų panašumą

BVP – bendrasis vidaus produktas

CPU – centrinis procesorius (kompiuterio)

DB – duomenų bazė

DNN/ANN – dirbtinis neuroninis tinklas (angl. Artificial Neural Network)

DT – sprendimų medis (angl., decision tree)

EAD – turto praradimo dydis įsipareigojimų neįvykdymo atveju (angl., Exposure at default)

EL, UL – numatytasis ir nenumatytasis nuostoliai, atitinkamai angl. expected ir unexcpected loss

Fuzzy logika – neraiškioji (angl. fuzzy) logika

GA – genetinis algoritmas (angl. Genetic Algorithm)

Ghz, Mhz – gigahercas, megahercas

KNN – artimiausio k kaimyno metodas *(*angl., k-nearest neighbor)

LGD – nuostolio ir pozicijos vertės santykis, išreikštas procentais (angl., Loss given default)

MCMC – Markov Chain Monte Carlo (paieškos metodas)

ML – mašininis mokymas (angl. Machine Learning)

NN –neuroninis tinklas (angl. Neural Network)

PD – įsipareigojimų neįvykdymo tikimybė (angl., Probability of default)

RAM - tiesioginės kreipties atmintis (angl., random-access memory)

RBF – radialinės bazės funkcija (angl. Radial Basis Function)

SOM – saveorganizuojantis žemėlapis (angl. Self-Organizing Map)

SQL – užklausų kūrimo kalba, naudojama duomenų bazėse (Structurized Query Language)

SV – atraminis vektorius (angl. Support vector)

SVM – atramos vektorių mašinos (angl. Support Vector Machines)

VaR – vertės pokyčio rizika (angl., Value-at-Risk)

VU KHF – Vilniaus universiteto Kauno humanitarinis fakultetas

ĮVADAS

Kredito rizikos vertinimas yra viena svarbiausių sričių bankininkystėje, kreditų unijų veikloje ir kitose finansų veiklose (pavyzdžiui, akcijų rinkose), kurios pagrindinis uždavinys yra rizikos, atsirandančios kreditoriui išduodant paskolą kuriam nors konkrečiam klientuojamam fiziniam ar juridiniam asmeniui, nustatymas bei įvertinimas, tam naudojant įvairius matematinius metodus bei modelius. Norint įvertinti rizikingumą, naudojami įvairūs modeliai, metodai bei algoritmai, skirti tokių duomenų analizei bei organizacijų klasifikavimui bei rangavimui pagal rizikos grupę. Išsivysčius dirbtinio intelekto sričiai (angl. *Artificial Intelligence*, sutr. AI), atsirado galimybė šioje srityje naudoti ir AI metodus, tokius, kaip neuroniniai tinklai, genetiniai algoritmai, saviorganizuojantys tinklai (*self-organizing maps*), neraiškioji (*fuzzy*) logika, Bajeso, medžių algoritmai bei kiti, o taip pat ir hibridiniai (apimantys dvi ar daugiau paminėtų sričių) metodai.

Tyrimo tema aktuali tuo, kad kreditų rizikos analizė bei valdymas yra svarbūs klausimai finansinėms institucijoms, teikiančioms paskolas tiek verslo objektams, tiek ir individualiems asmenims. Bankų blogai įvertintos rizikos ilgalaikiams kreditams gali privesti jį prie bankroto, o sėkminga bankų veikla parodo ekonominį šalies ar didesnio regiono stiprumą. Tyrimo sritis šiuo metu gana plačiai nagrinėjama visame pasaulyje, galima paminėti tokius šaltinius, kaip „defaultrisk.com“, „gloriamundi.org“, skirtus kredito bei finansinei rizikai, bei „CiteSeer“ ar IEEE duomenų bazę, kurios moksliniai straipsniai glaudžiai siejasi su informacinėmis technologijomis. Šiose mokslinėse duomenų bazėse galima rasti straipsnių, vienu ar kitu aspektu susijusiu su nagrinėjama tema. Lietuvoje ši sritis nagrinėjama taip pat plačiai, taikant įprastinę metodiką. Nagrinėjamos krypties kontekste galima išskirti Vilniaus universiteto Kauno Humanitarinio fakulteto (VU KHF) mokslininkų įnašą. Pagrindinis tyrimo privalumas būtų tas, kad jo rezultatus būtų galima pritaikyti ir praktiškai, kuriant ekspertinę sistemą, kurią būtų galima naudoti tiek moksliniais tikslais, tiek ir praktikoje, realizuojant ją kaip konkretaus banko informacinės sistemos atskirą modulį.

Darbo **objektas** – Bajeso metodo taikymas kredito rizikos valdyme.

Darbo **tikslas** – pritaikyti Bajeso metodą kredito rizikos valdyme.

Siekiant iškelto tikslo, yra apibrėžiami tokie darbo **uždaviniai**:

1. Atlikus informacijos šaltinių tyrimą, apibrėžti kredito rizikos valdymo esminius aspektus ir problemines sritis bei aprašyti metodus šioms problemoms spręsti;
2. Pagal anksčiau atliktus tyrimus ir palyginamąsias analizes įvertinti Bajeso metodo tinkamumą kredito rizikos vertinimui;

3. Bajeso metodo panaudojimo kredito rizikos vertinime galimybių atvejų stebėjimas ir jų analizavimas;
4. Praktiškai patikrinti atlikto tyrimo kokybinį įvertinimą:
    - Sumodeliuoti ir suprojektuoti AI Bajeso metodą, naudoti kredito rizikos vertinimui;
    - Realizuoti sumodeliuotą sistemą;
    - Ištirti ir įvertinti gautuosius empirinio tyrimo rezultatus, palyginti su alternatyvių sistemų gautais rezultatais.
5. Patvirtinti arba paneigti ginamuosius teiginius (darbinę tyrimo hipotezę);
6. Pateikti pasiūlymus, susijusius su nagrinėjama tema.

Metodologiniai metodai naudoti darbe:

- Visuotinio pažinimo metodas (darbo tikslų nustatymas, uždavinių formulavimas, informacijos apie dirbtinio intelekto metodus kredito rizikos vertinimui rinkimas ir analizė; duomenų apibendrinimas; išvadų formulavimas).
- Bendrieji mokslinio tyrimo metodai:
    - indukcijos metodas (darbo išvadų formulavimas);
    - dedukcijos metodas (naudotinas nuo bendro sprendimo prieiti prie atskirų dalių);
    - palyginimo metodas (metodų charakteristikų, panašumų ir skirtumų palyginimas).
- Duomenų analizės metodas (esamos situacijos vertinimas).
- Modeliavimo metodas (programinės realizacijos projektavimas ir modeliavimas).
- Apibendrinimo metodas (naudojamas medžiagos grupavimui, apibendrinimui bei išdėstymui).
- Abstrakcijos metodas – juo remiantis, daromos kiekvienos darbo dalies ir galutinės viso darbo išvados.

Darbą sudaro trys pagrindinės dalys – analizės dalis, eksperimentinės sistemos tyrimo dalis bei eksperimentinė dalis. Kiekvienoje iš šių dalių pateikiama atitinkama informacija apie atliekamą tyrimą ir kuriamą informacinę sistemą:

- *Analizės dalyje* aprašyta nagrinėjama problema, apžvelgtos pagrindinės su ja susijusios sąvokos, klausimai bei ankstesni tyrimai, pasirinktas metodas, pagal kurį bus atliktas tyrimas.
- *Eksperimento dalyje* iškelti techniniai reikalavimai būsimai sistemai, apibrėžtos funkcijos, kurias ji turi atlikti, pateikiama jos koncepcinė schema ir aptariamos jos išplėtimo ateityje galimybės; aprašomas su sukurta informacine sistema atliktas tyrimas, pateikiami jo rezultatai bei pasiūlymai tolimesniam tyrimui.

Darbe naudota interneto medžiaga, moksliniai straipsniai, paskaitų konspektai apie kredito riziką, finansus bei intelektinius metodus. Darbą sudaro x puslapių, x lentelių, x iliustracijų bei x priedų.

VILNIAUS UNIVERSITETO KAUNO HUMANITARINIO FAKULTETO INFORMATIKOS KATEDRA

VERSLO INFORMATIKOS MAGISTRANTŪROS PROGRAMOS

MOKSLO TIRIAMOJO DARBO PLANAS

Magistrantas: Gediminas Būzius, tel.: 863428301; 868611753

Magistrantūros trukmė nuo 2008m. iki 2010m.

TEMA: Bajeso metodo taikymas kreditų rizikos valdyme

Vadovas: Gintautas Garšva, daktaras, docentas, Vilniaus universiteto Kauno humanitarinio fakulteto informatikos katedra, 8 37 750539

Darbo anotacija:

Tikslas: išnagrinėti dirbtinio intelekto taikymo kredito rizikos valdyme galimybes ir pritaikyti Bajeso metodą.

Uždaviniai: Atlikti informacijos šaltinių tyrimą; Rastų dirbtinio intelekto metodų, o svarbiausia Bajeso tinklų, panaudojimo kredito rizikos vertinime galimybių atvejų stebėjimas ir jų analizavimas; Pagal anksčiau atliktus tyrimus ir palyginamąsias analizes įvertinti geriausiai tinkantį(-čius) kredito rizikos vertinimui dirbtinio intelekto metodą, įskaitant Bajeso metodą; Praktiškai patikrinti atlikto tyrimo kokybinį įvertinimą; Patvirtinti arba paneigti ginamuosius teiginius (darbinę tyrimo hipotezę); Pateikti pasiūlymus, susijusius su nagrinėjama tema.

Metodai, kuriuos ketinama ištirti ir panaudoti darbe: darbe bus tiriamas Bajeso metodas, o taip pat kiti dirbtinio intelekto metodai, kurie jau buvo panaudoti kreditų rizikai valdyti ar panaudoti analogiškai klasifikacijai.

Laukiami rezultatai: magistrinis mokslinio tiriamojo darbo aprašas ir Bajeso metodo principu veikianti sistema kreditų rizikai vertinti.

Mokslo - tiriamojo darbo planas

Semestras	(data)	Užduotys
S1	2008 rudens	Preliminarios magistrinio darbo temos formulavimas. Literatūros šaltinių parinkimas ir esamos padėties apžvalga, įvertinant, kas jau yra ištirta, kas žinoma. Literatūros šaltinių analizė, grupavimas, pagrindinių idėjų apžvalga. Tikslus darbo problemos formulavimas. Darbo objekto, darbo tikslo ir uždavinių numatymas. Magistrinio darbo preliminaraus plano parengimas. Pirmojo darbo etapo ataskaitos paruošimas.
S2	2009 pavasario	Teorinės darbo problemos sprendimo medžiagos ruošimas. Išsami esančių metodų, algoritmų, sprendimų analizė. Jų lyginamoji analizė: privalumai, trūkumai ir kritika. Pasiūlyti naują programinį sprendimą (įgalinantį Bajeso metodą). Išsamus jų aprašymas, preliminarus siūlomo sprendimo skyriaus parengimas.
S3	2009 rudens	Eksperimentinės tyrimo metodikos ruošimas. Eksperimentinės aplinkos formavimas Bajeso metodui realizuoti. Duomenų eksperimentiniam tyrimui rinkimas, apdorojimas ir įvertinimas. Eksperimentinio tyrimo atlikimas. Preliminarus eksperimentinio tyrimo rezultatų įvertinimas. Preliminarių išvadų formulavimas.
S4	2010 pavasario	Teorinio skyriaus papildymas, remiantis atliktais eksperimentais ir naujausiais literatūros šaltiniais. Išvados apie gautų rezultatų praktinį pritaikymą. Apibendrinančių išvadų, pasiūlymų bei rekomendacijų rengimas. Magistrinio darbo įvado (galutinės redakcijos), santraukos (užsienio kalba), literatūros ir šaltinių sąrašo parengimas. Galutinis magistrinio darbo sutvarkymas

Magistrantas ..... Vadovas: .....
(parašas) (parašas)

Visi darbo uždaviniai bus atliekami iš eilės pateikta tvarka, arba, esant galimybei, lygiagrečiai vienas su kitu (pavyzdžiui, sistemos kūrimas bei metodikos analizė; sistemos kūrimas ir testavimas). Darbas turi būti atliktas per dvejus metus.

Pirmiausias ir svarbiausias uždavinys yra išnagrinėti turimą informaciją bei tą jos dalį, kuri iki šiol buvo praktiškai neprieinama. Tam bus atrenkami moksliniai straipsniai iš keleto minėtų duomenų bazių ar paieškos sistemų. Visa literatūra kruopščiai atrenkama, grupuojama ir sisteminama, jog tiktų kitiems tiriamojo darbo etapams.

Toliau, ištyrus iki šiol padarytus darbus, atrinkti labiausiai dėmesio vertus dirbtinio intelekto metodus, kurie būtų taikomi kreditų rizikos vertinime, o tarp jų ir Bajeso tinklus. Išsiaiškinti kaip sėkmingai naudojami dirbtinio intelekto metodai kreditų rizikos vertinimui, o tam didžiąja dalimi bus remiamasi iki tol atliktais tyrimais ir jų rezultatais iš nagrinėtų straipsnių. Taip pat bus atsižvelgta į galimybę tuos metodus ir jų algoritmus realizuoti. Metodų efektyvumas turėtų būti vertinamas pagal jų realizacijų eksperimentinius rezultatus, pavyzdžiui, atsižvelgiant į tai, kiek procentiškai kreditorių įvertino teisingai arba ne, taip pat, jei įmanoma palyginti metodų realizavimo sudėtingumą, jų darbo laiką vertinant riziką ir kitus aspektus.

Po to, remiantis turimomis išvadomis, turėtų būti kuriama konkreti programinė realizacija informacijos empiriniam patikrinimui atlikti (ji gali būti kuriama ir lygiagrečiai, jei tai bus įmanoma) arba esamų ne komercinių atvirojo kodo programinių paketų pritaikymas. Projektavimui ir modeliavimui bus panagrinėta jau sukurta nemokama programinė įranga, pvz., Weka aplinkoje, į kurią bus integruota jau turima duomenų bazė su finansiniais organizacijų duomenimis.

Duomenys saugomi MySQL duomenų bazėje, kuri preliminariai jau sudaryta VU KHF doktoranto P.Danėno, tik ji gali būti papildoma naujais duomenimis. Sukūrus sistemą, ją testuoti, o vėliau atlikti lyginamąją analizę su kitomis realizacijomis, pvz., kad ir su minėta P.Danėno sukurta realizacija JAVA kalboje, kuri remiasi SVM metodu kreditų rizikai vertinti pagal firmų finansinius duomenis. Jei pavyks, šios sistemos gali būti apjungtos, o taip pat sukurta realizacija gali būti palyginta su analogiškomis Bajeso tinklus įgalinančiomis laisvai prieinamomis atvirojo kodo realizacijomis, kurių sąrašas ir nuorodos į juos pateikiamos: www.kdnuggets.com/software/bayesian.html. Realizuota sistema, pateikiant jai įvedimo duomenis (finansinius organizacijos duomenis) turėtų tą organizaciją tik priskirti prie mokios, vidutinės rizikos arba nemokios.

Galutinis darbo žingsnis – atlikti ginamųjų teiginių patvirtinimą arba atmetimą, be to, pateikti konkrečius pasiūlymus, kokie tyrimo aspektai galėtų būti patvirtinti praktikoje.

Darbo atlikimo preliminari eiga pagal darbo uždavinius pateikiama ir 1 lentelėje.

1 lentelė

Darbo atlikimo preliminari eiga

Eil. nr.	**Veikla/ Laiko intervalai**	2008.11.01	2008.12.30	2009.04.30	2009.05.31	2009.09.30	2010.04.30	2010.05.31
1	Informacijos šaltinių tyrimas							
2	Metodų stebėjimas ir analizė							
3	Tarpusavio metodų palyginimas							
4	Ekspertinės sistemos modeliavimas ir projektavimas							
5	Sistemos tyrimas ir įvertinimas							
6	Ginamųjų teiginių patvirtinimas arba paneigimas							

1. ANALITINĖ DARBO DALIS

Šiame skyriuje pateikiama literatūros apžvalga ir analizė, aprašant kreditų rizikos vertinimo ir valdymo aspektus, išvardijant kreditų rizikos vertinimo matematinius ir dirbtinio intelekto metodus bei juos aprašant, o taip pat pateikiant kai kurių iš jų panaudojimo atvejai šiai problemai spręsti. Pabaigoje pateikiami įvairių autorių tyrimai metodų klasifikavimo galimybėms įvertinti ir pateikiamos šios dalies išvados.

1.1. Kreditų rizikos vertinimo pagrindinės sąvokos ir principai

Finansinės organizacijos, prieš teikdamos paskolą, dažniausiai stengiasi surinkti apie savo klientus visą įmanomą finansinę informaciją, kurią naudoja, norėdami įvertinti galimos rizikos lygį. Ši informacija paprastai apima įvairius finansinius duomenis, tokius, kaip ketvirčių ar metų galutiniai balansai, pelno ataskaitos ir kiti rodikliai, parodantys tuometinę bendrovės ar asmens finansinę būklę.

Vertinant bet kokią riziką, svarbu atsiminti, kad kiekvienu atveju ji apima du pagrindinius aspektus: nepatikimumą ir pavojų prarasti investuotas lėšas bei neapibrėžtumą. Kredito rizikos atveju, ją vertindama finansinė institucija turi apsvarstyti tris klausimus [1]:

1. įsipareigojimų nevykdymo tikimybė (angl. *probability default*) – kokia galimybė, kad kita šalis nevykdys savo įsipareigojimų jų gyvavimo metu arba tam tikru jų laikotarpiu, pavyzdžiui, metais. Vienu metų laikotarpiu šis rodiklis gali būti apibrėžiamas kaip *tikėtinas įsipareigojimų nevykdymo dažnis*;
2. kredito praradimas (angl. *credit exposure*) – kokio dydžio bus neapmokėta skola, kai atsiras įsipareigojimų nevykdymas iš skolininko pusės;
3. atgavimo rodiklis (angl. *recovery rate*) – kokia prarasto kredito dalis įsipareigojimų nevykdymo atveju gali būti susigrąžinta per bankroto procedūrą ar kitą atsiskaitymo būdą.

Teikiant kai kuriuos kreditus, reikia atlikti papildomus veiksmus [11]:

- kai paskolą teikia keli bankai kartu (sindikuota paskola), prieš sudarydamas sutartį su kitais bankais kiekvienas bankas turi įvertinti kredito riziką, numatyti paskolos teikimo sąlygas ir kt.;
- kredituodamas investicinius ir panašius projektus, bankas įvertina ne tik ekonominę projekto sėkmę, bet ir technines jo įgyvendinimo galimybes, taip pat teisinius, aplinkosaugos ir kitus veiksnius, projekto plano laikymąsi. Jei banko darbuotojams nepakanka žinių, kad jie galėtų visapusiškai įvertinti

projektą, bankas turėtų samdyti reikiamą kvalifikaciją turinčius nepriklausomus ekspertus;

- paskolų su banku susijusiems asmenims teikimo sąlygas ir tvarką patvirtina banko stebėtojų taryba. Tokios paskolos neturėtų viršyti banko stebėtojų tarybos nustatytų dydžių. Paskolų teikimo sąlygos šiems asmenims negali būti palankesnės nei kitiems banko skolininkams;
- banke turėtų būti numatytos procedūros susijusiems skolininkams nustatyti, o paskolos jiems turėtų būti teikiamos pagal iš anksto nustatytą tvarką (Bazelio bankų priežiūros komiteto parengto dokumento (BCBS 2000c) 7 principas).

Kredito rizikoje taip pat naudojama ir sąvoka „kredito kokybė“, apibūdinanti skolininko galimybę atsiskaityti už įsipareigojimą. Ji apima tiek skolininko įsipareigojimų nevykdymo tikimybę, tiek ir atgavimo rodiklį [1].

Algoritminių metodų naudojimą kredito rizikos vertinimui galima apibrėžti sąvoka „kreditų rizikos modeliavimas“, kuri apima tokias sritis, kaip tradicinė kreditų analizė, metodai ir modeliai, naudojami įvertinti derivatyvas, bei portfelių kredito rizikos matavimus, naudojamus viso obligacijų portfelio analizei [1]. Svarbu pabrėžti ir tai, kad kai kuriais atvejais kredito rizikos vertinimas yra komplikuotas, unikalus, taip pat gali būti ir tokių atvejų, kai riziką geriau vertinti mažiau formaliu metodu. Pats kreditų kokybės vertinimo procesas apibrėžiamas kaip *kredito analizė* ir apima kredito rizikos vertinimą tiek algoritmiškai, tiek ir mažesnio formalumo metodais, o asmenys, atliekantys šį procesą, įvardinami kaip *kredito analitikai*. Šie asmenys pagal gautą informaciją apie būsimą skolininką įvertina kredito išdavimo galimybę, remdamiesi metiniais ir ketvirčių balansais, pajamų ataskaitomis, verslo šakos, kurioje veikia tiriamas subjektas, galimybėmis, ekonomine šalies situacija, įvairiais rodikliais ir kitais kriterijais, ir tuo remdamiesi sudaro *kredito reitingą,* kuriuo skolininką priskiria vienai ar kitai grupei pagal jo galimybes grąžinti kreditą. Šiuo reitingu remiantis, kreditorius gali atlikti kredito išdavimo sprendimus. Tokius reitingus kitoms įmonėms kuria ir pasaulyje žinomos bendrovės *Fitch ratings, Moody‘s, Dunn & Bradstreet, Standard & Poor‘s* ir kt., kurių reitingai pasaulyje yra vertinami ir pripažįstami įvairių finansinių institucijų, teikiančių kreditavimo paslaugas.

Įprastai bankai labiausiai koncentruojasi į pavienių kreditų rizikos vertinimą, t.y., prieš suteikiant kreditą įvertinamos skolininko galimybės jį grąžinti, suteikus kreditą stebima, kaip skolininkas vykdo įsipareigojimus bankui ir kokia yra rizika bei galimas nuostolis bankui, jei įsipareigojimai nebus tinkamai vykdomi. Bet egzistuoja ir viso kreditų portfelio rizika, kurią taip pat privalu vertinti ir valdyti – tai kreditų koncentracijos rizika, kuri susidaro būtent dėl pavienių kreditų koreliacijos. Kreditų koncentracijos rizika bankuose atsiranda dėl įvairių priežasčių: kai

kurie bankai specializuojasi tam tikrose veiklos srityse (kredituodami įmones, užsiimančias konkrečia ekonomine veikla), taip siekdami tapti rinkos lyderiais. Bankui pasirinkus tokią veiklos strategiją, kreditų koncentracijos rizika neišvengiama. Dėl didelės konkurencijos, siekdami neprarasti turimos rinkos dalies ar didesnio turto augimo, bankai gali būti priversti prisiimti didesnę koncentracijos riziką. Kartais kreditų koncentracijos rizika susidaro dėl to, kad, atsiradus kokiai nors sparčiai augančiai ekonominei veiklai, bankai labai teigiamai vertina jos perspektyvas ir tikisi uždirbti didesnį nei vidutinis pelnas (pavyzdžiui, iš turto kainų kilimo, didesnės kreditų maržos ar mokesčių). Kredito koncentracijos rizika gali atsirasti ir nuo dėl banko nepriklausančių priežasčių, pavyzdžiui, rinkos, kurioje veikia bankas, mažumo. Šios rizikos egzistavimas – tai galimybė bankui patirti santykinai didelį (palyginti su banko kapitalu, turtu, pajamomis ar, jei tai galima išmatuoti, bendra banko prisiimta arba jam priimtina rizika) nuostolį iš kreditų, jog sutriktų normali banko veikla.

Valdydamas kreditų portfelio riziką bankas siekia didžiausio pelno esant tam tikrai rizikai arba mažiausios rizikos esant konkrečiam pelnui. Tai darydamas bankas gali būti pasyvus arba aktyvus, t. y. pasirinkti vieną arba kitą strategiją [11]:

1. Pasyvi. Tai tradicinis požiūris į paskolų portfelio rizikos valdymą: suteiktos paskolos laikomos iki jų grąžinimo termino pabaigos, bankas nekeičia jau suteiktų paskolų portfelio sudėties. Laikantis pasyvios strategijos, paskolų portfelio rizika iš esmės valdoma tik taikant limitus, t. y. nesuteikiant per daug paskolų vienam skolininkui, susijusiems ar ta pačia ekonomine veikla užsiimantiems skolininkams. Pasyvios strategijos laikosi dauguma bankų.
2. Aktyvi. Bankas, išlaikydamas dalį paskolų iki jų grąžinimo termino pabaigos, aktyviai keičia paskolų portfelio sudėtį, kad gautų didžiausią pelną, esant priimtinai rizikai. Taikomos tokios priemonės, kaip paskolų pardavimas, paskolų keitimas vertybiniais popieriais (angl., *securitisation*), paskolų draudimas, taip pat išvestinės finansinės priemonės. Banke, kuris pasirenka aktyvią strategiją, paprastai įsteigiamas struktūrinis padalinys, atsakingas už paskolų portfelio rizikos valdymą.

Laikantis pasyvios strategijos, kreditų portfelio rizika daugiausia valdoma taikant limitus. Kredito rizikos limitai gali būti klasifikuojami įvairiais būdais [12]:

- *riziką ribojantys ir informuojantys limitai.* Nustatant kredito riziką ribojančius limitus siekiama riboti prisiimamą riziką, tačiau tuo pat metu daroma įtaka veiklos rezultatams, tiksliau jos apimčiai ir pelningumui.
- *apimties ir rizikos įvertinimu pagrįsti limitai.* Toks limitų skirstymas pagrįstas dvejopais kreditų koncentracijos rizikos mato vienetais. Apimties limitai nustatomi

atsižvelgiant į kreditų dydį, pavyzdžiui, limitas didžiausiam kreditui vienam skolininkui nustatomas kaip tam tikras banko kapitalo procentas. Apimties limitai pagrįsti prielaida, kad rizika tiesiogiai proporcinga kredito dydžiui, tačiau dėl kreditų rizikos koreliacijos taip nėra. Nustatant rizikos įvertinimu pagrįstus limitus atsižvelgiama į kredito riziką lemiančius veiksnius: kredito dydį, įsipareigojimų neįvykdymo tikimybę, nuostolį neįvykdžius įsipareigojimų, kredito trukmę.

- *absoliutūs ir santykiniai limitai.* Dauguma limitų yra santykiniai, paprastai skaičiuojami kaip tam tikras banko kapitalo procentas. Absoliutaus limito pavyzdys yra ilgiausia kredito trukmė, didžiausia tam tikro kredito suma;
- *išankstiniai ir paskesni limitai.* Limitai gali skirtis pagal tai, kada į juos atsižvelgiama, t. y. prieš suteikiant kreditą ar jį suteikus. Tačiau į daugumą, jei ne į visus kreditui taikytinus limitus, turėtų būti atsižvelgiama prieš suteikiant kreditą, t. y. įvertinama, kaip pasikeis banko kreditų portfelis ir kreditų koncentracijos rizika, jei bus suteiktas naujas kreditas.

Plačiau apie kreditų koncentracijos rizikos svarbą, vertinimą ir valdymą rašoma [12] šaltinyje, o vystymosi raida ir populiariausi darbai iki 1998 metų paminėti [13]. Nors kreditoriams svarbu vertinti ir valdyti abu minėtus rizikos aspektus (pavienius ir koncentracijos), šio tiriamojo darbo objektas yra pavienių kreditų rizikos vertinimas ir valdymas. Apie paskolų portfelio riziką daugiau aprašo V.Valvonis straipsnyje „Šiuolaikinis kredito rizikos vertinimas banke: paskolų portfelio rizika ir ekonominio kapitalo paskirstymas“, Pinigų studijos 2006/2, apžvalginiai straipsniai.

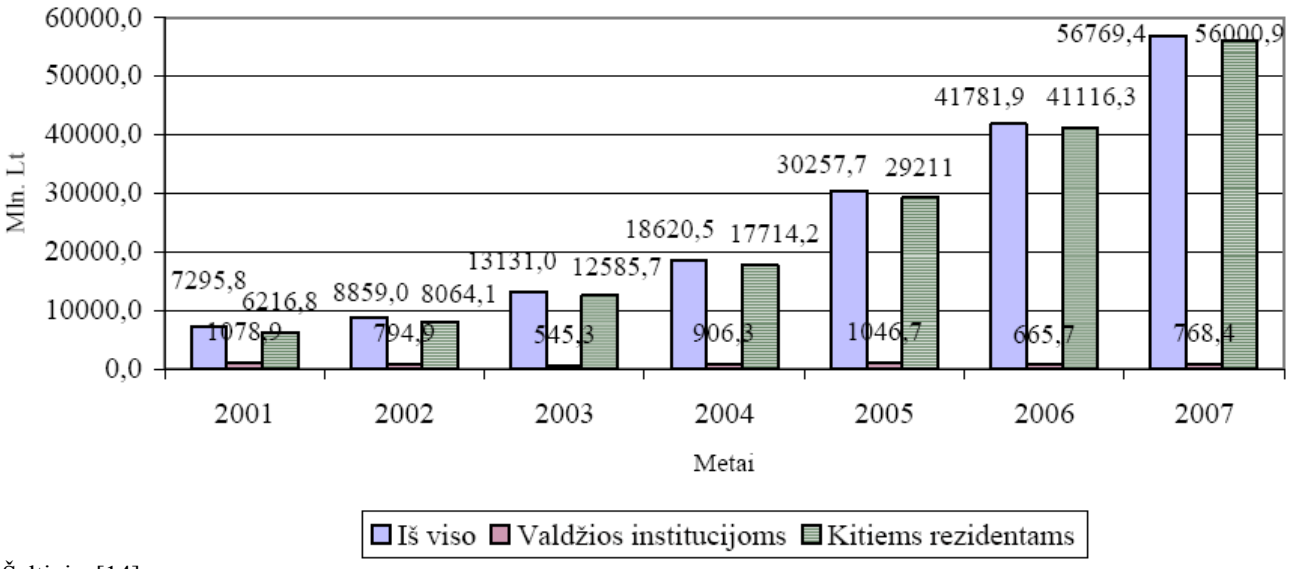
Vien Lietuvoje, kur pagal Statistikos departamento prie Lietuvos respublikos vyriausybės pateiktus 2008 metų duomenis gyvena 3366400 žmonių, kurių finansinius reikalus vienaip ar kitaip sieja 11 bankų:

- AB bankas „Snoras“
- Danske Bank A/S Lietuvos filialas
- AB DnB NORD bankas
- UAB Medicinos bankas
- Nordea Bank Finland Plc Lietuvos skyrius
- AB Parex bankas
- AB SEB bankas
- AB „Swedbank“
- AB Šiaulių bankas

- AS UniCredit Bank Lietuvos skyrius
- AB Ūkio bankas

Šių bankų įvairios paskirties paskolų likučiai nominalia verte (pagal Bankų asociacijos 2009-06-09 pateiktus duomenis) fiziniams asmenims yra 28,03 milijardai litų, apytiksliai 8327 litų vienam žmogui. Juridiniams asmenims bankai yra išdavę maždaug 40,25 milijardus litų.

Šaltinyje [14] pateiktos paskolų išdavimo tendencijos 2001-2007 metams (1 paveikslas). Kreditų išdavimo tendencijos kiekvienais metais vis labiau augo iki pastarosios ekonominės recesijos visame pasaulyje. Iš aukščiau pateiktų duomenų matoma, kad ir 2008 metais paskolų dydis išaugo.



Šaltinis: [14]

1 pav. Išduotų paskolų Lietuvoje augimo tendencijos

Paskolų augimas paaiškinamas per bankinių struktūrų ir ekonomikos plėtojimąsi, kuris turi taip vadinamą finansinio akseleratoriaus efektą, kuris cituojant [14]: „<...> sustiprina verslo ciklo svyravimus. Pirmajame etape dėl bankų kredito didėjimo ekonomikos augimas būna ženkliai spartesnis. Antrajame etape, kai verslo ciklas pereina į lėtėjimo fazę, dažnai kreditą pradedama normuoti, kyla palūkanų normos, stabilizuojasi arba ima mažėti turto kainos, tampa akivaizdi bankų prisiimta kredito rizika (Ramanauskas, 2005). Tai neigiamai veikia ūkio plėtrą, gali sukelti finansinę krizę.“, kas pasitvirtino vėliau 2008 metų pabaigoje. Žlugo nemažai bankų visame pasaulyje, kaip pavyzdžiui, 2009 metų gegužę didžiausias bankrutavęs Amerikos bankas BankUnited FSB, kuris jau buvo 34-as bankrutavęs Amerikos bankas (informacija pateikta Amerikos bankų priežiūros tarnybos), o daugeliui didžiųjų Jungtinių Amerikos ir kitų šalių bankų reikėjo valstybės paramos, t.y., finansinių injekcijų be kurių bankai būtų kapituliavę. Šią finansinę krizę kai kas vadina kreditų krize, kurią sukėlė didelės išduotų kreditų sumos su netinkamai

įvertinta rizika. Pastarieji faktai tik dar labiau pabrėžia šios tiriamojo darbo ir kreditų rizikos vertinimo srities aktualumą.

Apie dalies (šešių konkrečiai neįvardintų) Lietuvos bankų kreditų ir kreditų portfelių rizikos valdymo principus, naudojamus metodus ir politiką galima rasti [11] šaltinyje.

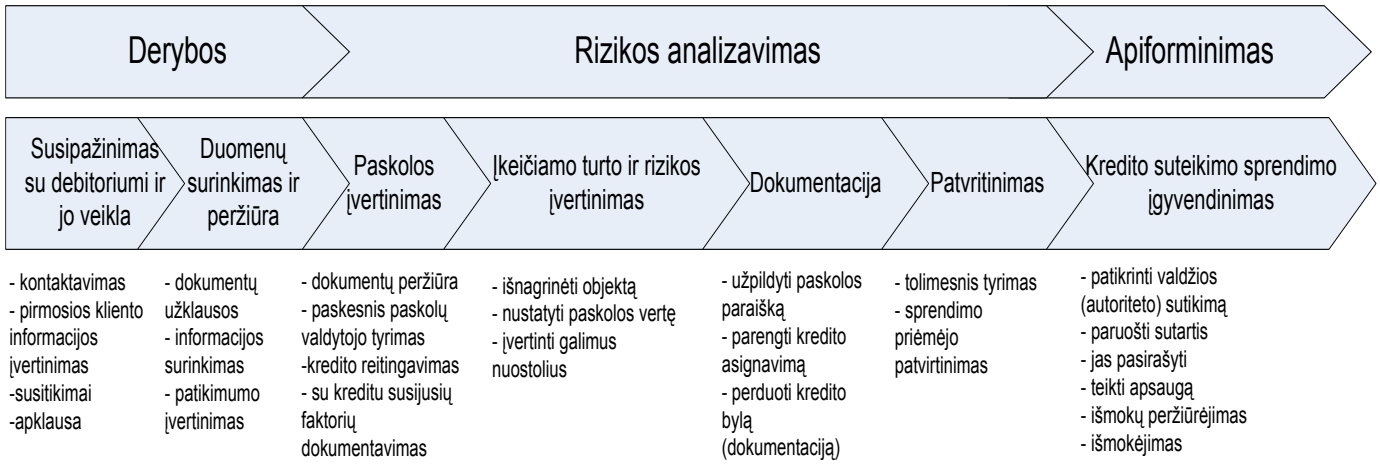
1.2. Pavienių kreditų rizikos vertinimas

Kreditų (paskolų) teikimas – tai pagrindinė įprasto banko veikla. Jog paskolos būtų sėkmingai teikiamos, bankui neužtenka vien tik turėti finansinių išteklių, bet reikia ir informacijos apie skolininkus, jų kredito riziką, t.y., finansiniai ištekliai yra būtina, bet ne vienintelė sąlyga, norint gauti pelno iš paskolų teikimo. Sėkmę teikiant paskolas gali lemti įvairi informacija: bendrieji duomenys apie skolininką, duomenys apie skolininko ankstesnių įsipareigojimų vykdymą, skolininko finansiniai duomenys, skolininko veiklos efektyvumo, jo patikimumo, reputacijos įvertinimas, rinkos ir šalies, kurioje veikia skolininkas, rizikos įvertinimas, informacija apie paskolos grąžinimo užtikrinimo priemones ir kita [15].

Finansinę riziką būtų galima skirstyti į tris grupes, t.y., riziką investuojant į valiutų biržas, riziką investuojat į vertybinius popierius (akcijas, obligacijas ir pan.) ir kreditų rizika (įvairių paskolų suteikimas), kuri šiuo atveju yra pagrindinis šio darbo objektas, o koncentruojamasi į kreditų suteikimą įvairioms kompanijoms (įvairiems juridiniams asmenims), kredito teikėjas – bankas, tačiau tai aktualu ir kitoms kreditus teikiančioms unijoms.

1.3. Kredito suteikimo procesas

Pagal Austrijos banką Oesterreichische Nationalbank (OeNB) ir Austrijos ministeriją [16] įvairioms bendrovėms (įmonėms, organizacijoms) paskolos teikiamos daugiau ar mažiau panašiu formaliu procesu, kuris suskaidytas ir pateiktas 2 paveiksle.



Šaltinis: sukurta autoriaus, remiantis [16]

2 pav. Kredito patvirtinimo proceso eiga

Informacija apie debitorių gaunama įvairiais būdais ir šaltiniais, nes reikia įvertinti ne tik ekonominę bendrovės būseną, bet ir kokybinę, o tai gali būti bendrovės apskaitos tikrinimas, finansinių dokumentų ir informacijos rinkimas, mokesčių inspekcijos duomenys, jei bendrovė prekiauja akcijomis, tai ir akciniai duomenys, kuriuos audituoja tam tikros organizacijos, galimas ir bendrovės klientų bei partnerių apklausos, pokalbiai su eiliniais bendrovės darbuotojais, empirinis paslaugų kokybės vertinimas (pvz., slaptieji pirkimo agentai), jei kreditas reikalingas kokio nors projekto įgyvendinimui, tai ir to projekto dokumentacija, turimos įrangos ir technikos vertinimas ir t.t.

Remiantis pastaruoju šaltiniu, kredito suteikimo proceso kokybė rizikos atžvilgiu yra nustatoma svarbiausiais įmanomais identifikavimo ir įvertinimo būdais. Kredito rizika įvertinama pagal *Basel* komiteto (Bazelio bankų priežiūros komitetas) rekomendacijas, kurias didžiąja dalimi adaptuoja Europos komisija ir yra įtraukta į Europos sąjungos direktyvas, todėl galioja visoms Europos sąjungos narėms. Išskiriamos keturios rizikos komponentės:

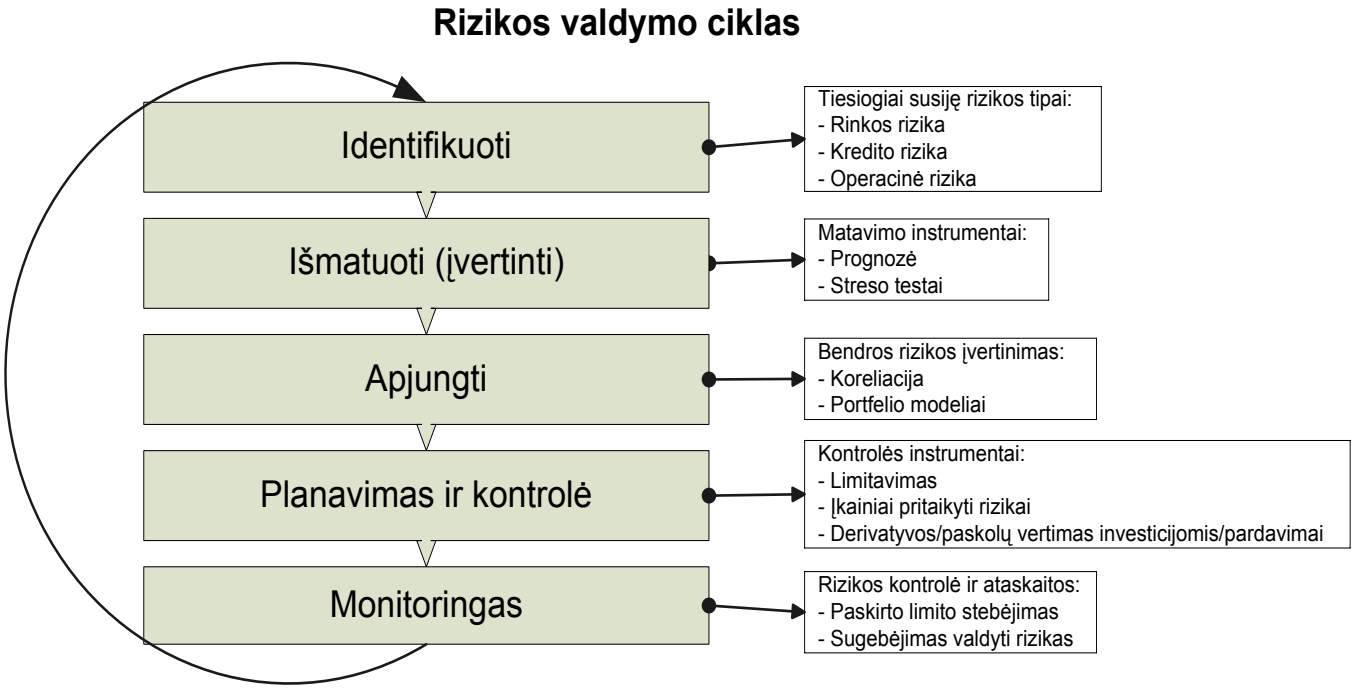
1. Tikimybė, jog kreditorius neįvykdys įsipareigojimų (angl., Probability of default, PD);
2. Dėl skolininko įsipareigojimų neįvykdymo atsiradusio nuostolio ir pozicijos vertės santykis, išreikštas procentais (angl., Loss given default, LGD);
3. Turto praradimo dydis įsipareigojimų neįvykdymo metu (angl., Exposure at default, EAD), kuris pagal Bazelio rekomendacijas dažniausiai skaičiuojamas vieneriems metams: iš nesumokėtos skolos dydžio (įsipareigojimo nevykdymo metu) atimant galimas pajamas iš įkeisto (užstatyto) turo;
4. Kredito suteikimo terminas (angl., Maturity, M);

Pastarasis komponentas yra mažiausiai svarbus iš visų išvardytų ir labiausiai prisideda prie reikalingo kapitalo skaičiavimo. Tuo tarpu pirmasis yra vertinamas pagal dabartinę ir ateities galimybes padengti kreditą ir palūkanas už jį bei kitus įsipareigojimus. Ypač svarbu įvertinti grynųjų pinigų srautus ir pajamas, kuriais ir turėtų būti padengti įsipareigojimai, o ne, tarkim, kitų kreditų pagalba. Šiuo atveju turi būti peržiūrėta ne tik kredituojama bendrovė, bet ir visas verslo modelis. Antrasis komponentas LGD vertinamas įkeisto turto pavidalu (kilnojamasis ir nekilnojamasis turtas), kur reikia atsižvelgti ne tik į įkeisto turto dabartinę vertę, bet ir vertę už kurią jis būtų vėliau realizuotas, parduotas, taigi atsižvelgiama į įkeisto turto tipą ir jo vertę. EAD atveju reikia atsižvelgti ne tik į skolininko tipą, grynųjų pinigų srautų šaltinį, įkeičiamo turto vertę ir tipą bet ir į pretenduojamo įkeisto turto dydį bendrovės atžvilgiu (ar įkeičiama visa bendrovė, ar tik dalis jos įrenginių, technikos, žemių, pastatų ir t.t.).

„Basel II“ skolintojo kapitalo įvertinimui siūlo du būdus, t.y., standartinį ir vidinio reitingavimo įvertinimą (angl., internal ratings-based approach - IRB approach), kuris remiasi 7-iomis turto klasėmis: priklausančio vien tik bendrovei, įkeisto bankui arba kitiems juridiniams ar fiziniams asmenims, bendras/kolektyvinis turtas, parduodamas turtas, akcininkų turtas, vertybinių popierių turtas, nekintantis turtas.

1.4. Kredito rizikos valdymo ciklas

Rizikos valdymas apima identifikavimą, įvertinimą, apjungimą, planavimą ir valdymą bei monitoringą (3 iliustracija). Visų pirma rizikos yra identifikuojamos, numatomi įvairūs scenarijai. Dažniausiai bankai išskiria į pačio kreditoriaus veiksmų riziką (netinkamų technologijų rizika, naujų produktų ar paslaugų rizika), rinkos riziką (akcijų rinkų, valiutų, palūkanų normų, kapitalo, likvidumo rizikos) ir operacinę riziką, o taip pat išskiria strategines bei reputacines rizikas, tačiau jas sunkiau išmatuoti kiekybiškai. Šiuo atveju lengviausiai vertinama rinkos rizika, kuri turi pakankamai istorinių duomenų. Kredito rizika skirstoma į numatytus nuostolius, kuriuos dažniausiai jau nuo pat pradžių turi numatyti pats skolintojas, ir netikėtus nuostolius, kurie dažniausiai būna didesni nukrypimai nuo numatytų nuostolių. Vėliau apjungiamos visos įmanomos rizikos, atsižvelgiama į jų tarpusavio koreliaciją ir įvertinama visa rizika kreditoriui. Kreditorius planuoja ir valdo galimas rizikas, valdymas galimas kaip limitų nustatymas sudarant individualius sandorius ar portfelius, garantijų, kredito draudimo naudojimas, kredito suteiktų lėšų vertimas investicijomis, turto pirkimas ir pardavimas ir pan.



Šaltinis: sukurta autoriaus, remiantis [16]

3 pav. Rizikos valdymo ciklas

Kreditų monitoringas (1 lentelė) apima įvairių kriterijų sekimą (pavyzdžiui, automatizuotai, naudojant IT priemones) tam tikrais laiko periodais, pvz., banko suteikto kredito viršijimas, nepakankamos kredito transakcijos sekamos kasdien, nauji įsipareigojimai, įmokų vėlavimas, finansinės ataskaitos – kas mėnesį, o pramonės šakos informacija – kas metus, šie visi faktoriai yra vieni iš pirmųjų skolininko nemokumo (rizikos) atsiradimo požymių. Monitoringo esmė yra pastebėti galimus debitoriaus nebesugebėjimus valdyti finansinės rizikos ir nebevykdyti įsipareigojimų. E. Bernhardsen apie monitoringą teigia, kad svarbiausia pateikti reikiamą informaciją apie kreditų sistemos funkcionavimą bei sekti, kaip laikui bėgant kinta kredituoto asmens galimybės grąžinti kreditą.

2 lentelė

Teiginiai apie kreditų monitoringą

Autorius	Metai	Apibrėžimas
Ari Hyytinen, Otto Toivanen	2004	*Kreditų monitoringas* – tai efektyvi priemonė kreditavimo proceso valdymo kokybei gerinti.
D.Jatkūnaitė, N.Stončiuvienė, N.Žaltauskienė	2000	*Kreditų monitoringas* prasideda iš karto po to, kai gaunama paraiška kreditui gauti, užtikrinanti savalaikį kredito grąžinimą bei palūkanų mokėjimą.
Maluis R., Salaski N.	1995	*Kreditų monitoringas* – tai nuolatinis kredituoto asmens stebėjimas, informacijos apie jo ekonominę būseną rinkimas, ateities tendencijų prognozė.
V. Valvonis	2004	*Banko kreditų monitoringas* – tai priemonė, padedanti įvertinti kreditų portfelio riziką.
R. Rukuižienė	1997	*Kreditų monitoringas* – tai kredito naudojimo sąlygų kontrolė, atsižvelgiant į faktinius kredito gavėjo ūkinės veiklos rodiklius, ir probleminių kreditų pasireiškimo aplinkybių ir veiksmų išaiškinimas siekiant išvengti kreditinės

		rizikos tolesnių padarinių.
S. Taraila	2001	*Kreditų monitoringas* - kredito gavėjo veiklos kontrolės ir probleminių kreditų, kuriems gresia nesavalaikis grąžinimas, ankstyvoje stadijoje išaiškinimo procesas.
D.Lauskienė,V.Snieška	2003	*Bankų kredito monitoringas* - tai krediuojamų asmenų vertinimas prieš išduodant kreditą, bei nuolatinė jų kontrolė iki kreditas grąžinamas.

Šaltinis: [15]

Remiantis [15] šaltiniu, bankų kredito monitoringas prasideda iš karto po to, kai gaunama paraiška kreditui gauti, tad kreditų monitoringo procese galima išskirti du etapus – kreditų monitoringas iki kredito išdavimo, ir kreditų monitoringas po kredito išdavimo, todėl banko kreditų monitoringo procesą iki ir po kredito išdavimo siūlo skirstyti į šiuos etapus:

1) monitoringo tikslų ir uždavinių nustatymas;

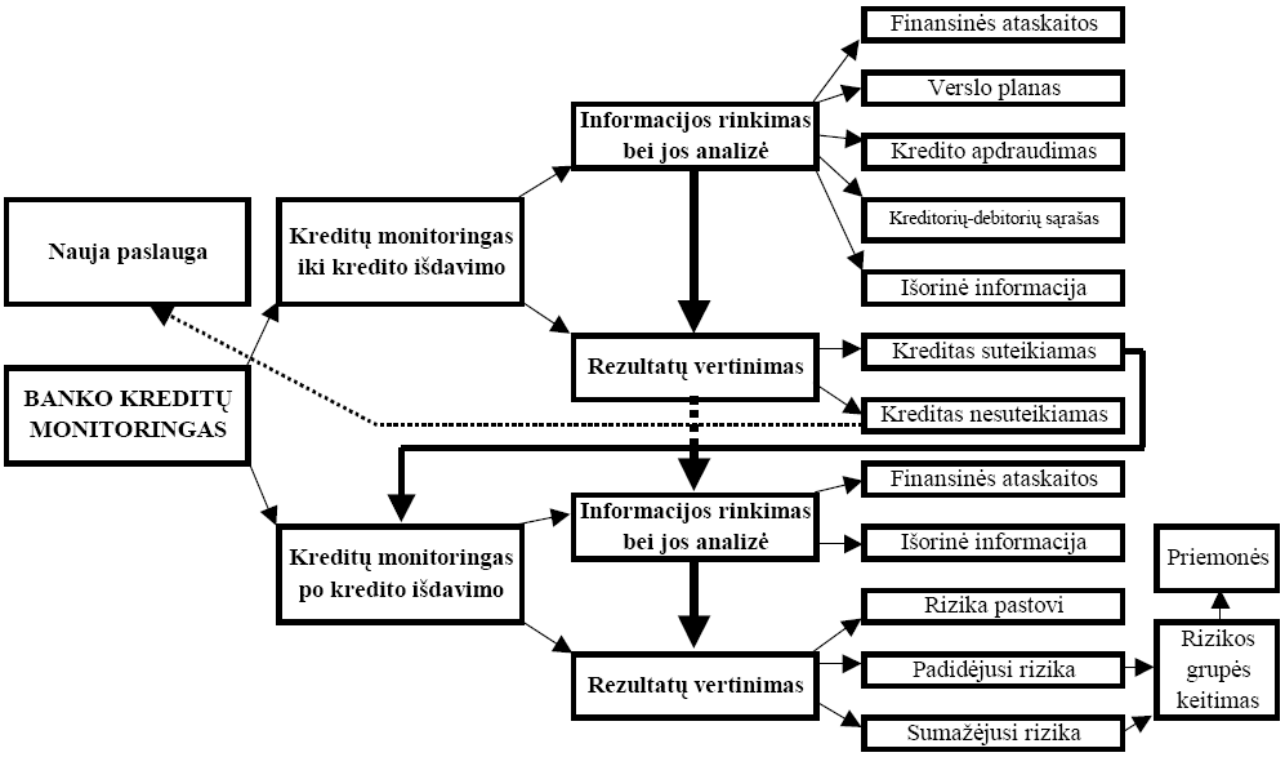
2) indikatorių sistemos kūrimas;

3) monitoringo plano sudarymas;

4) duomenų, informacijos apie kredito gavėjo ūkinę finansinę būklę, rinkimas ir analizė;

5) rezultatų įvertinimas.

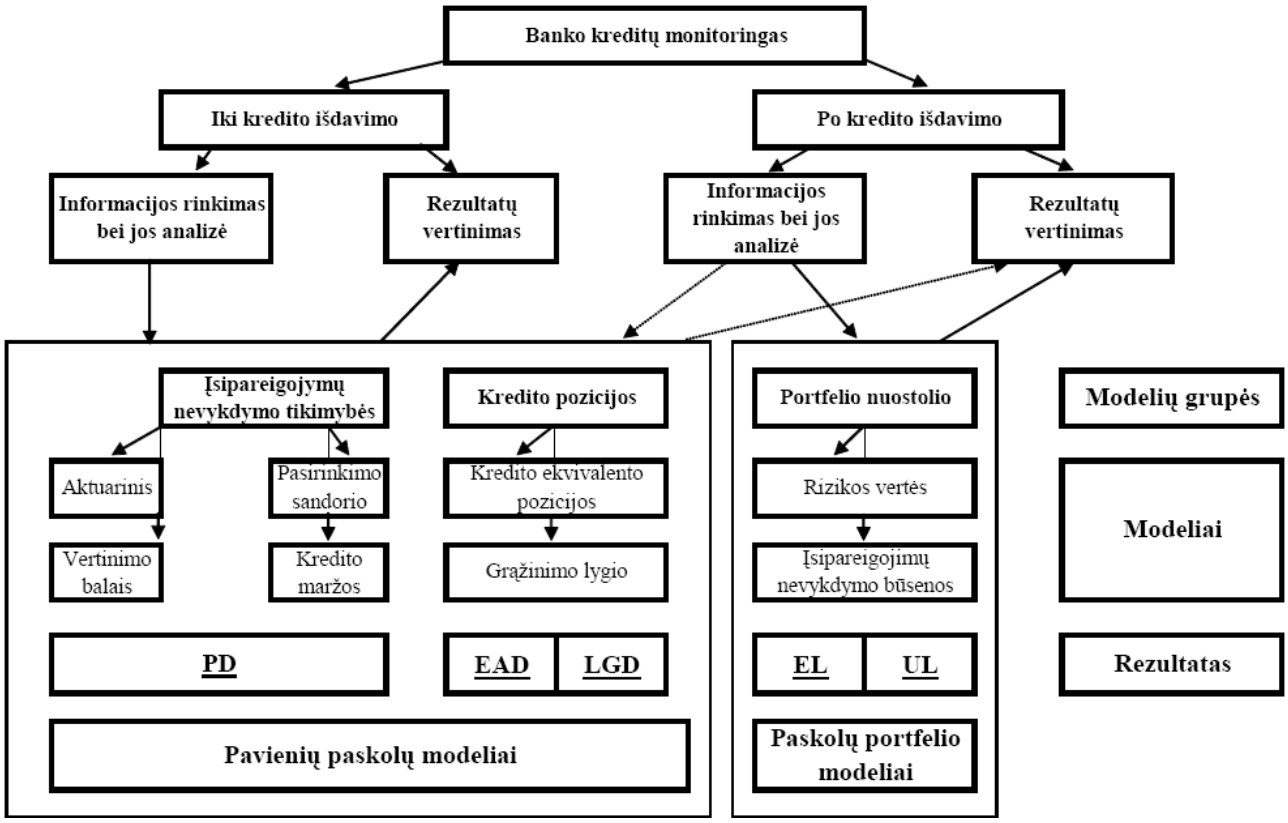
Nurodytas pirmasis etapas yra svarbus, nes klaidos šiame etape gali padaryti nerezultatyvų visą monitoringo procesą. Veiksmų vykdymas ir rezultatai stebimi ir vertinami tam tikrais kiekybiniais rodikliais, todėl svarbus antrasis banko kreditų monitoringo organizavimo proceso etapas, o siekiant maksimalaus efekto, kredito monitoringo pasiekimus apibūdinantys rodikliai tūrėtų būti pasirenkami iš anksto, jų pasirinkimą lemia informacijos prieinamumas, t.y. turi būti atsižvelgta, ar stebėjimo metu galima gauti vienokius ar kitokius duomenis. Šių teiginių grafinė interpretacija pateikiama 4 paveiksle.



Šaltinis: [15]

4 pav. Kreditų monitoringo procesas

Pastarasis autorius bando apibrėžti kredito rizikos vertinimo bei monitoringo sąsają, o tai siūlo padaryti per banko kreditų monitoringo tikslus ir uždavinius, indikatorių sistemą, renkamą ir apdorojamą informaciją, nukreiptą abiem atvejais į kredito rizikos mažinimą. Pagal anksčiau išreikštus teiginius banko kreditų monitoringo proceso metu surinkta informacija yra panaudojama kredito rizikos nustatymui.



Šaltinis: [15]

5 pav. Kredito rizikos vertinimo bei monitoringo sąsaja

Kreditų monitoringo proceso metu iki kredito išdavimo nustatomas tikėtinas ir netikėtinas nuostolis (atitinkamai angl., expected ir unexcpected loss – EL, UL), o tada priimamas sprendimas – suteikti kreditą ar ne. Priklausomai nuo to, kokio dydžio nustatytas tikėtinas nuostolis, parenkama atitinkama palūkanų norma. Kaip jau buvo minėta, vertinant kredito riziką, jo sudedamosios dalys: įsipareigojimų nevykdymo tikimybė (PD), nuostolis įsipareigojimų nevykdymo atveju (LGD), kredito pozicija įsipareigojimų nevykdymo atveju (EAD), pagal kuriuos galima apskaičiuoti EL ir UL (5 pav.).

1.5. Rizikos padengimas

Rizikos padengimas (angl., *Risk-bearing Capacity*) šiuo atveju reiškia kreditoriaus galimybes padengti rizikas iš savo turimų finansinių resursų (iš akcijų, valiutų kurso padidinimo, rezervų, pelno). Visi resursai, kuriais gali kreditorius manipuliuoti rizikų padengimui laikomi padengimo kapitalu (fondu, turtu). Šio kapitalo dydis apriboja skaičių derybų, kurios yra nesaugios rizikos atžvilgiu ir į kurias turėtų leistis kreditorius. Šio kapitalo dydis svariai prisideda prie banko verslo politikos ir rizikos valdymo politikos. Žinoma, jei bankas turi didesnius piniginius resursus, jis gali labiau plėtotis, skolinti rizikingesniems klientams, priimti didesnius nuostolius, jei rizika visgi išsipildytų [16].

1.6. Rizikos išmatavimas

Yra du pagrindiniai metodai, kuriais remiantis galima išmatuoti netikėtus nuostolius (UL):

- Vertės pokyčio rizika (Value-at-Risk, VaR), kuris yra portfelio galimų nuostolių dėl rinkos kainos kitimo kiekybinis įvertinimas tam tikru laikotarpiu su tam tikra tikimybe.
- Scenarijaus technikos (angl., scenario techniques).

Pirmasis iš jų yra tikslesnis, tačiau jį sudėtingiau apskaičiuoti, o antrasis dažniausiai naudojamas tada, kai nėra galimybės apskaičiuoti VaR. Abu metodai aprašyti žemiau.

1.7. VaR

VaR nurodo galimų nuostolių kiekybinį įvertinimą tam tikram laikotarpiui su tam tikra neviršytina tikimybe. Tam nustatoma patikimumo riba, kuri dažniausiai yra tarp 95% ir 99,95%, o tai reiškia, kad didesni nuostoliai yra įmanomi, tačiau tik su tikimybe tarp 5% ir 0,05%. VaR apskaičiavimui reikia nustatyti potencialų galimų nuostolių pasiskirstymą, o tam yra daromos prielaidos apie galimą (ateities) įsipareigojimų nevykdymą ir EAD (Exposure at default).

Šis rodiklis, deja, neturi jokio nuostolių dydžio nustatymo likusiems 5%, o be to neatsižvelgia į jokius staigius rinkos ar ekonomikos pokyčius (pvz., krizę), kurios metu staigiai išauga įsipareigojimų nevykdymo tikimybė. Skirtingoms rizikoms vertinti skiriasi ir apskaičiavimo sudėtingumas, kaip, tarkim, minėtoms rinkoms, operacinei ir kredito rizikos tipams. Dar vienas neigiamas aspektas gali būti nepakankamas istorinių duomenų apie skolininką kiekis.

Smulkiau apie vertės pokyčio riziką, jos atsiradimą, pritaikymą ir kitus aspektus galima rasti šaltinyje [17], o taip pat taikymą valiutų rinkoje, įskaitant ir CVaR (conditional VaR) šaltinyje [18].

1.8. Scenarijaus technikos

Šiai rizikos įvertinimo analizei taip pat reikia istorinių rinkos arba vidinių kreditoriaus duomenų, taip kuriant tam tikrus galimus scenarijus apie galimus įsipareigojimų nevykdymus.

Kaip ir VaR atveju bandoma nuspėti „normalius atvejus“, kur tariama, kad analogiškų nuostolių jau yra buvę tam tikruose istoriniuose duomenyse, ir blogiausius galimus atvejus, kur bandoma nuspėti didžiausius įmanomus nuostolius. Šie du scenarijai naudojami nustatyti svyravimus (nepastovumą). Tokie pasikeitimai gali atsirasti dėl eilės priežasčių, pvz., pinigų vertės kaita, įkeisto turto vertės keitimasis ir t.t. Dėl savo mažesnio parametrų, naudojamų rizikai nustatyti, kiekio teikia prastesnius rezultatus nei VaR.

Taigi, aptarus įvairius su kreditų rizikos vertinimu ir valdymu susijusius aspektus, pateikiami konkretūs vertinimo metodai, kurie skiriami į dvi dalis – matematinius statistinius ir dirbtinio intelekto metodus, pastarieji dėl savo objektiškumo, lyginant su matematiniais metodais,

yra šio tiriamojo darbo objektas, kai tuo tarpu matematiniams statistiniams metodams reikia taip vadinamų ekspertinių žinių, kurie, kaip jau buvo aprašyta aukščiau skyriuose, yra pagrindiniai, kurie naudojami realiuose bankuose ar kitose kreditais užsiimančiose kompanijose bei yra labai plačiai ištirti.

1.9. Kredito rizikos vertinimo metodai

Remiantis [2, 13] šaltiniais, šiuo metu finansinei rizikai vertinti yra nagrinėti tokie matematiniai statistiniai metodai ir sritys, kaip:

- tiesinė tikimybių analizė (angl., *the linear probability analysis*);
- tiesinė diskriminantinė analizė (angl., *linear discriminant analysis*);
- *logit* analizė (angl., *logit analysis*);
- *probit* analizė (angl., *probit analysis*);
- tiesinis programavimas;
- *integer* programavimas;
- „*risk of ruin*“ ir *option pricing* (sutr. OPM) modeliai;
- *mortality default rate* modeliai;

bei dirbtinio intelekto metodai ir sritys:

- artimiausio k kaimyno metodas *(*angl., *k-nearest neighbor*, sutr. KNN);
- klasifikavimo medis (angl., *classification tree*);
- dirbtiniai neuroniniai tinklai (angl., *artificial neural networks*, sutr. ANN);
- genetiniai algoritmai;
- paramos vektorių mašinos (angl., *support vector machines*, sutr. SVM);
- neuro-fuzzy sistemos;
- Bajeso metodas;
- hibridiniai ir kiti metodai.

Kitame skyriuje apžvelgiamas kreditų kokybės vertinimas balais ir aprašomas tiesinės diskriminantinės analizės metodas ir jo veikimo principai, o taip pat paliečiama logit analizė. Toliau skyriuose bus aptarti dirbtinio intelekto metodai.

1.10. Tiesinė diskriminantinė analizė ir kiti vertinimo balais metodai

Rizikos, atsirandančios išduodant paskolas individualiems asmenims ar mažoms organizacijoms, vertinimo atveju kredito kokybė paprastai įvertinama kredito vertinimu balais (angl., *credit scoring*): kreditorius iš turimos informacijos apie subjektą, kuriam išduodama paskola, apskaičiuoja kredito įvertinimą balais, tam naudodamas tam tikrą standartizuotą formulę. Pagal šį rodiklį kreditorius gali spręsti, ar jam verta suteikti kreditą konkrečiam subjektui, ar to daryti

neverta. Kaip jau minėta, reitingavimą naudoja ir kredito reitingavimo agentūros, tam naudodamos skirtingus rodiklius. Pavyzdžiui, „Standard & Poor's" reitinguodama skaičiuoja bendrą įsipareigotojo pajėgumą įvykdyti finansinį įsipareigojimą, ir tai atlieka apskaičiuodama įsipareigojimų neįvykdymo rodiklį. Tuo tarpu „Moody's" į vertinimą įtraukia ir sprendimą apie skolos atgavimą esant nuostoliams, todėl skaičiavimai artimi numatytų praradimų rodiklio skaičiavimams. Šios bendrovės, reitinguodamos bendroves, vadovaujasi tokiais kriterijais, kaip verslo šakos ar pramonės šakos rizika, subjekto vieta šioje šakoje, jo veiklos ir efektyvumo rodikliai, finansinis pajėgumas bei lankstumas ir kt. Šių bendrovių reitingavimo algoritmai nėra viešinami, todėl informacija apie juos prieinama tik šių bendrovių analitikams.

Kreditų rizikos vertinimas balais apima 4 aukščiau minėtus metodus: tiesinė tikimybių analizė, tiesinė diskriminantinė analizė, logit analizė, probit analizė.

Populiariausias kreditų vertinimo metodas, kuris paremtas kreditų vertinimu balais, yra diskriminantinė analizė (antra seka logit analizė, abu metodai teikia panašius efektyvumo rodiklius), kuri remiantis kai kuriais tyrimais rodo geresnius rezultatus už ekspertines bankų sistemas, kurios yra gana subjektyvios ir iš esmės remiasi keturiais kriterijais: skolininko reputacija, kapitalu, finansiniu produktyvumu ir įkeičiamo turto verte, taip vadinamų 4 „Cs" (angl., borrower character (reputation), capital (leverage), capacity (volatility of earnings), collateral) [13]. E.I.Altman ko gero yra žinomiausias mokslininkas su savo straipsniais tradiciniame kreditų rizikos valdyme, o ypač plėtojant diskriminantinę analizę, kuri jau 1968 metais buvo publikuota „Journal of Banking & Finance" [1], o vėliau plėtota dėl savo populiarumo eilę metų daugelio mokslininkų [13].

Diskriminantinės analizės principas – suskirstyti tam tikrus įvairius informacijos teikiamus atributus į grupes, kurie vienaip ar kitaip gali prisidėti prie rizikos įvertinimo. Vėliau vertinamos tų grubių reikšmingumas, daroma įtaka, paskiriant tam tikrus koeficientus. Iš esmės diskriminantinė funkcija atrodo taip:

$$Z=v_1x_1+v_2x_2+\ldots+v_nx_n, \qquad (1)$$

kur Z yra diskriminantinis įvertis (reikšmė), n – narių skaičius, $v_1,v_2,\ldots v_n$ yra diskriminantiniai koeficientai, o $x_1,x_2,\ldots x_n$ yra nepriklausomi kintamieji. Vėliau šis vienai kompanijai skaičiuojamas įvertis naudojamas kitam koeficientui apskaičiuoti, kuris vertina eilės kompanijų tarpusavio tendencijas. Kitaip tariant, diskriminantinė analizė bando susieti apskaitos ir rinkos duomenis tiesine priklausomybe su dvejomis skolininkų grupėmis – mokiais ir nemokiais. Tuo tarpu logit metodas apima tik apskaitos duomenų analizę nustatyti tikimybę, kad skolininkas neįvykdys įsipareigojimų, tariant, kad ta tikimybė yra pasiskirsčiusi logistiškai (angl., logistically distributed), t.y., įgyja reikšmes tarp 0 ir 1. Logit modelio tikimybės radimo formulė (2) pateikta žemiau:

$$\mathrm{E}[y_i|x_i] = \Phi\left(a_0 + a_1 x_{i1} + a_2 x_{i2} + ... + a_d x_{id}\right), \quad y_i = \{0, 1\} \tag{2}$$

2 formulėje *Φ(x)* yra kumuliacinė pasiskirstymo funkcija (angl., cumulative distribution function), kuri reiškia, kad su kiekvienu realiu x, galioja $x \mapsto F_X(x) = \mathrm{P}(X \leq x)$, kas reiškia, jog atsitiktinis X bus mažesnis arba lygus x. Analogiškai pateikiama probit modelio formulė (3) tikimybei rasti, jog objektas priklauso vienai ar kitai grupei (pvz., 0 ir 1):

$$\mathrm{E}[y_i|x_i] = \frac{1}{1 + \exp(-a_0 - a_1 x_{i1} - \ldots - a_d x_{id})} \tag{3}$$

Ir logit ir probit modeliams kredito vertinimo balais formulė yra tokia pati kaip ir tiesinei diskriminantinei analizei, t.y., formulė 1, o 2 ir 3 formulės atitinkamai parodo tikimybę, kad skolininkas neįvykdys įsipareigojimų.

Kiekvienas kredito riziką vertinantis ekspertas pats nustatinėja ir nepriklausomų kintamųjų skaičių, ir diskriminantinius koeficientus, tačiau E.I.Altman 1968 metais kaip optimaliausią variantą siūlė 5 kintamuosius: likvidumas, pelningumas, svertinės sistemos (angl., leverage, pelningumo įvertinimas), mokumas ir aktyvumas (tarkim, pardavimų). Šie kintamieji yra paprasčiausios grupės, sudarytos iš 22 atributų, rodiklių, gaunamų iš įvairių finansinių duomenų. Atitinkamai šiems kintamiesiems buvo siūlomos tokie diskriminantiniai koeficientai: 0.012, 0.014, 0.033, 0,006 ir 0.999. Jei Z koeficiento reikšmė yra mažesnė nei 2.675, ta kompanija skiriama prie bankrutuojančių grupės. Vėliau (1977) tas pats autorius siūlė naudoti diskriminantinį modelį su 7 kintamaisiais, kuris dar labiau buvo tobulinimas kitų autorių ir vienas iš tokių buvo Zeta modelis, kurį plėtojo viena firma, o jos efektyvumą lyginant su kitais analogiškais sprendimais dar kartą patvirtino Scott (1981) [13].

Nors esant tvirtoms ekspertinėms žinioms diskriminantinis modelis gali rodyti gerus rezultatus, tačiau šis metodas (kaip ir visi likę, paremti vertinimu balais) turi minusų, kurių pagrindiniai – subjektyvus didelio finansinių rodiklių kiekio įvertinimas bei gana sudėtingas tinkamos koreliacijos nustatymas, lėtas reagavimas į pokyčius rinkoje ir ekonomikoje ar pačios firmos, kadangi vertinimas vyksta tam tikrais laiko intervalais, tuo pačiu visi vertinimo kriterijai bandomi susieti tiesiškai, nors realus pasaulis toks nėra, o taip pat šiems metodams nėra konkretaus vieningo teorinio modelio.

1.11. Dirbtinis intelektas ir mašininis mokymas

Dirbtinio intelekto sąvoka atsirado prieš pusę amžiaus (1956 m.). Šeštajame XX amžiaus dešimtmetyje pradėti naudoti kompiuteriai šią sąvoką pavertė rimta teorine ir praktine disciplina, kuri pasirodė gerokai sudėtingesnė nei manyta iki tol. Tuo metu buvo sukurtos pirmosios dirbtiniu intelektu pagrįstos programos, *Logic Theorist, General Problem Solver (GPS),* leidusios panaudoti

šios srities žinias praktiniam problemų sprendimui. Dirbtinio intelekto sąvoka yra plati ir galbūt todėl sunkiai apibrėžiama. Dėl šios priežasties nėra vienareikšmio jos apibrėžimo, nes beveik kiekvienas šaltinis pateikia savo apibrėžimą ar net kelis apibrėžimus. Ši sąvoka gali būti apibrėžta ir kaip „AI – tai mokslas, kuris apima protinių sugebėjimų panaudojimą naudojantis kompiuteriniais modeliais“ (*Charniak and McDermott*)[3] arba „Dirbtinis intelektas – tai mokslas kaip išmokyti kompiuterius daryti kažką, ką tuo momentu geriau daro žmogus“ (*Rich*)[3], kas aprašo dirbtinį intelektą kaip „protingų“ mašinų kūrimo procesą. Sistemų, kurios elgiasi kaip žmonės, sąvoką atspindi *Tiuringo testas*, kurio pagrindinis subjektas yra pripažinimas, kad mašina gali mąstyti, atsakydama į žmogaus klausimus taip, kad jis nesupras, ar bendravo su kompiuteriu, ar su žmogumi. Tačiau dirbtinio intelekto sąvoką galima apibrėžti ir kaip „kompiuterių mokslo šaka, kuri automatizuoja protingą elgesį“ (*Luger and Stubblefield*)[3], kas jau savaime apibrėžia AI kaip atskirą mokslo šaką, nagrinėjančią įvairius sudėtingus skaičiavimus, kas leidžia šią sritį taikyti tiek finansuose, tiek ir įvairiose kitose srityse.

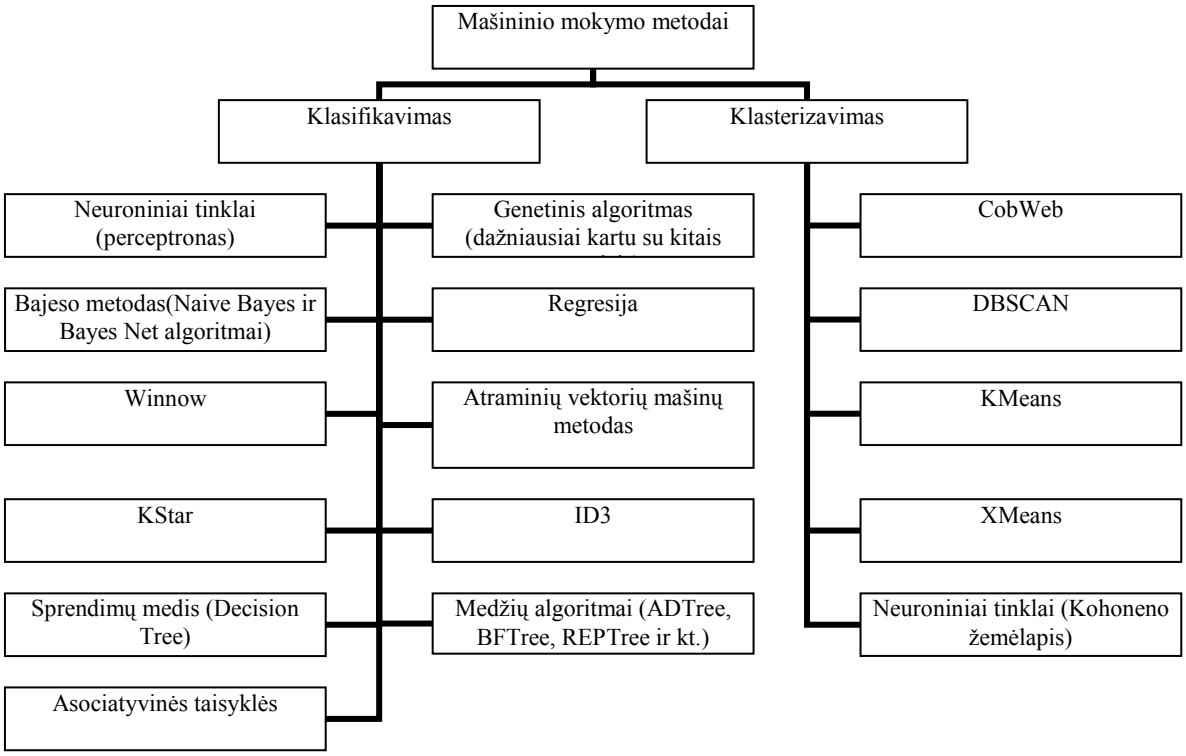
Dirbtinis intelektas kredito rizikos vertinime taikyti pradėtas palyginti neseniai, nors finansuose šios srities taikymas siekia IX dešimtmečio pradžią, kai buvo sukurta pirmoji finansinė ekspertinė sistema FOLIO, skirta vertybinių popierių portfelio analizei[2]. Vėliau buvo kuriama vis daugiau dirbtiniu intelektu pagrįstų sistemų, skirtų įvairiems finansiniams skaičiavimams bei prognozėms, kaip TARA, INSPECTOR, ARMAX ir t.t. Šiuo atveju gali būti priimami tokie sprendimai, kaip investavimas realiu laiku, akcijų kainų bei indeksų prognozavimas, vertybinių popierių pasirinkimas ir pan. Į tokių sprendimų grupę galima įtraukti ir nagrinėjamą sritį, kadangi tiek bankroto prognozavimą, tiek kredituojamų subjektų pasirinkimą galima priskirti prie tokio tipo uždavinių. Lietuvoje taip pat yra sukurta ekspertinė sistema akcijų rinkoms – STRASS [4]. Kaip atskirą tokių modelių grupę galima įvardinti ir dirbtinio intelekto srities modelius bei metodus, IX-ajame praėjusio amžiaus dešimtmetyje, atsiradus pirmosioms dirbtiniu intelektu paremtoms sistemoms, į šią kompiuterių mokslo šaką buvo atkreiptas didesnis dėmesys ir ji buvo imta rimčiau taikyti ekspertinių sistemų kūrimui. Problemų sprendimo laipsnis pagrįstas duomenų kokybiškumu bei žmogaus eksperto suformuluotomis taisyklėmis. Ekspertinės taisyklės sukurtos veikti žmogaus eksperto lygmenyje. Akivaizdu, kad tokio tipo sistemos gali būti labai naudingos žmogui, siekiančiam priimti svarbų sprendimą. Kaip jau buvo minėta, Amerikoje kai kurios finansinės institucijos jau dabar linkusios pasikliauti ne tik žmogiškąja išmintimi, bet ir programine įranga (TARA, INSPECTOR, kt.), paremta dirbtinio intelekto metodais (neuroniniais tinklais, fuzzy logika, saviorganizuojančiais tinklais ir kt.) [4].

Ši sritis vis dar tiriama ir plėtojama, todėl ekspertinės sistemos gali būti naudojamos tik kaip patariančios vartotojui, jų sprendimu negalima visiškai pasitikėti. Labiau ištobulėjus šiai sričiai, galima tikėtis, kad ateityje bus sukurta ir sistemų, kurios galės pateikti optimalų sprendimą.

Laisvoji enciklopedija Vikipedija apibrėžia mašininį (sistemos) mokymą taip: tai dirbtinio intelekto sritis, kuri apima metodų kūrimą, mokinančių kompiuterius „mąstyti“. Tiksliau kalbant, tai yra programų kūrimo būdas, kai sukurta sistema prisitaiko prie duomenų („apsimoko“). Sistemos mokymasis yra stipriai susijęs su statistika, nes sistemos mokymasis ir statistika nagrinėja duomenų analizę, bet skirtingai nuo statistikos, sistemų mokymasis yra susijęs su skaičiavimams naudojamų algoritmų sudėtingumais. Sistemos mokymosi algoritmai yra sistematizuoti pagal jų rezultatus. Bendri algoritmų tipai:

- prižiūrimas mokymasis (mokymasis su mokytoju) – algoritmas generuoja funkciją, kuri sieja įvedamus duomenis su tinkamais išvedimo duomenimis. Viena iš standartinių prižiūrimo mokymosi formuluočių yra klasifikavimo problema: sistema turi išmokti (kad aproksimuotų funkcijos elgesį) funkciją, kuri susieja vektorių $[X_1, X_2, \ldots X_N]$ su skirtingomis klasėmis, tikrinant kelis funkcijos įvedimo-išvedimo rezultatus.
- dalinai prižiūrimas mokymasis – apjungia pažymėtus ir nepažymėtus pavyzdžius atitinkamai funkcijai arba klasifikatoriui sugeneruoti.
- neprižiūrimas mokymasis – modeliuoja aibę įvedimų: pažymėti pavyzdžiai neprieinami.
- mokymasis su paskatinimu - algoritmas keičia elgsenos strategiją, priklausomai nuo pateiktų žinių apie pasaulį. Kiekvienas veiksmas turi įtakos aplinkai, o aplinka gražina informaciją, pagal kuria vadovaujasi besimokantis algoritmas.
- signalo keitimas – algoritmas panašus į prižiūrimą mokymąsi, bet skirtingai nuo jo nesiekia sukonstruoti tikslią funkciją. Bando nuspėti naujus išvedimus pagrįstus bandymų įvedimais, bandymų išvedimais ir naujais įvedimais.
- mokymasis mokytis – algoritmas nagrinėja savo induktyvų tendencingumą, pagrįstą ankstesne patirtimi.

Žemiau pateikiami vieni populiariausių klasifikavimo ir klasterizavimo metodai ().



Šaltinis: [19]

6 pav. Klasifikavimo ir klasterizavimo metodai

Pagal [20], mašininio mokymosi (angl., machine learning) tikslas yra kompiuterinės sistemos, kurios pačios galėtų mokytis save tobulinant su įgaunama patirtimi veikimo metu. Tokių sistemų sėkmingas kūrimas pagelbėtų daugelyje kompiuterių mokslo sričių ir ne tik: (namų) robotų taikymas, kalbos atpažinimo srityje, kur tokios sistemos galėtų prisitaikyti prie naujų kalbančiųjų ar aplinkos poveikio, žiniomis grindžiamų (angl., knowledge-based) sistemų bendradarbiaujant su ekspertais taikymas įvairiose sudėtingose sistemose ar net kompiuterinės programos, sugebančios skaityti ir vėliau pagal tai atlikti matematinius ar fizikos skaičiavimo veiksmus.

Žinoma, tai nėra vien tik tikslai ir svajonės, jau yra atlikta nemažai sėkmingų mašininio mokymo programų, kaip žymiai spartesnis skaičiavimų atlikimas išmokstant reikalingas taisykles, taisyklių įsisavinimas medicinoje – ligų nustatymas pagal turimus simptomus, kalbų atpažinimai ir pan., pateikiami keletas iš jų:

- Medžių algoritmais besiremianti ID3 sistema, kuri 99% tikslumu atpažįsta skydliaukės ligas ir yra gretinama ekspertinėms sistemoms. Panašius rezultatus pateikia analogiškos sistemos geltai ir limfinėms ligoms nustatyti.
- Soar paieškos sistema, kuri vietoje reikalingų 1731 žingsnių panaudoja 7 paieškos žingsnius ir yra tapatinama žmogaus programuotoms paieškos sistemoms.

- Klasifikavimo programos (dažniausiai naudojančios statistinius metodus), apdorojančios didelius kiekius informacijos, kaip tarkim astronomijoje atpažįstant naujus objektus.

Tai tik keletas iš įvairių įgyvendintų sprendinių ir dalis jų yra įgyvendinta jau iki 1990, todėl šiuo metu tikėtina yra efektyvesnių analogiškų realizacijų.

Gana lėtas šios srities vystymasis susijęs su daugeliu objektyvių priežasčių, kaip kelerių mokslo sričių apjungimas ir suderinimas, generalizavimo savybių neefektyvumas, kai apmokytas objektas patenka į naują aplinką ir nebegali atlikti analogiškų uždavinių ir t.t. Žemiau pateikiami keletas pavyzdžių, perteikiančių įvairias kylančias problemas sėkmingam mašininiui mokymui.

Taigi kaip pavyzdys aptariamas besimokantis namų robotas, kuris tarkim vykdytų iš pažiūros paprastą veiksmą: namų šeimininkui liepus atnešti akinius ar telefoną, jis tai ir padarytų, bet tokiai užduočiai reikia įgyvendinti daug aspektų, kaip kelio radimas, kliūčių išvengimas, suvokimas ir manipuliacijas. Žinoma, užprogramuoti robotą daryti keletą užduočių viename ar keleriuose namuose vienam ar keliems šeimininkams yra įgyvendinama užduotis, tačiau bendru atveju jis turi suprasti kiekvieną komandą, nors tai kiekvieną kartą pasakytų kitas žmogus su savo akcentu ir pan., atpažinti įvairius objektus įvairiomis apšvietimo sąlygomis, sugebėti skirtingų išmatavimų daiktą (pvz., stiklinę vandens, grandinėlę, laikraštį ir t.t.) paimti ir juo manipuliuoti. Tokioms užduotims įvykdyti robotas turėtų sugebėti mokytis iš besikeičiančios aplinkos ir įvairių variantų kaip galima įgyvendinti problemą toje aplinkoje, tarkim, kur dažniausiai būna akiniai, kurios durys dažniausiai užrakintos ir t.t. To galėtų mokytis keleriais būdais, kaip tiesioginiu aplinkos tyrinėjimu, gaunant nurodymus ir patarimus iš žmogaus, aktyvaus eksperimentavimo toje aplinkoje. Taigi iš pažiūros paprastai užduočiai įgyvendinti reikia eliminuoti eilę problemų.

Kitu pavyzdžiu galėtų būti įvairių sistemų gebėjimas bendrauti su žmogumi, t.y., ne tik atpažinti žmogaus kalbą, tačiau ją ir suprasti, suvokti, tada būtų galimas dialogas bei tam tikri veiksmai (pvz., mašinų remonto darbai, automatinis klientų aptarnavimas ir pan.), bet tam įgyvendinti kyla eilė problemų: atpažinti skirtingų kalbančiųjų šnekas, akcentus, dialektus ar jau įsisavintus kalbančiuosius naujoje aplinkoje, gebėti mokytis naują žodyną ir gramatiką, o taip pat gramatinius natūralios kalbos aspektus. Taip pat kyla problemos ne tik su kalbos atpažinimu, bet ir suvokimu ko nori žmogus, kokio tipo problemą reikia spręsti bei naujų problemų įsisavinimas [20].

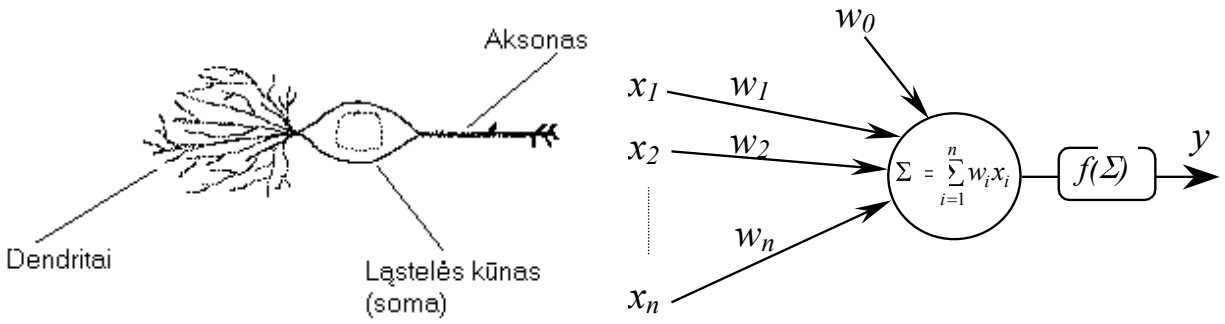
Taigi įvairiose mašininio mokymo srityse egzistuoja aibės skirtingų problemų, kurias išsprendus po vieną būtų galima jungti į skirtingų kompleksinių užduočių sprendimus įvairiose gyvenimo srityse kaip ekonomikoje, medicinoje, kariuomenėje, švietime ir t.t.

1.12. Dirbtiniai neuroniniai tinklai

Toliau aprašant dirbtinius neuroninius tinklus (angl., artifiacial neural netwok, sutr. ANN) remiamasi R.Simučio paskaitų medžiaga [4], kuri parengta pagal (A. Verikas ir kt., 2002).

Dirbtiniai neuroniniai tinklai pradėti nagrinėti šeštajame dešimtmetyje, tačiau iki devintojo dešimtmečio vidurio jie nebuvo plačiai naudojami. Tik išradus greitus ir galingus apmokymo mechanizmus, dirbtiniai neuroniniai tinklai galėjo spręsti realius uždavinius. Šiuo metu neuroniniai tinklai naudojami signalų apdorojimui, triukšmo eliminavimui, duomenų klasifikavimui, sistemų modeliavimui, identifikavimui, prognozei ir kontrolei.

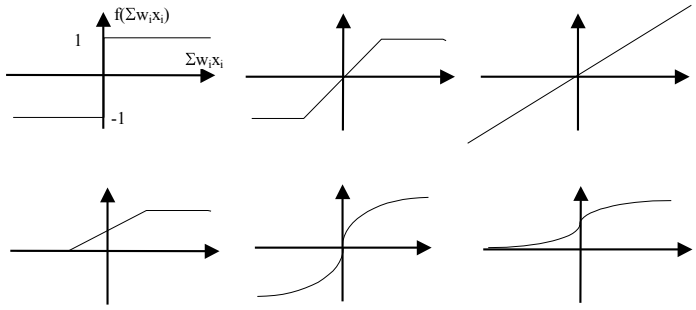
Biologinio neurono ir dirbtinio neurono sandaros pateiktos žemiau esančiame paveiksle (6 pav.), kur neurono įėjimai pažymėti $x_1 \ldots x_n$, svoriai pažymėti $w_1 \ldots w_n$, w_0 – slenksčio reikšmė, $f()$ – perdavimo funkcija, y – neurono išėjimas.



Šaltinis: [4]

7 pav. Biologinis (kairėje) ir dirbtinis neuronai

Neurono veikimas: neuronas gauna keletą įėjimo reikšmių, kiekviena įėjimo jungtis turi savo perdavimo koeficientą (svorį w), šie svoriai atitinka biologinio neurono sinapsių efektyvumą. Kiekvienas neuronas turi savo slenksčio reikšmę. Neurono sužadinimo reikšmė formuojama skaičiuojant svorinę įėjimo signalų sumą ir atimant slenksčio reikšmę. Pagal sužadinimo signalą, naudojant neurono perdavimo funkciją skaičiuojama neurono išėjimo reikšmė. Jeigu naudojama šuolinė neurono perdavimo funkcija (neurono išėjimas lygus 0, jei aktyvacijos reikšmė mažiau už nulį, ir lygus 1, jei reikšmė didesnė ar lygi nuliui), neuronas veikia kaip biologinis neuronas. Neurono svoriai gali būti neigiami. Neigiamas svoris reiškia, kad jungtis turi slopinantį efektą, tokie neuronai egzistuoja ir biologinėse sistemose

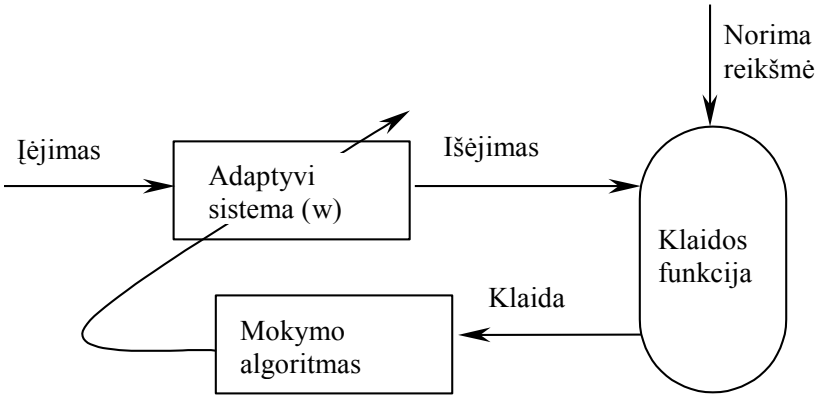


Šaltinis: [4]

8 pav. Neuronų perdavimo funkcijos

7 paveiksle parodyta keletas neurono perdavimo funkcijų. Kai kurios funkcijos būdingos tam tikro tipo neuroniniams tinklams, kitos sąlygojamos mokymo taisyklių arba parenkamos pagal sprendžiamą uždavinį, tačiau dažniausiai naudojama hiperbolinio tangento funkcija.

Neuroninių tinklų adaptyvumas (8 pav.) yra geriausioji jų savybė. Neuroninės sistemos kaip ir matematiniai metodai, tokie kaip diskriminantinė analizė, siekia susieti įėjimo parametrus (kintamuosius), tačiau skirtingai nuo diskriminantinės analizės, įėjimai siejami ne tiesiškai, taigi šios sistemos nėra sudaromos naudojant išankstines žinias pagal specifikaciją, formules ar aprašymą. Vietoje to, sistema naudoja išorinius duomenis savo parametrų nustatymui. Kaip jau buvo minėta, neuroniniai tinklai apmokomi žinant įėjimo ir atitinkamas išėjimo reikšmes (užduoties reikšmės), grąžinamas per ryšį, kuriame naudojama klaidos funkcija.



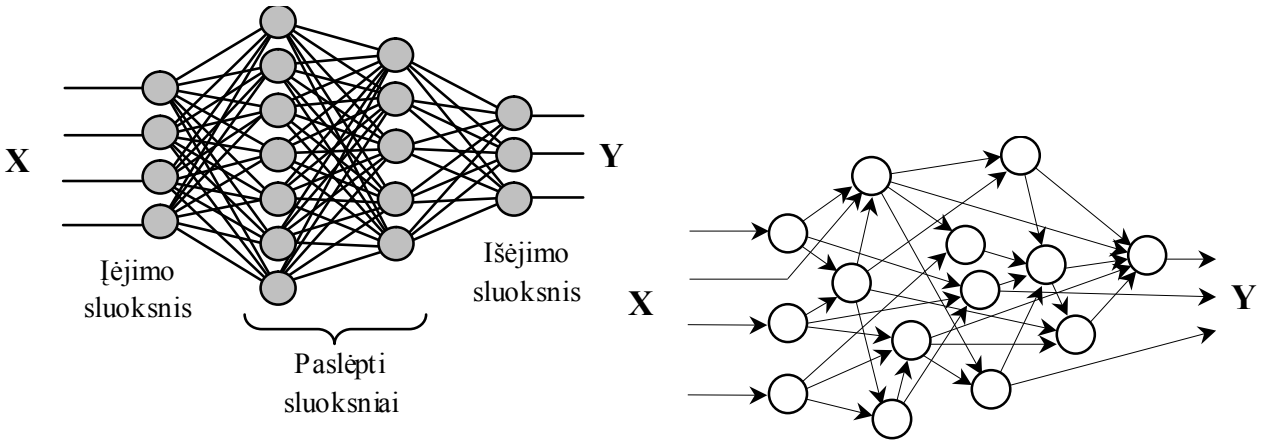
Šaltinis: [4]

9 pav. Adaptyvios sistemos kūrimas

Klaidos funkcija labai dažnai yra skirtumo funkcija tarp neuroninio tinklo išėjimo ir užduoties reikšmės. Neuroninio tinklo atsako tikslumas tiesiogiai naudojamas parametrų (tinklo svorių) keitimui. Mokymo metu svoriai keičiami taip, kad sistemos išėjimo reikšmės artėtų prie norimų reikšmių (mažėtų klaida).

Aukščiau aprašyti pavieniai neuronai jungiami į tinklus. Bet kokios paskirties neuroninis tinklas turi įėjimus, perduodančius kintamųjų reikšmes iš išorės, ir išėjimus, formuojančius tinklo

atsaką. Dažnai egzistuoja ir tarpiniai (paslėpti) neuronai atliekantys vidinį vaidmenį tinkle. Įėjimo, paslėpti ir išėjimo neuronai jungiami vieni su kitais. Galimi du jų jungimo variantai: jungiant juos į sluoksnius ir nejungiant (9 pav.)



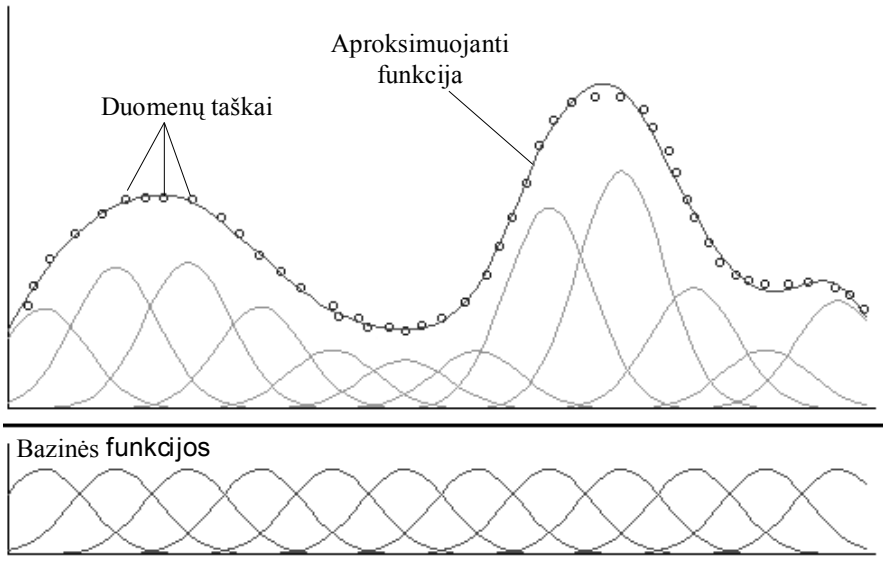
Šaltinis: [4]

10 pav. Sluoksniuotas (kairėje) ir nesluoksniuotas neuroniniai tinklai

Piešinyje (kairėje) parodytas įėjimo sluoksnis nėra skaičiuojančių neuronų sluoksnis. Šio sluoksnio paskirtis tėra įvesti į tinklą įėjimo kintamųjų reikšmes. Toks pirmojo sluoksnio tinkle vaizdavimas yra įprastas, kadangi prieš duomenų apdorojimą neuroniniu tinklu dažnai seka pirminio duomenų apdorojimo etapas (pavyzdžiui, normavimas, centravimas). Pavyzdyje paslėptų ir išėjimo sluoksnių neuronai sujungti su visais prieš tai esančio sluoksnio neuronais, bet neuronai gali būti sujungiami ne su visais gretimų sluoksnių neuronais. Vieno sluoksnio neuronai dažniausiai turi tą pačią perdavimo funkciją.

Įėjimo reikšmės patenka į įėjimo sluoksnį, vėliau paeiliui skaičiuojamos paslėptų sluoksnių neuronų išėjimo reikšmės, po to išėjimo sluoksnio neuronų išėjimo reikšmės. Kiekvienas neuronas skaičiuoja svorinę prieš tai esančio sluoksnio neuronų išėjimų sumą ir atima slenksčio reikšmę, taip gaudamas sužadinimo lygį. Pagal gautąją sužadinimo reikšmę skaičiuojama neurono perdavimo funkcijos (išėjimo) reikšmė. Išėjimo sluoksnio neuronų išėjimo reikšmės laikomos neuroninio tinklo išėjimo reikšmėmis.

Kita populiari neuroninių skaičiavimų funkcija – radialinė bazinė funkcija. Naudojamos kelios bazinių funkcijų formos, kaip sinusoidės, bangelės, tačiau labiausiai paplitusi Gausinė, pavyzdys pateiktas 10 paveiksle [4].

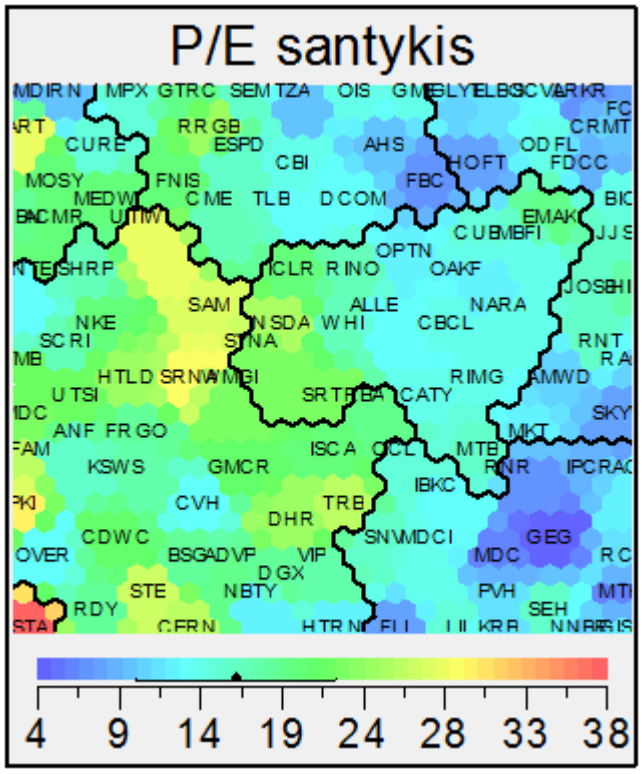


Šaltinis: [4]

11 pav. Radialinių bazinių (gausinių) funkcijų naudojimas

Įvairūs neuroninių tinklų tipai naudoja skirtingus mokymo metodus. Egzistuoja du pagrindiniai neuroninių tinklų apmokymo tipai: mokymas su mokytoju (neuroninis tinklas turi išmokti nežinomą funkciją iš pavyzdžių, kurie yra mokymo imties įėjimo ir išėjimo vektorių rinkiniai X ir Y) ir mokymas be mokytojo (turima tik įėjimo duomenų imtis X), iš kurių labiau paplitęs yra mokymas su mokytoju (egzistuoja ir trečias mokymo tipas: *reinforcement learning*). Mokymas be mokytojo dažniausiai naudojamas atpažinimui arba klasifikacijai. Populiariausi mokymosi be mokytojo principai Hebbo ir Kohoneno, kitaip dar vadinami save-organizuojančiais tinklais (angl., self-organizing maps, sutr. SOM) [4]. Pastarojo principais remiasi Viscovery® SOMine produktas, skirtas duomenų klasterizavimui, palaikantis įprastus *.txt ar *.xls failų formatus. Žemiau pateiktas šio produkto naudojimo pavyzdys (11 pav.), kur kompanijos suskirstytos į klasterius pagal skirtingus parametrus, kurie skirtingai įtakojo klasterių atsiradimą.

Šiuo metu žinomi tokie mokymo (svorių nustatymo) algoritmai: gradientinis nusileidimas (įtraukiant ir momentum narį bei mokymosi greičio keitimą), tyrimas tiesėje (vienparametrinė paieška), jungtinis gradientas (keičiant svorių paieškos kryptį, gradientas senąja kryptimi nekinta), Niutono (paremtas antros eilės išvestinių taikymu) ir Kvazi-Niutono metodai.



Šaltinis: sukurta autoriaus

12 pav. Klasterizavimas naudojant Viscovery® SOMine

Reiktų atkreipti dėmesį, kad SOM tyrimai buvo atlikti ir Vilniaus universitete, Kauno humanitariniame fakultete (Merkevičius, Garšva, Simutis, 2004; „Forecasting Of Credit Classes With The Self-Organizing Maps“), kur bandyta suskirstyti kreditus 5 „gerus“ ir „blogus“. R.Simutis yra prisidėjęs ir daugiau prie dirbtinių neuroninių tinklų panaudojimo atvejų tyrimų, galima paminėti keletą iš jo darbų: „Estimation of biomass concentrations in fermentation processes for recombinant protein production“, „New Methods and Technologies for Deflection Yoke Tuning“, „Discrete Eye Tracking Based on PCA and Neural Classifier“, „Cash Demand Forecasting for ATM Using Neural Networks and Support Vector Regression Algorithms“, „Intelligent Cash Management System for an ATM network“, „Application of Particle Swarm Optimization Algorithm in Stock Markets“, „Development and Evaluation of Decision-Making Model for Stock Markets“. Taip pat reiktų paminėti E.Merkevičiaus daktaro disertaciją „Savitvarkių neuroninių tinklų-diskriminantinio modelio tyrimai kredito rizikos vertinimo sprendimų paramos sistemoje“.

Dirbtinių neuroninių tinklų pritaikymų yra be galo daug, tarp jų nemažai finansų rinkoje, įskaitant kreditų rizikos valdyme, kai kurie iš jų pateikiami žemiau esančioje lentelėje.

3 lentelė

ANN taikymas kreditų rizikai vertinti

Autoriai	**Metai**	**Aprašymas**
Glorfeld, Hardgrave	1996	Naudoti ANN su Gauso maksimalaus tinkamumo klasifikatoriais.
Deng	1993	Sinergetinis naujų faktų gavimo iš esamų atmintyje (*Memory-Based Reasoning*)

Autoriai	Metai	Aprašymas
		bei taisyklėmis paremto induktyvaus mokymosi (*Rule-Based Inductive Learning*) metodų integravimas su neuroniniais tinklais į ekspertinę sistemą.
Williamson	1995	Genetinio algoritmo, automatiškai parenkančio optimalią konfigūraciją bei rezultatus, panaudojimas.
Arminger, Enache, Bonne	1997	Klasifikavimo metodų apjungimas.
Handzic, Tjandrawibawa, Yeo	2003	Naudotas daugiasluoksnio perceptrono neuroninio tinklo tipas kartu su vidurkinimu (*Ensemble Averaging)* bei pagreitinimu filtruojant (*Boosting by Filtering),* kas leido gauti tikslius rezultatus (paskutiniu atveju procentinė klaida siekė 1.32 %).
di Tollo	2006	Naudota mokymo su mokytoju paradigma.
Besens, Van Gestel, Stepanova, Van den Poel	2004	Naudota "išlikimo analizės" paradigma, paremta tam tikros populiacijos išlikimo rodikliu.
Keung Lai, Yu, Wang, Zhou	2006	Naudojamas trijų fazių neuroninių tinklų metodas blogų ir gerų kreditorių atskyrimui. Jame naudojama daug skirtingų neuroninių tinklų modelių, kuriuos sukūrus, taikomas iškoreliavimo maksimizavimo algoritmas tinkamų aibės narių išrinkimui; čia klasifikavimui naudojamas ir patikimumu paremtas metodas. Kaip teigia patys autoriai, atlikę eksperimentą, jie gavo gerus klasifikavimo rezultatus.
Keung Lai, Yu, Wang, Zhou	2006	Šis darbas traktuoja neuroninius tinklus kaip metamokymo metodą, kurio esmė yra skirtingų neuroninių tinklų pagal skirtingus apsimokymo duomenų rinkinius su skirtingomis pradinėmis sąlygomis bei apsimokymo algoritmais apmokymas skirtingų kredito vertinimo modelių (bazinių modelių) formulavimui. Iš šių bazinių modelių gali būti generuojamas metamodelis, kas leidžia pagerinti patikimumą, t.y., tikslumą prognozuojant praradimus.
B.M.M. Dissananayake, C. H. Hendahewa, A. S. Karunananda	2007	ANN taikymas atskiroms pramonės šakoms, šiuo atveju buvo pasirinkta viešbučių (turizmo) rinka, įvertinant finansinius ir ekonominius faktorius, naudojant du neuroninius tinklus: vieną grynųjų pinigų cirkuliacijai numatyti, kitą skolų numatymui, abejiems naudojant sklidimo atgal apmokymą (angl., backpropagation), suderinant neuroninių tinklų modelius su web aplikacijomis duomenims gauti.
Lean Yu, Shouyang Wang and Kin Keung Lai	2007	Daugiafazio (6 fazės) neuroninio modelio taikymas kreditų rizikai vertinti. Pirmame etape vykdomas duomenų generavimas, jų prastinimas, vidurkinimas ir t.t., antrame – sukuriama keletas skirtingų neuroninių tinklų su skirtingais treniravimosi duomenų rinkiniais, tinklai treniruojami, gaunami klasifikavimo ir patikimumo įvertinimai, dekoreliacijos maksimizavimo algoritmo naudojimas tinkamų bendrų narių parinkimui, penktame – pastarųjų narių sudarytų neuroninių tinklų patikimumo įvertinimų skirstymas į intervalus, šeštame – sudaromas galutinis modelis ir vertinamas jo patikimumas. Galiausiai modelis testuojamas su realiais duomenimis.
Eliana Angelini, Giacomo di Tollo and Andrea Roli	2008	Atsižvelgta į Bazelio II rekomendacijas kapitalo reikalavimams bankų knygoms, atsižvelgiant į kreditų riziką. Sukurtos dvi neuroninės sistemos (viena įprasta feedforward, kur neuronai turi vienos krypties ir neturi grįžtamųjų ryšių). Parodyta, kad dirbant su realiais duomenimis gauti neblogi skolininko galimų įsipareigojimo nevykdymo nuspėjimo rezultatai.
Maximilian J.B. Hall, Dadang Muljawan, Suprayogi, Lolita Moorena	2008	ANN naudojimas bankų kreditų portfelių rizikai vertinti, naudojant makroekonominius kriterijus. Įtrauktos Bajeso reguliacijos technikos, jog duomenys nepersimokytų (per daug neprisitaikytų prie apmokymo duomenų). Parodyta, jog iš naudotų parametrų (BVP augimas, akcijų kainosinfliacijos dydis, valiutų kursai ir pinigų apyvarta) didžiausią įtaką daro akcijų kainos.
Ping Yao, Chong Wu, Minghui Yao	2009	Šešių lygių fuzzy neuroniniais tinklas su 4 faktorių įėjimais. Parodyti, autorių teigimu, žymiai geresni rezultatai nei naudojant įprastą ANN modelį.

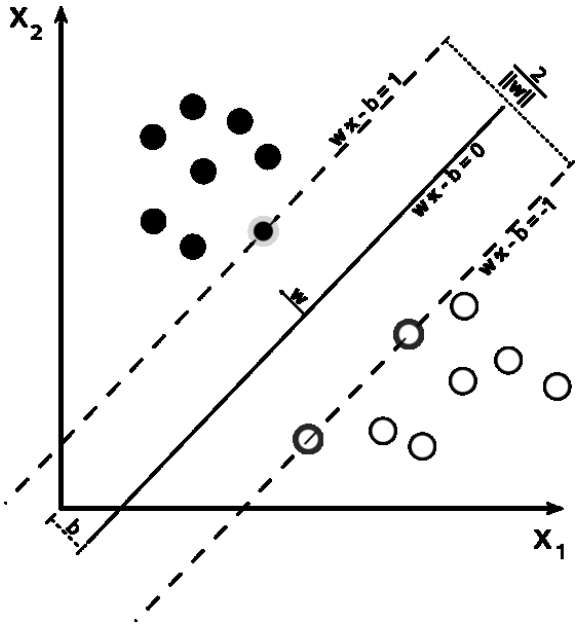
Šaltinis: [19], papildyta autoriaus

Apibendrinant neuroninius tinklus, galima pasakyti, kad neuroninis tinklas negali padaryti nieko, kas negali būti padaroma tradiciniais skaičiavimo metodais. Tačiau neuroniniu tinklu galima

padaryti tai, ką kitais metodais išgauti kartais labai sunku. Neuroninis tinklas gali formuoti modelį iš pavyzdžių. Tai ypač naudinga apdorojant daviklių duomenis arba duomenis iš sudėtingų (cheminių, gamybos, komercinių) procesų. Galbūt, daugeliu atvejų ir egzistuoja duomenų apdorojimo algoritmas, bet dažnai jis nėra žinomas arba turi pernelyg daug kintamųjų, todėl paprasčiau, naudojant turimus duomenų pavyzdžius, apmokyti neuroninį tinklą. Akcentuojamos problemos su ANN yra šios: neaiškus rezultatų gavimas (angl., ad hoc theoretical foundation), neaiškus koreliacijos tarp įvedamų kintamųjų tarpusavyje ir su išėjimais nustatymas, tų koreliacijų ar principų iš modelio išgavimas, lokalių minimumų sprendimas, neuroninio tinklo optimalus dydis nors yra įrodyta (Hecht-Nielsen ir Kolmogorov teoremos) kad su dideliais techniniais ištekliais visas problemas galima spręsti su dviem paslėptais neuroniniais sluoksniais, o taip pat neuroniniai modeliai nenumatytai elgiasi atvejais, kai gauti duomenys visiškai neatitinka apmokymo duomenų.

1.13. Atramos vektorių mašinos (SVM)

Atramos vektorių mašinos (engl., support vector machines) yra vienas iš populiarių mašininio mokymosi klasifikavimo metodų. Kaip ir ANN, SVM pagrįstas mokymusi iš duotų pavyzdžių, kurie šiuo atveju skiriami į du vektorius n-dimensinėje erdvėje, kur SVM sukuria hiperplokštumą, kuri atskiria pastaruosius du įėjimo duomenų rinkinius. Dažniausiai klasifikuojama į dvi klases, sąlyginai vieną žymint kaip teigiamą, kitą kaip neigiamą, analogiškai 1 ir -1. Vizualiai SVM principas pateiktas 12 paveiksle, kur w – tai normalinis vektorius (angl., normal vector), išraiška {x: wx+b = +1} – tai "teigiama" plokštuma, o "neigiama" plokštuma žymima kaip {x: wx+b = -1}, klasifikatoriaus riba – {x: wx+b = 0}. Parametras $\frac{b}{\|\mathbf{w}\|}$ yra renkamas toks, kad atstumas tarp teigiamų ir neigiamų plokštumų būtų kuo didesnis, taip gaunamos mažesnės generalizavimo klaidos [19, 21, 22].



Šaltinis: http://upload.wikimedia.org/wikipedia/commons/2/2a/Svm_max_sep_hyperplane_with_margin.png

13 pav. Tiesinis SVM

Įprastai SVM metodas aprašomas [21]: duotos apmokymo duomenų poros (x_i,y_i), i=1,...,l, kur $x_i \in R^l$, o $y \in \{1,-1\}^l$, tada SVM optimizavimo formuluotė atrodo taip (4):

$$\begin{aligned} \min_{\mathbf{w},b,\boldsymbol{\xi}} \quad & \frac{1}{2}\mathbf{w}^T\mathbf{w} + C\sum_{i=1}^{l}\xi_i \\ \text{subject to} \quad & y_i(\mathbf{w}^T\phi(\mathbf{x}_i) + b) \geq 1 - \xi_i, \\ & \xi_i \geq 0. \end{aligned} \tag{4}$$

Pateiktoje išraiškoje C>0 yra baudos parametras klaidos atžvilgiu. Funkcija *Φ* treniravimosi duomenis x_i žymi (angl., mapping) į aukštesnės dimensijos erdvę. $K(\mathbf{x}_i,\mathbf{x}_j) \equiv \phi(\mathbf{x}_i)^T\phi(\mathbf{x}_j)$ yra branduolio, kitaip vadinama kernel, funkcija. Yra keturios pagrindinės kernel funkcijos[21]:

- tiesinė: $K(\mathbf{x}_i,\mathbf{x}_j) = \mathbf{x}_i^T\mathbf{x}_j$;
- polinominė: $K(\mathbf{x}_i,\mathbf{x}_j) = (\gamma\mathbf{x}_i{}^T\mathbf{x}_j + r)^d,\ \gamma > 0$;
- radialinė bazinė funkcija (RBF): $K(\mathbf{x}_i,\mathbf{x}_j) = \exp(-\gamma\|\mathbf{x}_i - \mathbf{x}_j\|^2),\ \gamma > 0$;
- sigmoidinė: $K(\mathbf{x}_i,\mathbf{x}_j) = \tanh(\gamma\mathbf{x}_i{}^T\mathbf{x}_j + r)$;

Šiose funkcijose γ, *r* ir *d* yra kernel parametrai. Polinominės dar būna homogeninės ir nehomogeninės, atitinkamai kai r=0 ir r=1 [22].

Šaltinis [21] siūlo tokią SVM modelių kūrimo eigą, ypatingai tiems, kurie nėra itin patyrę šioje srityje:

- Eksperimentinius duomenis transformuoti į SVM formatą. Visi duomenys SVM metodui privalo būti transformuojami į realių skaičių vektorius, tarkim yra trijų kategorijų (reikšmių) atributas {atmestas, tikrinamas, patvirtintas}, tada jį būtų tikslinga (0,0,1), (0,1,0) ir (1,0,0), tačiau jei atributas turi daug reikšmių, tada tikslinga tik vieną kategoriją laikyti lygią 1, o kitas kaip 0.
- Atlikti duomenų normavimą į reikšmes nuo -1 iki 1 arba nuo 0 iki 1. Tai svarbus etapas, kuris naudojamas ne tik SVM metode, jo esmė yra ta, kad didelės arba stipriai aptriukšmintos reikšmės nebūtų dominuojančios visų kitų atžvilgiu, o taip pat šis etapas palengvina aritmetinius skaičiavimus, ypač naudojant polinominius ir tiesinius branduolius.
- Pradiniu branduoliu pasirinkti RBF, kur $K(\mathbf{x}, \mathbf{y}) = e^{-\gamma\|\mathbf{x}-\mathbf{y}\|^2}$. RBF yra turi pranašumą prieš tiesinį branduolį būtent dėl tiesinių ir netiesinių ryšių tarp klasių ir atributų. Tiesinis branduolys yra specialus RBF atvejis, o tinkamų parametrų parametrų parinkimas atsiremia į vienodų rezultatų gavimą. Be to, sigmoidinis branduolys su atitinkamais parametrais rodo panašius rezultatus kaip ir RBF. Tačiau, jei funkcijų skaičius yra didelis, RBF gali taip pat netikti, tada geriau rinktis tiesinį kernel.
- Naudoti kryžminį C ir γ parametrų patvirtinimą (atrinkimą, angl., cross-validation), tai daroma atsižvelgiant į generalizavimo savybes, o ne tik į apmokymo tikslumą, t.y., treniravimo duomenis padalinti į dvi dalis, kurių vienos dėka žiūrima kaip tiksliai klasifikuojami žinomi įėjimo duomenys, o kita – nematyti duomenys (generalizavimo įvertinimas). Siekiama, jog generalizavimo savybės būtų kuo geresnės. Tai apsaugo nuo modelio prisitaikymo prie apmokymo duomenų. Parametrų nustatymui rekomenduojama naudoti įprastą gridinę paiešką (pvz., keičiant C ir γ eksponentiškai didėjimo tvarka, angl., grid-search), nors galimi ir kiti įvairūs euristiniai metodai, kurie gali būti kur kas spartesni.
- Naudoti geriausius gautus C ir γ parametrus visam treniravimo duomenų rinkiniui.
- Testuoti;

4 lentelė

SVM taikymas kreditų rizikai vertinti pastaraisiais metais

Autoriai	Metai	Aprašymas
Liwei Wei, Jianping Li, Zhenyu Chen	2007	Kredito rizikos vertinimas naudojant SVM metodą su sumaišytais branduoliais, skirstant kreditus į „gerus“ ir „blogus“ ir, autorių teigimu, sėkmingai išbandyta su realaus JAV banko duomenimis.

Yanwen Dong	2007	SVM pritaikymas skolininkų rizikai vertinti. Pritaikyta mažoms kompanijoms vertinti, kurių finansiniai duomenys yra sunkiai gaunami, todėl vertinama naudojantis kasdienių operacijų duomenimis, tokiais kaip pardavimai, klientų mokėjimai, vėluojančių mokėjimų sumos ir kt. Pasak autorių, gauti geri rezultatai, o taip pat buvo palyginta su LDA metodu.
Härdle Wolfgang, Moro Rouslan A., Schäfer Dorothea	2008	SVM metodas taikytas kreditų rizikai įvertinti realiais Deutsche Bundesbank duomenimis. Gauti rezultatai palyginti su tradiciniais diskriminantinės analizės ir logit regresijos metodais, SVM metodas pasirodė geriau visais testuotais kintamaisiais.
Pei-Yi Hao, Min-Shiu Lin, Lung-Biao Tsai	2008	SVM metodas apjungiamas su Fuzzy teorija, kuri naudojama įėjimo duomenims ir optimizavimo problemai spręsti, taip kad kiekvienas įėjimo taškas turėtų įtakos sprendimų paviršiui. Parametrai taip pat parenkami Fuzzy nariais.
Leilei Zhang, Xiaofeng Hui	2009	Sukurtos kelios aplikacijos (SVM ir BNN - backpropagation neural network), kurių nuspėjimo tikslumas siekia 80% su tikrais Australijos ir Vokietijos duomenų rinkiniais.
Ligang Zhou, Kin Keung Lai, Lean Yu	2009	Orientuojamasi į paieškos metodus SVM optimaliausiems parametrams nustatyti, epicentre tiesioginis paieškos metodas, lyginant su gridine paieška, genetine paieška ir eksperimentiniu dizainu. Testai atlikti su dvejais realaus pasaulio duomenų rinkiniais.

Šaltinis: sukurta autoriaus

Aukščiau pateiktoje lentelėje pateikta tik keletas iš daugelio SVM metodo taikymo kreditų rizikai vertinti darbų, taigi SVM kreditų rizikoje yra itin plačiai naudojamas ir plėtojamas šiai problemai spręsti. Daugelyje straipsnių sprendžiama problema yra parametrų optimizavimas, taisyklių išgavimas, apjungimas su kitais metodais geresniems rezultatams gauti, tačiau reikia paminėti vieną iš didžiausių šio metodo trūkumų – tai praleistų, tuščių duomenų netoleravimas.

1.14. Bajeso metodas

Pirmą kartą Bajeso metodo sąvoka paminėta Pearl (1985), nors Bajeso teorema jau žinoma nuo 1763, kai savo moksliniame rašte (*„An Essay towards solving a Problem in the Doctrine of Chances. <...>“*) ją aprašė Thomas Bayes, kurio vardu ši tikimybių teorema ir pavadinta. Dabar Bajeso metodas yra labai plačiai naudojamas įvairiems sprendimo priėmimo uždaviniams spręsti, tarp jų ir finansinio sektoriaus problemoms spręsti.

Bajeso metodas tampa vis svarbesniu objektu dirbtiniame intelekte. Pirmiausia Bajeso tinklai atsirado ir išsivystė siekiant įvesti tikimybes į ekspertines sistemas ir tai vis dar yra vienas iš labiausiai naudojamų metodų. Žymus pavyzdys yra ekspertinė sistema sukurta medicininiais tikslais – QMR-DT (*a decision-theoretic reformulation of the Quick Medical Reference* (QMR) *model*), kurios tikslas – kuo tiksliau nustatyti kiekvienos ligos tikimybę, žinant tam tikrus simptomus, o kai kurių nežinant [9].

Bajeso tinklai labai plačiai naudojami buvo Microsoft produktuose, tokiuose, kaip Answer Wizard of Office 95 (pradėtas kurti 93 metais), Office Assistant of Office 97, virš 30 Technical Support Troubleshooters [9, 10].

Dar vienas žymus realizavimas yra Eric Horvitz kurtoje programoje „Vista“, kuri yra sprendimų teorijos sistema, naudota NASA misijų kontrolės centre keletą metų (plačiau:

http://www.research.microsoft.com/research/dtg/horvitz/vista.htm) [9]. NASA į Bajeso panaudojimo tyrimus yra investavusi gana didelius pinigus, pvz., ji naudota kuriant sistemas pilotavimui, kurios pagalba pilotui būtų suteikiama kuo daugiau informacijos apie objektus, į kuriuos tuo metu jis kreipia didžiausią dėmesį (akių žvilgsniai, galvos pasukimo trajektorija ir pan.). Esmė – šio metodo pagalba atskirti, suklasifikuoti visiškai „žalią“ informaciją (neįsikišant į sistemos darbą) [10].

Specializuoti Bajeso tinklų variantai nepriklausomai naudojami daugelyje sričių, kaip pvz., genetikoje (ryšių analizėse), kalbos atpažinime (angl., *speech recognition*), sekime, stebėjime (Kalman fitering), duomenų suspaudimui ir kodavimui, žaidimuose, kur įtraukiama tikimybė, įvairių duomenų bazių naršymui, robotikos kūrimui ir t.t., o šio tiriamojo mokslinio darbo tikslas – panaudoti Bajeso metodą finansų srityje.

Visgi reiktų apibrėžti kai kuriuos aspektus, kaip, pavyzdžiui, kas yra Bajeso teorema, o kas Bajeso tinklas, Naivus Bajesas ir Bajeso optimalus klasifikatorius, kurie yra tarpusavyje koreliuoti.

Toliau pateikiamas Bajeso teoremos pavyzdys-apibrėžimas: tarkim, jog turime du įvykius – A ir B. Šie įvykiai yra tarpusavyje susiję, ir tarkim, jog B kaip nors įtakoja įvykio A atsiradimą ar buvimą. Norime sužinoti įvykio A tikimybę (P), žinant, jog įvykis B įvyko ar vyksta (t.y., B egzistavimas, buvimas) yra tiesa. Tai žymima taip – P(A|B), taigi tai parodoma 5 formulėje:

$$P(A \mid B) = \frac{P(B, A)}{P(B)} \qquad (5)$$

P(A|B) yra lygi bendrai įvykių B ir A įvykimo tikimybei padalintai iš įvykio B tikimybės.

Ši formulė šiuo atveju yra labai paprasta, ji gali būti kur kas sudėtingesnės formos, pvz., įvykį A gali įtakoti ne vien tik įvykis B, bet ir C, D, E ir t.t. Kitas variantas: įtakojamieji įvykiai yra ne tik A, bet ir daugiau. Kaip paprastą pavyzdį galima apibūdinti taip: jeigu žmogus viršija greitį naktį, tikimybė, kad jis padarys avariją yra palyginti maža, bet jei jis vairuoja naktį ir yra ką tik vartojęs nemažai alkoholio, tai tikimybė, kad jis padarys avariją yra labai didelė, o jei dar pridėtume trečią įvykį, jog jis kurį tai laiką bando važiuoti užsimerkęs, tai avarijos tikimybė bus arti 100%. Analogiškai galima šią teoremą pritaikyti kreditų rizikos vertinime: tikslas nustatyti ar kreditorius yra mokus, tai tarkim, jei jis turi nedaug likvidaus kapitalo, tai tikimybė, kad jis nemokus bus maža, bet jei jo pelno vienai akcijai augimas per paskutinius 3 metus yra labai mažas ar net nuostolingas, tai tikimybė didėja, jei dar pridėsime, kad jo ilgalaikių paskolų ir turto santykis yra labai didelis (tarkim, kad dvigubai ar trigubai) bei grynojo pelno ir akcinio kapitalo santykis bus labai mažas, tai toks kreditorius gali būti skiriamas, prie rizikingų, o tų kintamųjų gali būti dar daugiau. Kita vertus, jei rodikliai atvirkščiai yra geri, jį galima priskirti prie mažos rizikos grupės.

Mašininiame mokyme 5 formulė perrašoma taip:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)} \quad (6)$$

čia *h* yra hipotezė (tarkim, kreditorius skiriamas į rizikingųjų grupę), o *D* šiuo atveju yra apmokymo duomenys (pvz., istoriniai finansiniai skolininko duomenys). P(*h*|*D*) yra vadinama h *posterior* tikimybe, nes skaitoma, jog hipotezės patvirtinimas (arba paneigimas) yra nustatomas tik po to, kai treniravimo duomenys *D* yra jau žinomi, kai tuo tarpu pati P(*h*) tikimybė yra *prior* [9, 29].

Bajeso optimalaus klasifikatoriaus (angl., Bayesian optimal classifier, kitaip dar vadinami BayesNet arba Bajeso tinklais) esmė yra nustatyti kokia labiausiai tikėtina klasė priklauso naujam atvejui, žinant treniravimo duomenis. Naujam atvejui v_j teisingo klasifikavimo tikimybės P(v_j|*D*) išraiška yra pateikta kaip 7 formulė:

$$P(v_j|D) = \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) \quad (7)$$

Bajeso optimalaus klasifikatoriaus išraiška pateikiama žemiau (8 formulė):

$$\underset{v_j \in V}{\operatorname{argmax}} \sum_{h_i \in H} P(v_j|h_i)P(h_i|D) \quad (8)$$

Bajeso optimalus klasifikatorius naują atvejį skiria į vieną ar kitą klasę apjungdamas visas galimas hipotezių prognozes, atsižvelgiant (kaip svorių priskyrimas) į jų posterior tikimybes.

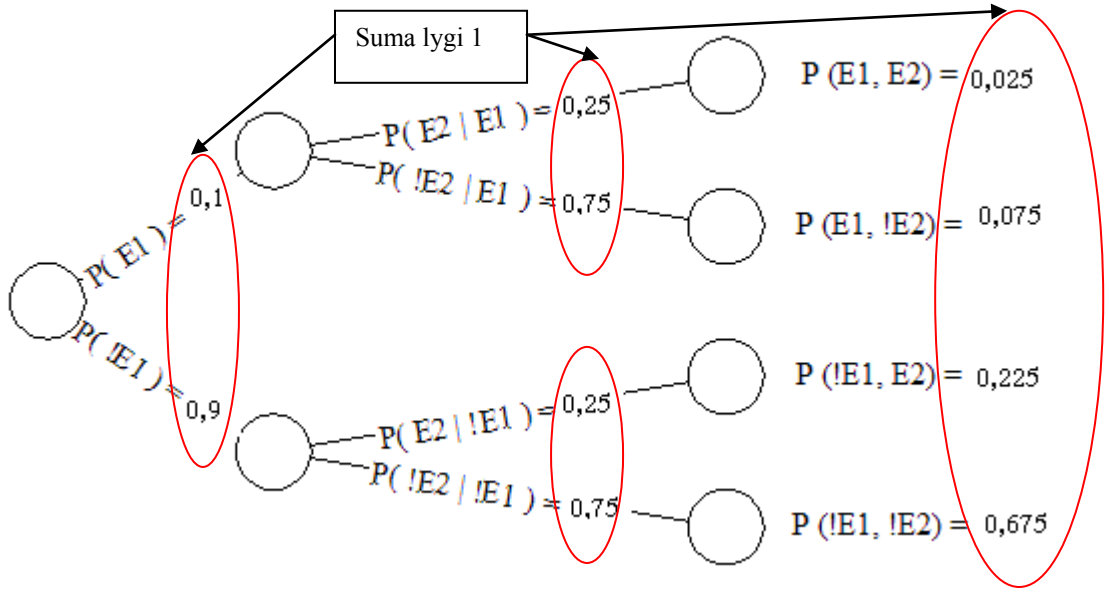
Naivaus Bajeso klasifikatorius (angl., naive Bayes classifier) remiasi atvejų atributais (a_i), pagal kuriuos atvejis yra skiriamas į vieną ar kitą klasę, o patys atributai tarpusavyje laikoma, jog nėra koreliuoti, todėl šis metodas ir vadinamas „naiviu“. Šio klasifikatoriaus formulė (9) pateikiama žemiau:

$$v_{NB} = \underset{v_j \in V}{\operatorname{argmax}} P(v_j) \prod_i P(a_i|v_j) \quad (9)$$

kur v_{NB} yra naivaus Bajeso klasifikatoriaus siekiama nustatyti reikšmė (angl., target value). Formulė aiškinama taip: imama didžiausia gauta reikšmė, dauginant atvejo tikimybę su visų atributų sąlyginių tikimybių su atveju sandauga. Pastarosios reikšmės P(v_j) ir P(a_i| v_j) gaunamos iš jų dažnumo treniravimo duomenyse [29].

Skirtingai nuo naivaus Bajeso klasifikatoriaus, bajeso tinklai atsižvelgia į atributų tarpusavio priklausomybę ir tam kuria kryptinius grafus su mazgų įtakojimu kitų buvimui [9, 30]. Taigi naivaus Bajeso metodas yra paprastesnė Bajeso tinklų atmaina. Pagal pirmąjį pavyzdį, apibrėžiantį 5 formulę, galima sudaryti mažą Bajeso tinklą, kuris atvaizduotų pateiktą situaciją, tai tarkim, kad

yra du įvykiai E1 ir E2, atitinkamai {smarkiai viršija greitį blaivus} ir {vairuoja dieną}. Ir tarkime, kad vairuotojas jau padarė avariją, tada tikimybė, jog padaryta avarija todėl, kad smarkiai viršijo greitį blaivus yra 1 iš 10, t.y, 10%, analogiškai, jog neblaivus bus 90 % ir šį įvykį žymėsime priešingu įvykiui E1, t.y., !E1. Toliau, jei tikimybė, kad jis važiavo naktį, žinant, kad padarė avariją, viršydamas greitį neblaivus, yra 75% ir žymima P(!E2|E1), tai Bajeso tinklas atrodys kaip pavaizduota 1 paveiksle.



Šaltinis: sukurta autoriaus, remiantis [10]

14 pav. Bajeso tinklas

Iš paveikslėlio matome, kad didžiausia tikimybė, jog avarija padaryta yra dėl to, kad vairuotojas viršijo greitį išgėręs naktį, o mažiausia, jog viršijo greitį blaivus dieną, atitinkamai 67,5% ir 2,5%. Žinoma, šis pavyzdys yra labai elementarus, tačiau priežastinių, įtakojančių įvykių gali būti žymiai daugiau ir vieni gali įtakoti arba visus likusius įvykius ar tik dalį jų, taip susidaro didelis ir painus tinklas su daug priežastinių ryšių.

Bajeso metodas taip pat kaip ir kiti minėti metodai yra taikyti kreditų rizikai vertinti, keletas iš jų pateikiama toliau: autoriai A. C. Antonakis ir M. E. Sfakianakis (2009) lygino naivaus Bajeso metodą su kitais penkiais metodais, naudojant realių bankų duomenis, parodyta, kad šis metodas nerodo geriausių rezultatų lyginant su kitais ir pateikiamos priežastys dėl ko taip yra. Kiti autoriai Sanjiv R . Das , Rong Fan ir Gary Geng savo darbe (2002) Bajeso metodą taikė PD dinamikos atžvilgiu. Gunter Loffler, Peter N. Posch, Christiane Schone (2005) parodė, kad Bajeso metodas gali būti sėkmingai taikomas bankinėse sistemose, kur dažniausiai naudojamos tradicinės rizikos vertinimo metodologijos, kaip diskriminantinė ar logit analizės; autoriai parodė, jog naudojant Bajeso metodą logit analizės vertinimo tikslumas gali būti pagerintas kartu naudojant ir Bajeso

metodą nuo 2 iki 9 %, įskaitant atvejus, kai duomenų rinkiniuose trūksta tam tikrų duomenų. Bajeso tinklų taikymas kreditų įvertinimui taip pat buvo aprašyti autorių B. Baesens, M. Egmont-Petersen, R. Castelo, J. Vanthienen (2002), kurie taip šį metodą apjungė su Markov Chain Monte Carlo paieška geresniems rezultatams gauti. 1993 metais Terence C. Fogarty ir Neil S. Ireson Bajeso metodą kreditų rizikai vertinti taikė apjungiant jį su genetiniais algoritmais. Egzistuoja ir hibridinių Bajeso metodo taikymų, pvz., Gerhard Paass ir Jörg Kindermann (1998) kreditų vertinimui apjungė du metodus: Bajeso ir klasifikavimo medžių. Taigi Bajeso metodas tinkamas ir sėkmingai pritaikytas analizuojamai problemai spręsti, t.y., kreditų rizikai vertinti.

Toliau glaustai apibrėžiami žinomi klasifikavimui tinkantys metodai, nesigilinant į jų specifines savybes, pritaikymą realiems uždaviniams ir pan., aptariant tik jų veikimo principus.

1.15. KNN

Šio metodo esmė yra atskirti duomenis į tam tikras klases pagal tų klasių bruožų skirtumus, ieškant jų pateiktų duomenų masyvuose. Kiekvienas duomenų vienetas yra skirstomas erdvėje, ir matuojami atstumai tarp jų kiekvieną porą žymint kaip (x,y). Tada skaičiuojamas euklidinis atstumas tarp dviejų porų pagal formulę [30]:

$$\mathbf{d} = \sqrt{\sum_{i=0}^{i=n} (\mathbf{x}_i - \mathbf{y}_i)^2} \qquad (10)$$

Naudojant šiuos atstumus sudaroma atstumų matrica tarp visų įmanomų duomenų porų, kiekvienas duomenų atvejis (vienetas, taškas) turi klasės atributą $C=\{c_1,...,c_n\}$. Kiekvienam duomenų atvejui k yra kaimyninių atvejų skaičius, kuriuos randama pagal sukurtąją atstumų matricą. Tada analizuojant kaimyninius duomenų vienetus nustatoma kokioms klasėms jie priklauso ir tam analizuojamam atvejui tada taip pat skiriama ta klasė, kurią daugiausia turi kaimyniniai duomenų atvejai. Jei pasitaiko toks atvejis, jog yra dvi ar daugiau klasių, turinčių tą patį didžiausią skaičių artimų kaimynų, tada atmetamas vienas iš kaimynų ir tikrinama priklausomybė likusiems duomenų vienetams (k-1), reikiant ir (k-2), ir t.t., analizuojamam duomenų vienetui klasė priskiriama tik tada, kai tik viena gretimų vienetų klasė dominuoja [31, 32].

1.16. Medžių algoritmai

Sprendimų medžiai arba kitaip klasifikavimo medžiai remiasi sprendimų išdėstymu medžiais, kur egzistuoja mazgai „lapai“, kurie daugiau neturi jokių atsišakojimų ir yra klasės reikšmė arba dar yra „sprendinių“ mazgai, kurie nusako tam tikrą „testą“ atributui-reikšmei nustatyti. Visi klasifikavimo medžio mazgai, išskyrus aukščiausiąjį, kuris neturi tėvinio mazgo, turi tik vieną tėvinį mazgą. Sprendimui priimti naudojant medžių metodą, einama nuo medžio šaknies į

žemesnius mazgus, kurie tenkina tam tikrą sąlygą (išlaiko testą), tada jau nuo tenkinančio sąlygą mazgo einama į dar žemesnius mazgus, kur vėl žiūrima, kuris tenkina norimą sąlygą ir taip iki paskutinio mazgo – lapo, kuris nusako į kokią klasę patenka testuojamas atributas. Kaip ir daugeliui kitų metodų, šiam taip pat atributai ir klasės turi būti diskrečių reikšmių (tolydžius išėjimus duoda regresijos metodai), o apmokymo duomenų turi būti kuo daugiau [33].

1.17. Evoliuciniai skaičiavimai

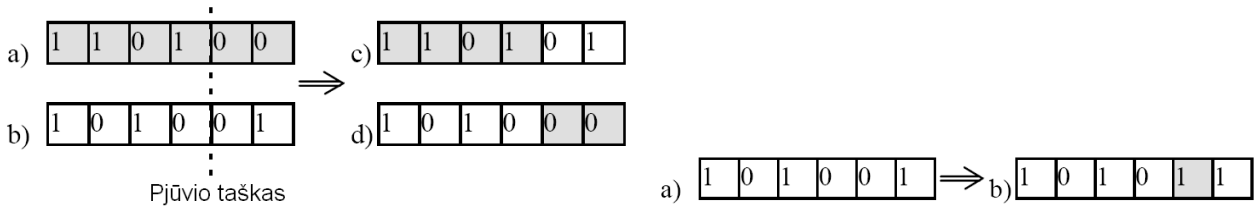
Evoliucinius skaičiavimus apima 4 populiariausi algoritmai, kurie dažniausiai naudojami realioms problemoms spręsti: genetiniai algoritmai, genetinis programavimas, evoliucinės strategijos, evoliucinis programavimas. Pastarieji algoritmai turi tą pačią metodologinę bazę, skiriasi tik tam tikri komponentai jų veikimo schemose, reprodukcijos bei mutacijos operatoriuose, bei populiacijos individų atrankos mechanizmuose [4]. Bendros jų savybės pateikiamos žemiau:

1) Suformuojama pradinė individų populiacija;
2) Įvertinama kiekvieno populiacijos nario kokybės funkcija;
3) Geriausios kokybės individai atrenkami sekančiai generacijai;
4) Atliekamas kryžminimas tarp atrinktų populiacijos narių ir suformuojama nauja individų generacija;
5) Atliekama naujos generacijos individų mutacija;
6) Apskaičiuojama naujos generacijos individų kokybės funkcijos. Jei geriausio populiacijos nario kokybės funkcija atitinka užsiduotą reikšmę arba viršijamas skaičiavimų limitas, skaičiavimai nutraukiami. Priešingu atveju grįžtama į trečią procedūros žingsnį.

Žingsniai 4 ir 5 gali būti naudojami kartu arba tik vienas iš jų.

Genetiniai algoritmai iš kitų evoliucinių skaičiavimų išsiskiria trimis požymiais:

- Individų savybės genetiniuose algoritmuose atvaizduojamos bitų eilutėmis (angl., bitstrings);
- Individų atrinkimui į sekančią generaciją taikomas proporcinės selekcijos metodas;
- Pagrindinis individų variacijos kūrimo instrumentas yra individų kryžminimas, tačiau naudojamas ir mutacijos būdas (15 pav.).

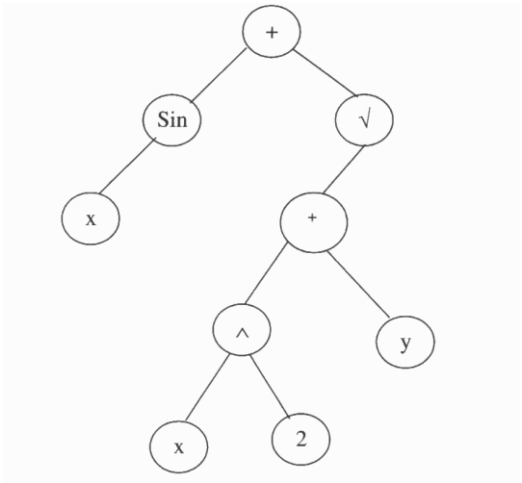


Šaltinis: [4]

15 pav. Kryžminimo (kairė) ir mutacijos operatorių, taikomų genetiniuose algoritmuose, iliustracija

Reikia paminėti, kad toks operatorius kaip kryžminimas, neįneša jokios naujos informacijos, kai tuo tarpu mutacijos operatorius įneša į populiaciją naują informaciją.

Genetinis programavimas yra speciali genetinių algoritmų rūšis, kurių tikslas yra rasti ne optimalias tam tikrai problemai spręsti bitų eilutes, bet rasti optimalų programos kodą ar optimalią analitinę funkciją, kuri leistų efektyviai išspręsti analizuojamą problemą. Kaip pavyzdys pateikiamas 16 pav., kur vaizduojamas funkcijos $f(x) = \sin(x) + \sqrt{(x^2 + y)}$ medis.



Šaltinis: [4]

16 pav. Genetinio programavimo pavyzdys

Kiekvienas medžio (16 pav.) lapas yra tam tikra reikšmė, pasirenkama iš galimų reikšmių rinkinio. Kiekvienas vidinis medžio mazgas yra tam tikra funkcija, kuri atrenkama iš galimų funkcijų sąrašo. Visas medis atitinka vieną sudėtingą funkciją, kuri savyje gali apimti eilę elementarių funkcijų ir jų reikšmių. Šios funkcijos reikšmę galima apskaičiuoti atitinkamai apskaičiuojant viso medžio mazgų ir lapų reikšmes. Paprastai medis pradedamas skaičiuoti nuo kairiosios žemiausios šakos. Genetiniame programavime dažniausiai naudojamas medžio šakų kryžminimo operatorius.

Evoliucinio programavimo ir **evoliucinių strategijų** metodai, skirtingai nuo genetinių algoritmų, objekto savybių aprašymui vietoje fiksuoto ilgio bitų eilučių naudoja fiksuoto ilgio realių skaičių vektorius. Pagrindinis pastarųjų dvejų metodų skirtumas yra tame, kad evoliucinio programavimo metoduose dažniausiai pagrindinis populiacijos keitimo operatorius yra mutacija, o tuo tarpu evoliucinių strategijų metodai populiacijos keitimui taiko tiek individų kryžminimo, tiek mutacijos operatorius. Šiuo atveju kryžminimas yra dvejopas: paprastas (kaip ir genetiniuose algoritmuose) ir aritmetinis, išvedant aritmetinį vidurkį tam tikroms reikšmėms.

Taigi evoliuciniai skaičiavimai naudojami daugelyje sričių, įskaitant ir klasifikavimą, o taip pat paieškos optimizavimui, naudojant geriausiųjų galimų variantų vystymą tolimesnei raidai.

1.18. Neraiškios aibės ir ekspertinės sistemos

Įprastose aibėse objektas priklauso aibei arba ne. Neraiškios aibės turi priklausomybės aibei funkciją (priklausomumo funkciją). Tarkime, norima skolininkus skirstyti į rizikingus ir ne, knygas į storas ir plonas…ir t.t. Neraiškios aibės padeda operuoti sąvokomis, tokiomis kaip ‘nelabai didelis’, ‘labai aukštas’.

Neraiškios aibės tinkamos atvaizduoti nepatikimai informacijai ar nepatikimumo lygiui. Neraiškios aibės aprašomos poromis:

$$A = \{x, \mu_A(x)\}, \qquad (11)$$

kur x yra dalis visos įmanomos X reikšmių srities, $x \in X$ ir $0 \le \mu_A \le 1$. Priklausomumo funkcija $\mu_A(x)$ nusako x priklausomybės aibei A laipsnį. $\mu_A(x)=0$ reiškia, jog x nepriklauso aibei A, ir $\mu_A(x)=1$ reiškia pilną priklausomybę. Operacijos su neraiškiomis aibėmis atliekamos panašios kaip ir įprastoms aibėms (sankirta, sąjunga, priklausomybė) [4].

Pati *fuzzy* logika viena retai kur taikoma. Dažniausiai ji taikoma kartu su neuroniniais tinklais, genetiniais algoritmais, nes apjungus keletą metodų, galima gauti geresnius rezultatus įnešant į juos ekspertų ar specialistų žinias.

Klasikinės ekspertinės sistemos neturi apsimokymo savybių, tačiau jos sugeba paaiškinti savo daromus sprendimus, o kiti metodai, kaip dirbtiniai neuroniniai tinklai gali apsimokyti iš pavyzdžių, tačiau jie funkcionuoja „juodos dėžės“ principu ir negali paaiškinti savo pateikiamų sprendimų. Hibridinė sistema, vadinama neuro-ekspertinė sistema, apjungia abiejų technikų gerąsias savybes [4]. Šioje sistemoje žinių bazė yra formuojama neuroninio tinklo pagalba. Taisyklių formavimo blokas analizuoja neuroninio tinklo įėjimus ir išėjimus ir pagal juos suformuoja „*JEI ... TAI*“ taisykles. *Fuzzy* logika yra bandoma taikyti kartu su įvairiais metodais, įskaitant Bajeso, medžių, KNN ar SVM.

1.19. Klasifikavimo metodų tarpusavyje palyginimas

Šio skyriaus eigoje pateikiama eilė tyrimų, atliktų lyginant įvairius mašininio mokymo algoritmus (metodus) ir įvertinant įvairius jų aspektus, kriterijus ir pan., juos aprašant ir įvardijant gautus rezultatus.

1997 m. buvo atliktas vienas iš platesnių klasifikavimo algoritmų testų, kur pagrindinis dėmesys buvo skirtas algoritmo klasifikavimo tikslumui, sudėtingumui ir treniravimosi laikui, o iš viso buvo ištestuoti 33 algoritmai, nors kai kurie iš jų buvo tiesiog vieno metodo atmainos, pavyzdžiui, buvo testuojami keturi Bajeso metodu paremti algoritmai: bayes, bayes opt, mml ir mml opt (atitinkamai sutrumpinimai IB, IBO, IM ir IMO). Iš esmės tai du Bajeso metodai (be „opt“), o

kiti du yra jų išplėtimai, kurie yra šiek tiek sudėtingesni, dirba ilgiau ir dėl papildomų sukuriamų tinklų (medžių) skaičiaus užima daugiau vietos, tačiau jų klasifikavimo tikslumas yra didesnis. Visi 33 algoritmai suskirstyti pagal savo savybes į 3 grupes: 22 medžiais ir taisyklėmis paremti algoritmai, 9 statistiniai algoritmai ir 2 neuroniniais tinklais paremti algoritmai.

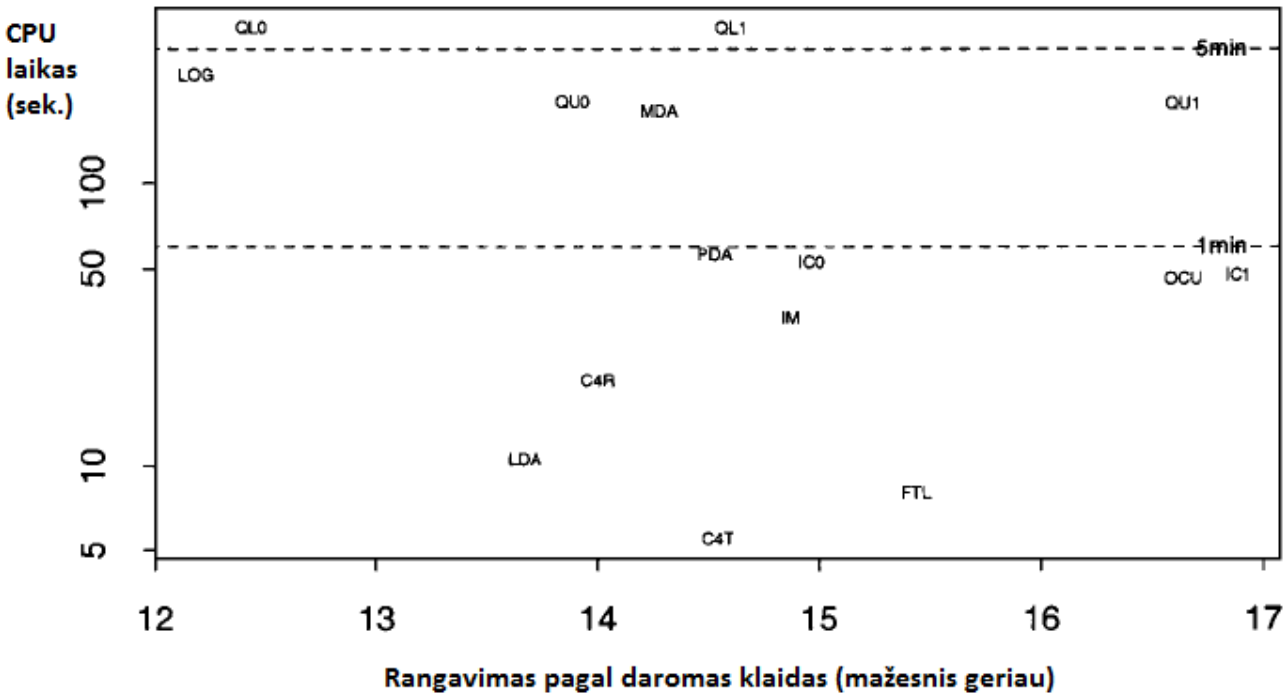
Algoritmams testuoti buvo pasirinkta 16 testinių originalių duomenų grupių (problemų, kaip krūtinės vėžio atpažinimas, paveikslėlių segmentacija, diabeto atpažinimas, DNR sekos ir t.t.) ir 16 aptriukšmintų tų pačių duomenų grupių, kurių egzemplioriai turi nuo 5 iki 60 atributų, o grupėje esančių egzempliorių skaičiai svyruoja nuo 151 iki 4435. Kadangi ne visi algoritmai pritaikyti įprastiems kategoriniams atributams, jie buvo paversti į vektorinius atributus, t.y., į 0 ir 1. Testai buvo atlikti trimis skirtingais kompiuteriais Unix platformoje.

POL	LOG	MDA	QLO	LDA	QL1	PDA	IC0	FM2	IBO	IMO
.195	.204	.207	.207	.208	.211	.213	.215	.218	.219	.219
C4R	IM	LMT	C4T	QUO	QU1	OCU	IC1	IB	OCM	ST0
.220	.220	.220	.220	.221	.226	.227	.227	.229	.230	.232
ST1	FTL	FTU	FM1	RBF	OCL	LVQ	CAL	NN	QDA	T1
.233	.234	.238	.242	.257	.260	.269	.270	.281	.301	.354

Šaltinis: [23]

17 pav. 33 klasifikavimo algoritmų atliktų klaidų vidurkiai

Atlikus visus testus rezultatai (17 pav.) parodė, kad daugelio algoritmų statistinis teisingo klasifikavimo efektyvumas yra gana panašus, tačiau geriausiai pasirodė POL (POLYCLASS) algoritmas, kuris yra iš statistinės grupės ir remiasi LOG (angl., logistic discriminant analysis) algoritmu bei pateikia tikimybinius rezultatus, tačiau šis metodas yra vienas iš pačių lėčiausių (ilgiau nei 3 valandos apsitreniravimui).



Šaltinis: [23]

18 pav. 15 geriausių klasifikavimo algoritmų vertinant tikslumą ir greitį

Pagal autorių išvadas, nors POL algoritmas klasifikuoja tiksliausiai, visgi, jei reikalingas greitesnis algoritmas didelėms sistemoms spręsti didesnius uždavinius, geriau rinktis kitus, kaip LOG ar QL0 (vienas iš klasifikavimo medžių QUEST metodų, paremtas tiesiniu diskriminantiniu metodu LDA, angl., linear discriminant analysis), o iš viso tokių algoritmų yra 15, kurie panašius rezultatus pasiekia per mažiau nei 10 minučių (18 pav.), tarp kurių yra ir vienas iš Bajeso metodu paremtas algoritmas.

Vienas iš paskutinių platesnių besimokančių algoritmų palyginimas atliktas autorių Rich Caruana, Alexandru Niculescu-Mizil, kurie savo eksperimente tarpusavyje lygino 10 metodų: SVM, neuroniniai tinklai, mechaninė regresija (logistic regression), naivus Bajeso metodas (naive bayes, t.y., nemodifikuotas, neapjungtas su kitais metodais), atmintimi grindžiamas mokymas (memory-based learning, KNN), atsitiktinio išsišakojimo (random forests), sprendimo medžių (decision trees), plėstų medžių (bagged trees), paskatintų medžių (boosted trees), paskatinto kapojimo (boosted stums). Taip pat šie visi metodai buvo apjungti su Platt Scaling ir Isotonic Regression modeliais geresnių rezultatų gavimui. Šie du modeliai iš esmės skirti metodams neteikiantiems tikimybinių skirstymo į tam tikras klases reikšmių. Pirmasis iš jų paremtas transformuojamą metodą perkelti į sigmoidę, o antrasis pasižymi apribojimais, kad žymėjimas „žemėlapyje“ yra monotoniškai augantis.

Metodų testavimui buvo pasirinkti įvairūs jau realizuoti klasifikatoriai:

- SVM – SVMLight (Joachims, 1999) su įvairiais kriterijais ir tipais (radialinių bazinių funkcijų, tiesinis, polinominis);
- ANN – neuroniniai tinklai su gradientiniu nusileidimu atgal, skirtingais paslėptais neuronų sluoksniais, skirtingais „momentum“, skirtingu treniravimo iteracijų (angl., epochs) skaičiumi;
- Logistic Regression (LOGREG) – naudoti ir „regulized“ (su skirtingais apribojimais), ir „unregulized“ modeliai;
- Naive Bayes (NB) – naudotas Weka (Witten & Frank, 2005) klasifikatorius išbandant visus tris siūlomus variantus: su normalizavimu, kernel nustatymais ir diskretizacija;
- KNN – naudoja keletą svorių parinkimo variantų ir skirtingas kernel reikšmes;
- Random Forests (RF) – bandyti du klasifikatoriai: Breiman- Cutler ir Weka, bet geriau pasirodė pirmasis su 1024 medžiais, tad jis ir naudotas, požymių reikšmės taip pat skirtingos;
- Decision trees (DT) – naudotas Buntine's IND paketas (Buntine & Caruana, 1991) su visais siūlomais modeliais BAYES, ID3, CART, CART0, C4, MML, SMML bei tipais C44LS, C44BS ir MMLLS;
- Bagged trees (BAG-DT) – tas pats kaip ir DT , tik skirtingas medžių skaičius bei parenkami kriterijai;
- Boosted trees (BST-DT) – tas pats kaip ir DT , tik skirtingas medžių skaičius bei parenkami kriterijai;
- Boosted stumps (BST-STMP) – tas pats kaip ir DT , tik skirtingas medžių skaičius bei parenkami kriterijai;

Paveiksle 19 pateikti testavimo duomenys, kurių yra 11 grupių (sprendimo problemų), kiekvienas iš jų yra paimtas iš jau anksčiau kitų autorių vykdytų eksperimentų kaip, pavyzdžiui, raidžių, medicininių ar nominalių duomenų atskyrimas. Paveiksle pateikta 11 klasifikavimo problemų, kur stulpeliai paeiliui reiškia: pavadinimas, treniravimo ir testavimo duomenų kiekiai.

PROBLEM	TRAIN SIZE	TEST SIZE
ADULT	5000	35222
BACT	5000	34262
COD	5000	14000
CALHOUS	5000	14640
COV_TYPE	5000	25000
HS	5000	4366
LETTER.P1	5000	14000
LETTER.P2	5000	14000
MEDIS	5000	8199
MG	5000	12807
SLAC	5000	25000

Šaltinis: [24]

19 pav. Testavimo duomenys

Testuotojai metodų efektyvumą vertino aštuoniais kriterijais suskirstytais į tris grupes:

- Slenkstis (angl., threshold), kurį sudaro trys kriterijai – tikslumas (ACC), F-score (FSC) bei lift (LFT) kreivės. Jei metodo reikšmės yra aukščiau slenksčio, jis turės teigiamą vertinimą, jei žemiau – neigiamą, čia neatsižvelgiama kaip arti slenksčio yra reikšmės.
- Tvarkos (angl., ordering/rank), kur žiūrima kaip tiksliai išsidėsčiusios reikšmės tarpusavyje neatsižvelgiant į pačių reikšmių skaitinį tikslumą, kaip gerai atskiriamos teigiamos nuo neigiamų reikšmių, vienos klasės nuo kitos. Šią grupę sudaro ROC kreivė, vidutinis tikslumas (APR) bei BEP (angl., precision/recall break even pint), kurių kreivių plote turi būti metodų reikšmės.
- Tikimybių, kuri nusako tikimybę, jog reikšmės patenka į vieną ar kitą klasę. Šią grupę sudaro kvadratinė klaida (RMS) ir susikertanti entropija (angl., cross-entropy, MXE).

Pagal pateiktus vertinimo kriterijus pateikti testavimo rezultatai (Pav. 20), kurie yra nominalių reikšmių (nuo -1 iki 1, kur neigiamas, jei netenkino slenksčio). Eilutėse yra metodų pavadinimai, o stulpeliuose paeiliui: modelio apjungimas su Platt ar Isotonic modeliais (atitinkamai PLT ir ISO, o brūkšnelis reiškia neapjuntą metodą), toliau rikiuojasi 8 aukščiau išvardyti kriterijai, priešpaskutinis stulpelis rodo vidurkį, o paskutinis stulpelis rodo vidurkį su optimaliausiais metodo nustatymais, kuriuos jau keičia po pirmojo testo.

MODEL	CAL	ACC	FSC	LFT	ROC	APR	BEP	RMS	MXE	MEAN	OPT-SEL
BST-DT	PLT	.843*	.779	**.939**	**.963**	**.938**	.929*	**.880**	**.896**	**.896**	**.917**
RF	PLT	.872*	.805	.934*	.957	.931	**.930**	.851	.858	.892	.898
BAG-DT	–	.846	.781	.938*	.962*	.937*	.918	.845	.872	.887*	.899
BST-DT	ISO	.826*	.860*	.929*	.952	.921	.925*	.854	.815	.885	.917*
RF	–	**.872**	.790	.934*	.957	.931	**.930**	.829	.830	.884	.890
BAG-DT	PLT	.841	.774	.938*	.962*	.937*	.918	.836	.852	.882	.895
RF	ISO	.861*	**.861**	.923	.946	.910	.925	.836	.776	.880	.895
BAG-DT	ISO	.826	.843*	.933*	.954	.921	.915	.832	.791	.877	.894
SVM	PLT	.824	.760	.895	.938	.898	.913	.831	.836	.862	.880
ANN	–	.803	.762	.910	.936	.892	.899	.811	.821	.854	.885
SVM	ISO	.813	.836*	.892	.925	.882	.911	.814	.744	.852	.882
ANN	PLT	.815	.748	.910	.936	.892	.899	.783	.785	.846	.875
ANN	ISO	.803	.836	.908	.924	.876	.891	.777	.718	.842	.884
BST-DT	–	.834*	.816	**.939**	**.963**	**.938**	.929*	.598	.605	.828	.851
KNN	PLT	.757	.707	.889	.918	.872	.872	.742	.764	.815	.837
KNN	–	.756	.728	.889	.918	.872	.872	.729	.718	.810	.830
KNN	ISO	.755	.758	.882	.907	.854	.869	.738	.706	.809	.844
BST-STMP	PLT	.724	.651	.876	.908	.853	.845	.716	.754	.791	.808
SVM	–	.817	.804	.895	.938	.899	.913	.514	.467	.781	.810
BST-STMP	ISO	.709	.744	.873	.899	.835	.840	.695	.646	.780	.810
BST-STMP	–	.741	.684	.876	.908	.853	.845	.394	.382	.710	.726
DT	ISO	.648	.654	.818	.838	.756	.778	.590	.589	.709	.774
DT	–	.647	.639	.824	.843	.762	.777	.562	.607	.708	.763
DT	PLT	.651	.618	.824	.843	.762	.777	.575	.594	.706	.761
LR	–	.636	.545	.823	.852	.743	.734	.620	.645	.700	.710
LR	ISO	.627	.567	.818	.847	.735	.742	.608	.589	.692	.703
LR	PLT	.630	.500	.823	.852	.743	.734	.593	.604	.685	.695
NB	ISO	.579	.468	.779	.820	.727	.733	.572	.555	.654	.661
NB	PLT	.576	.448	.780	.824	.738	.735	.537	.559	.650	.654
NB	–	.496	.562	.781	.825	.738	.735	.347	-.633	.481	.489

Šaltinis: [24]

20 pav. Eksperimento rezultatai pagal vertinimo kriterijus

Pagal problemas rezultatai atrodo taip, kaip pateikta 21 paveiksle žemiau, kur laukų reikšmės tokios pačios kaip 19 paveiksle, tik vietoje vertinimo kriterijų yra sprendžiamų problemų sutrumpinimai (pav. 21)

MODEL	CAL	COVT	ADULT	LTR.P1	LTR.P2	MEDIS	SLAC	HS	MG	CALHOUS	COD	BACT	MEAN
BST-DT	PLT	**.938**	.857	**.959**	**.976**	.700	.869	**.933**	.855	**.974**	**.915**	.878*	**.896***
RF	PLT	.876	.930	.897	.941	**.810**	.907*	.884	.883	.937	.903*	.847	.892
BAG-DT	–	.878	.944*	.883	.911	.762	.898*	.856	**.898**	.948	.856	**.926**	.887*
BST-DT	ISO	.922*	.865	.901*	.969	.692*	.878	.927	.845	.965	.912*	.861	.885*
RF	–	.876	.946*	.883	.922	.785	.912*	.871	.891*	.941	.874	.824	.884
BAG-DT	PLT	.873	.931	.877	.920	.752	.885	.863	.884	.944	.865	.912*	.882
RF	ISO	.865	.934	.851	.935	.767*	**.920**	.877	.876	.933	.897*	.821	.880
BAG-DT	ISO	.867	.933	.840	.915	.749	.897	.856	.884	.940	.859	.907*	.877
SVM	PLT	.765	.886	.936	.962	.733	.866	.913*	.816	.897	.900*	.807	.862
ANN	–	.764	.884	.913	.901	.791*	.881	.932*	.859	.923	.667	.882	.854
SVM	ISO	.758	.882	.899	.954	.693*	.878	.907	.827	.897	.900*	.778	.852
ANN	PLT	.766	.872	.898	.894	.775	.871	.929*	.846	.919	.665	.871	.846
ANN	ISO	.767	.882	.821	.891	.785*	.895	.926*	.841	.915	.672	.862	.842
BST-DT	–	.874	.842	.875	.913	.523	.807	.860	.785	.933	.835	.858	.828
KNN	PLT	.819	.785	.920	.937	.626	.777	.803	.844	.827	.774	.855	.815
KNN	–	.807	.780	.912	.936	.598	.800	.801	.853	.827	.748	.852	.810
KNN	ISO	.814	.784	.879	.935	.633	.791	.794	.832	.824	.777	.833	.809
BST-STMP	PLT	.644	**.949**	.767	.688	.723	.806	.800	.862	.923	.622	.915*	.791
SVM	–	.696	.819	.731	.860	.600	.859	.788	.776	.833	.864	.763	.781
BST-STMP	ISO	.639	.941	.700	.681	.711	.807	.793	.862	.912	.632	.902*	.780
BST-STMP	–	.605	.865	.540	.615	.624	.779	.683	.799	.817	.581	.906*	.710
DT	ISO	.671	.869	.729	.760	.424	.777	.622	.815	.832	.415	.884	.709
DT	–	.652	.872	.723	.763	.449	.769	.609	.829	.831	.389	.899*	.708
DT	PLT	.661	.863	.734	.756	.416	.779	.607	.822	.826	.407	.890*	.706
LR	–	.625	.886	.195	.448	.777*	.852	.675	.849	.838	.647	.905*	.700
LR	ISO	.616	.881	.229	.440	.763*	.834	.659	.827	.833	.636	.889*	.692
LR	PLT	.610	.870	.185	.446	.738	.835	.667	.823	.832	.633	.895	.685
NB	ISO	.574	.904	.674	.557	.709	.724	.205	.687	.758	.633	.770	.654
NB	PLT	.572	.892	.648	.561	.694	.732	.213	.690	.755	.632	.756	.650
NB	–	.552	.843	.534	.556	.011	.714	-.654	.655	.759	.636	.688	.481

Šaltinis: [24]

21 pav. Eksperimento rezultatai pagal sprendimo problemas

Apibendrinant atliktą eksperimentą, galima teigti jog klasifikavimo medžių principais veikiantys metodai „boosting“, „random forests“, „bagging“ ir SVM metodas visuose problemų sprendimuose atrodo efektyviausiai. Apjungus metodus su Platt ir Isotonic regresijos modeliais pagerėja daugelio metodų rezultatai, o didžiausią įtaką daro „boosted trees“, SVM, „boosted stumps“ ir „Naive Bayes“ metodams. Autoriai pranašiausiu metodu pripažino „boosted trees“, o antras seka „random forests“. Pasak [24], kadangi vieni metodai geriau sprendžia vienas klasifikavimo problemas, kiti kitas, analogiškuose testuose, naudojant skirtingas problemas galima gauti skirtingus rezultatus, nors jie neturėtų labai skirtis.

Darbe [25] buvo atliktas eksperimentas kaip teigiama trijų populiariausių klasifikavimo metodų, t.y., SVM (Support Vector Machines), Bajeso ir kNN (k-Nearest Neighbors), paveikslėlių turiniui atstatyti, o šie metodai pasirinkti būtent paveikslėlių turinio atstatymo kontekste (angl., Content-Based Image Retrieval, CBIR). kNN metodas pasirinktas todėl, kad ankstesniuose CBIR tesuose, kur buvo klasifikuoti paveikslėlių pikseliai, jis jau parodė geresnius rezultatus nei neuroniniai tinklai ir tiesinės diskriminantinės analizės bei kiti metodai.

Tikslas buvo išbandyti naują autorių strategiją RETIN AL (RETIN Active Learning), kuri tinkama įvairiems metodams ir šiame eksperimente pateisino autorių lūkesčius pasirodydama geriau nei įprasti klasifikavimo metodai ir viena išskirta SVM strategija – Tong. Eksperimentas atliktas naudojant COREL duomenų bazę su 50000 paveikslėlių, iš kurių pasirinkta 6000 eksperimentui. Kaip teigia autoriai, eksperimentui buvo pasirinkti sunkiausiai klasifikuojami paveikslėliai su mažai treniruojamų duomenų. Šie paveikslėliai buvo suskirstyti į vienuolika klasių, kurių pasiskirstymas atrodo taip:

Klasė	**Dydis**	**Pastaba**
birds	219	Įvairūs viso pasaulio paukščiai
castles	191	Modernios ir viduramžių pilys
caverns	121	Urvai ir uolos iš vidaus
dogs	111	Įvairių veislių šunys
doors	199	Paryžiaus ir San Francisko durys
Europe	627	Europos miestai ir gamtovaizdžiai
flowers	506	Įvairios viso pasaulio gėlės
food	315	Indai ir vaisiai
mountains	265	Kalnai
objects	116	Įvairūs pavieniai objektai nesikeičiančiame fone
savannah	408	Afrikos dykumų gyvūnai

Šaltinis: sukurta autoriaus, remiantis [25]

Pagal gautus eksperimento rezultatus, geriausiai pasirodė SVM metodas, o Bajeso ir kNN metodai klasifikavo prasčiau ir gana panašiai (šiek tiek geriau pasirodė kNN) su įprastomis ir RETIN AL strategijomis.

2006 m. šaltinyje [26] tirtas penkių mašininio mokymo algoritmų gebėjimas sparčiai klasifikuoti IP transportavimo srautus (pagal paketų dydžius, atvykimo laiką ir pan.). Tirti šie metodai: Naive Bayes (naivaus Bajeso, kuris tiria priklausomybę tarp atributų ir klasės bei sieja su kiekvienu atveju, o taip pat teikia tikimybes priklausomumui klasėms), C4.5 (medžių metodo algoritmas), Bajeso tinklai (tinklai iš kryptinių neciklinių grafų, sudarytų iš mazgų, reiškiančių funkcijas arba klases, ir ryšių bei iš sąlyginių tikimybių sudarytos lentelės, nurodančios ryšių stiprumą), Naive Bayes Tree (medžių algoritmas apjungtas su naivaus Bajeso algoritmu) ir kreiptas dėmesys ne į tikslumą, o į efektyvumą (skaičiavimo greitį, angl., performance), nes, pasak autorių, algoritmų tikslumas yra daugmaž panašus.

Atlikus testus pasitvirtino teiginiai, kad visų algoritmų tikslumas yra labai panašus, tačiau greičio atžvilgiu jie rodo skirtingus rezultatus. Vienos iš išvadų yra, jog įvairių funkcijų nenaudojimas menkai paveikia klasifikavimo tikslumą, tačiau žymiai paspartina skaičiavimo greitį. Greičiausias metodas klasifikuojant buvo C4.5, po jo seka Naive Bayes, Bajeso tinklų ir Naive Bayes Tree algoritmai. Naive Bayes Tree tiksliausiai klasifikavo, tačiau vienas šio metodo trūkumas, jog jis ilgiausiai užtrunka kol sukuria savo struktūrą (apsimoko), tuo tarpu šiuo požiūriu greičiausi paeiliui Naive Bayes, Bajeso tinklai ir C4.5.

Vienas iš palyginimų yra darbe [27], kur buvo pateikiamas naujas (2003 m.) Bajeso neuroninio tinklo klasifikatorius (angl., Bayesian Neural Network Classifier, BNN) ir lyginami jo rezultatai su iki tol buvusiais analogiškais klasifikatoriais bei su SVM klasifikavimo metodu, kuris, pasak autorių, pasirinktas dėl to, jog įvairiuose eksperimentuose rodė geresnius klasifikavimo rezultatus už kitus analogiškus metodus, kaip artimiausio kaimyno (angl., nearest neighbor), diskriminantinės analizės (angl., discriminant analysis) bei neuroninius tinklus (angl., neural networks), o taip pat yra efektyviai naudojamas realiose aplikacijose.

Nepaisant teigiamų SVM metodo savybių: praktiškumas naudojant realiose aplikacijose, išplėtota teorinė bazė, sąlyginai paprasti skaičiavimai ir geri rezultatai, jis turi keletą minusų – tai jautrumas dideliems triukšmams bei netoleravimas neapibrėžtoms reikšmėms, kas lemia prastesnes generalizavimo savybes, t.y., prastesni rezultatai dirbant su testuojamais duomenimis nei su apmokymo.

Kalbant apie neuroninius tinklus, jie turi geras vidines savybes, bet jas sunkiau realizuoti, todėl neuroniniai tinklai klasifikuojant rodo prastesnę kokybę lyginant su SVM ar BNN. Pagrindinės dvi priežastys kodėl sudėtinga sėkmingai realizuoti neuroninius tinklus yra vis dar neaiški neuroninio tinklo optimali struktūra specifinei užduočiai spręsti bei reikalingų duomenų kiekiai apsimokyti, todėl jie arba per daug pritampa prie mokymosi duomenų arba per mažai ir tai lemia prastas generalizavimo savybes. Kita NN (neuroninių tinklų) problema – treniravimo algoritmų efektyvumas, nes deterministiniai treniravimo algoritmai, kaip skleidimo atgal (angl., backpropagation) ir jungtinio gradiento (angl., conjugate gradient) bei jų įvairios variacijos yra linkusios patekti į lokalius funkcijos minimumus ir ši problema opesnė su sudėtingesnės struktūros NN. Šias problemas padeda spręsti Bajeso neuroniniai tinklai: pradinė tinklo struktūra ir svorių jungtys veikia reguliacija tinklo apmokymams, kontroliuojant įėjimo kintamųjų, paslėpto sluoksnio neuronų skaičiaus ir svorių srities parinkimą; Monte Karlo Markovo grandinės (angl., Markov chain Monte Carlo, MCMC) metodas leidžia išvengti patekimo į lokalius minimumus, kur jungtinės posterior tinklų struktūros ir svorių jungtys yra atrenkami, o trečia – galimas duomenų neapibrėžtumas, kur praleistiems duomenims naudojamas vidurkinimas.

Pastarajame [27] darbe buvo įvykdyti du eksperimentai, kurių viename buvo testuojamas tik BNN, o kitame visi trys aukščiau minėti metodai. Kiekviename eksperimente daryta po 10 testų ir imamos vidurkinės reikšmės, kiekviename teste buvo parinkta po 6000 iteracijų, o kiti nustatymai buvo keičiami. Naujajam BNN metodui buvo naudojamas autorių sukurtas produktas, o tuo tarpu senesniam BNN buvo pasirinktas jau sukurtas produktas, kurio autorius R.M.Neal ir jį galima parsisiųsti iš http://www.cs.toronto.edu/. Šiam klasifikatoriui taip pat atlikta 10 testų, tačiau iteracijų skaičius buvo 30, bet jis yra adekvatus naujajam BNN metodui su 6000 iteracijų, nes jis

turi įvairių po-žingsnių (angl., substeps), o taip pat naudoja įvairius atrankos metodus, skiriasi ir paslėptų neuronų skaičius, kurių tikslūs duomenys pateikti žemiau esančiame paveiksle (pav. 22). Abejų metodų CPU sunaudojamas laikas yra panašus. SVM klasifikatorius (LIBSVM) buvo taip pat parsisiųstas iš http://www.csie.ntu.edu.tw/»cjlin, kurio autoriai Chang and Lin. Pastarajam branduolio (angl., kernel) reikšmės C buvo parinktos nuo 0.001 iki 10, o rezultatai taip pat pateikti paveiksle 22, o kadangi SVM metodai yra kur kas greitesni skaičiavimo atžvilgiu jų CPU sunaudotas laikas nebuvo įtrauktas ir buvo atliktas po 1 testą kiekvienam C, be to testavimo klaidos standartinių nuokrypų vidurkiai laikomi artimais nuliui. Eksperimentui pasirinktas tiesinis branduolys, nes prieš tai išbandyti radialinių bazinių funkcijų bei polinominiai branduoliai ir jų generalizavimo savybės buvo prastesnės už tiesinę funkciją.

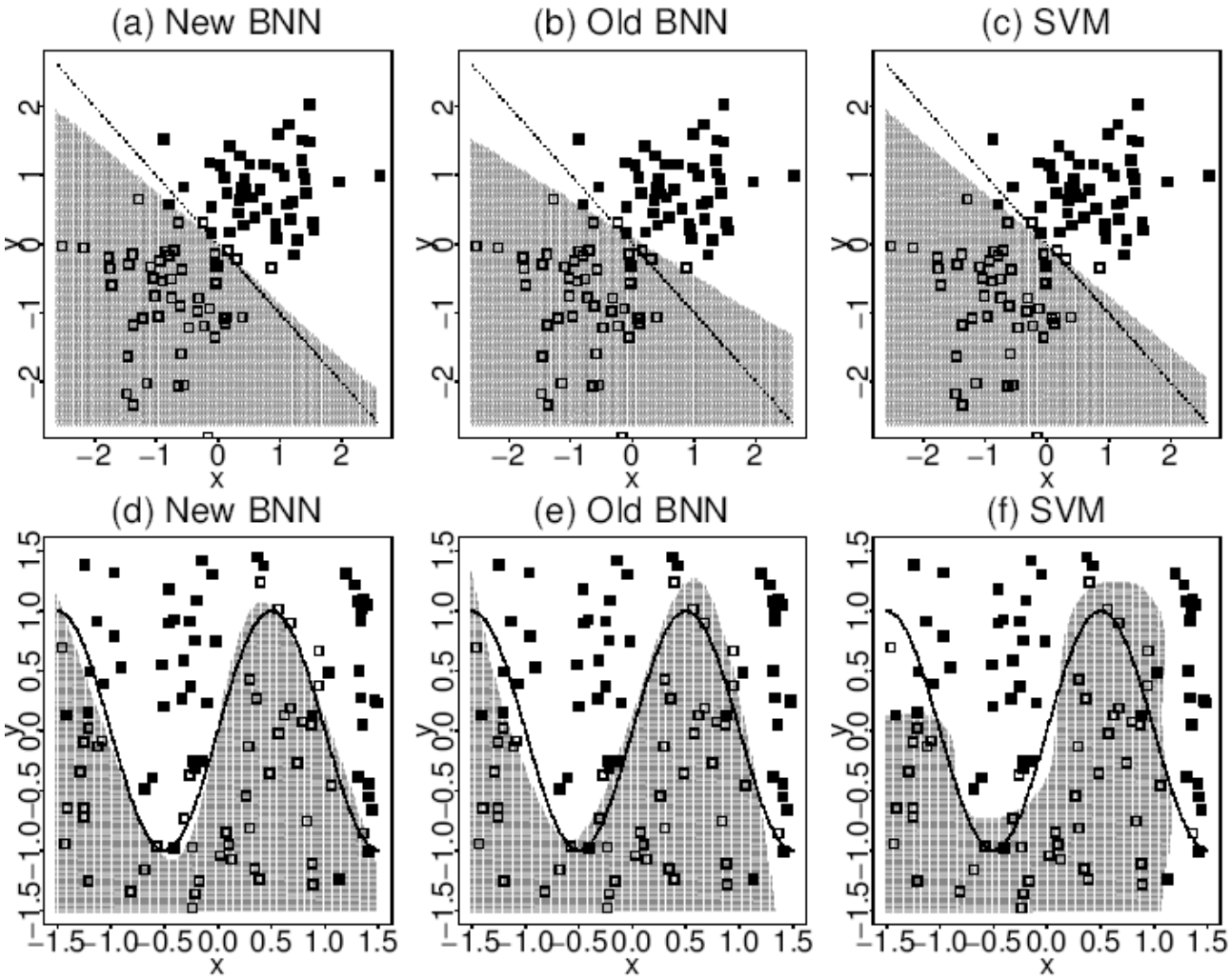
	Parametrai	CPU	Treniravimas	Testavimas
Naujasis BNN	$\lambda = 10$	47.1	5.426(0.026)	62.032(0.097)
	$\lambda = 15$	47.3	5.416(0.026)	61.708(0.100)
	$\lambda = 20$	47.5	5.388(0.023)	61.726(0.096)
	$\lambda = 25$	47.6	5.364(0.024)	61.706(0.099)
Senasis BNN	$M = 2$	39.9	5.206(0.026)	63.52(0.11)
	$M = 3$	47.6	5.192(0.029)	63.43(0.12)
	$M = 4$	55.3	5.130(0.030)	63.62(0.11)
	$M = 5$	62.6	5.166(0.031)	63.63(0.12)
(Tiesinis) SVM	$C = 0.001$	—	7.24	72.32
	$C = 0.01$	—	5.36	61.94
	$C = 0.10$	—	5.60	61.64
	$C = 1.0$	—	5.40	63.40
	$C = 10.0$	—	5.18	64.16

Šaltinis: [27]

22 pav. Eksperimento rezultatai (BNN ir SVM)

Paveiksle 22 grafoje CPU nurodyti CPU apkrovimo laikai sekundėmis dešimčiai testų, o stulpeliai pavadinimais treniravimas ir testavimas rodo suvidurkintas klaidas ir šalia skliausteliuose standartinės nuokrypos. Reziumuojant 22 paveikslą, matoma, kad naujasis BNN turi geresnes generalizavimo savybes už ankstesnįjį BNN, nors jis ir darė mažiau klaidų apmokant, tai sąlygoja per didelį prisitaikymą prie apmokymo duomenų. SVM šiuo atveju rodo panašius rezultatus į naująjį BNN, bet, sprendžiant labiau komplikuotas užduotis, SVM rezultatai rodo prastesnius rezultatus. Vizualiai šie duomenys pateikti kitame paveiksle (23), kur paimti du skirtingi duomenų rinkiniai ir stulpeliais paeiliui parodyta naujojo BNN, ankstesniojo BNN ir SVM metodų gebėjimas

klasifikuoti. Tuščiaviduriai ir pilnaviduriai kvadratukai rodo dvi skirtingas klases, o linijos parodo realų klasių pasiskirstymą, kai tuo tarpu metodų gebėjimas klasifikuoti vaizduojamas pilkame fone.



Šaltinis: [27]

23 pav. Klasifikavimo rezultatai (BNN ir SVM)

Pastarajame eksperimente, BNN klasifikavimo tikslumas buvo apie 96%. Tame pačiame darbe buvo atlikti dar keletas eksperimentų, kurie buvo kur kas sudėtingesni, jų metu klasifikuoti krūties vėžio atvejai bei proteino struktūros, tačiau smulkiau šie eksperimentai šiame darbe nebus aptariami, tik galima pažymėti, jog visuose geriausiai generalizavimo savybėmis pasižymėjo naujasis BNN metodas, o taip pat buvo įtrauktas SVM paremtas radialinėmis bazinėmis funkcijomis krūties vėžio eksperimente ir paliktas tik jis proteino struktūros klasifikavimui, nes pasirodė geriau nei tiesinio branduolio atveju.

Taigi šio darbo metu empiriškai parodyta, kad Bajeso neuroniniai tinklai gali būti efektyviai naudojami klasifikavimui ir pristatytas naujas metodas, kuris turi geresnes generalizavimo savybes nei prieš tai naudotas metodas bei SVM metodų variacijos, o tai lėmė dvi priežastys: nauja

tikimybinė funkcija ir globalios duomenų informacijos naudojimas sprendimų riboms (angl., decision boundary) nustatyti, kai tuo tarpu SVM naudoja lokalią informaciją. BNN sėkmingai klasifikuoja tiek tiesinius tiek netiesinius duomenis.

Dar kitame darbe [28] parodyta, kad neraiškių aibių (angl., Fuzzy sets) ir tikimybinio Bajeso metodais ekspertų žiniomis sumodeliuotos sistemos gali būti efektyviai naudojamos spektroskopiniuose tyrimuose. Šio eksperimento tikslas – turint pradinius duomenis sumodeliuoti galimą vibracijos dažnių informaciją. Galiausiai prieita išvadų, jog abu metodai rodo panašius rezultatus.

1.20. Analitinės dalies išvados

1. Aptartos pagrindinės kredito rizikos sąvokos, valdymo bei vertinimo būdai. Kredito rizikos vertinimas gali apimti ir kitų rizikų vertinimą (pvz., viso portfelio rizika), atkreiptas dėmesys ir į rizikos išmatavimo būdus, kaip VaR bei scenarijų technikos. Šiame darbe apsiribojama pavienio kredito rizikos vertinimu.
2. Apžvelgti pagrindiniai klasikiniai kredito rizikos vertinimo metodai, tiek matematiniai-statistiniai, tiek ir modernūs, aptarti jų naudojimo privalumai ir trūkumai, įvertintos dirbtinio intelekto metodų (skirtų klasifikavimui bei klasterizavimui) panaudojimo su jais galimybės.
3. Atlikus klasikinių ir DI metodų taikymo kredito rizikos vertinimui lyginamąją analizę, išryškinti dirbtinio intelekto metodų privalumai, todėl buvo atlikta jų platesnė analizė.
4. Išskirti pagrindiniai ir plačiausiai naudojami mašininio mokymo metodai (neuroniniai tinklai, atramos vektorių mašinos, Bajeso metodas), apžvelgti jų ankstesni taikymai nagrinėjamai problemai bei jų privalumai ir trūkumai, o taip pat aptarti kiti žinomi dirbtinio intelekto metodai: evoliuciniai skaičiavimai, *fuzzy* logika, k-artimiausio kaimyno ir klasifikavimo medžių metodai).
5. Aptarta eilė išsamių tyrimų atliktų įvairių autorių su skirtingais klasifikavimo metodais. Galima daryti išvadas, kad vieni geriausių ir plačiausiai naudojamų yra SVM, neuroninių tinklų ir medžių bei KNN metodai, kurie rodo tikslesnius klasifikavimo rezultatus, nuo jų kiekybiškai ir kokybiškai šiek tiek atsilieka Bajeso metodas, tačiau parodyta, jog specifinėms problemoms spręsti jis rodo geresnius rezultatus už visus paminėtus metodus bei bendru požiūriu turi pranašumo greičio ir paprastumo atžvilgiu, todėl vėlesniems tyrimams pasirinktas Bajeso metodas.
6. Pasirinkto metodo tinkamumu analizuojamai problemai spręsti galimi daugelis įvairių kriterijų ir vieni iš jų – klasifikavimo tikslumas ir klaidų skaičius. Iš pateiktų metodų tarpusavio analizių matyti, kad tikslumas svyruoja nuo 10% iki beveik 100% (klasifikuojant į dvi klases), tačiau

daugeliu atveju įvairūs metodai klasifikuoja nuo 50% iki 90%, todėl pasiekus bent 70% tikslumą būtų galima tvirtinti, kad lyginant su kitais metodais Bajeso metodas yra tinkamas. Klaidų svyravimas šaltiniuose nuo 0.195 iki beveik 0.4. Žinoma, reiktų atsižvelgti ir į faktą, kad šio eksperimento metu bus klasifikuojama ne į dvi, o tris klases, sudarant sudėtingesnes sąlygas klasifikatoriui.

7. Taigi šio darbo metu planuojama panaudoti Bajeso metodą kredito rizikai valdyti, t.y, turint kompanijų finansinius duomenis, pagal juos suskirstyti jas į mokias ir nemokias. Duomenys bus paimti iš ankstesnio VU KHF doktoranto P.Danėno daryto magistrinio darbo, taikant SVM metodą tokiam pačiam tikslui, tik galbūt juos papildant, o galiausiai, esant galimybei palyginti gautus rezultatus. Prieš tai tuos pačius duomenis savo daktaro laipsnio disertacijai naudojo ir kitas VU KHF doktorantas E.Merkevičius. Visiems tyrimams vadovavo dr. doc. Gintautas Garšva.

2. BAJESO METODU PAREMTO EKSPERIMENTO APRAŠYMAS

Šiame skyriuje aprašomas siūlomas eksperimentinis sprendimas, jo prototipas ir įvairūs veiklos aspektai. Aptariama, kaip šią sistemą vėliau būtų galima patobulinti bei išplėsti.

2.1. Naudojamų duomenų ir programinės įrangos aprašymas

Sistemos realizacijai pasirinktas moksliniuose tyrimuose plačiai naudojamas JAVA programavimo kalba sukurtas Weka 3.6.2 programinis atvirojo kodo įrankis, kuris leidžia realizuoti jo aplinkoje įvairius specifinius modulius (tokius, kaip duomenų importavimas, duomenų įvedimas ir pan.) bei sąveiką su duomenų bazėmis ir įvairių formatų duomenų failais. Šis paketas naudojamas daugelyje įvairių tyrimų susijusių su klasifikavimu, klasterizavimu, prognozavimu ir t.t., dalis jų pateikiama literatūros sąraše [19, 24, 26, 35, 37, 38, 39, 40, 41].

Weka paketas leidžia jungtis tiek prie duomenų bazių, tiek importuoti duomenis iš failų, todėl pasirinktas antrasis variantas, eksportuojant duomenis iš SQL į CSV failo formatą. Duomenų saugojimui pasirinkta Firebird duomenų bazė, nors galimas variantas ir atvirojo kodo MySQL duomenų bazė, nes abi jos leidžia greitai apdoroti didelius duomenų kiekius bei pasižymi patikimumu, o taip pat yra daugiaplatformės (Windows, Linux, kitos operacinės sistemos), abi gali eksportuoti duomenis į įvairius failo formatus.

Eksperimentui naudojami EDGAR tarptautinės duomenų bazės 1999 – 2003 m. duomenys, kurie apima beveik 10000 tarptautinių įmonių įvairius finansinius duomenis, o pagrindą sudaro metinių ir ketvirtinių balansų bei pelno ataskaitų duomenys. Visi duomenys suskirstyti į 9 stambesnius sektorius, kurie pateikti žemiau esančioje lentelėje. Kokių veiklų firmos patenka į šiuos sektorius pateikta 2 priede.

5 lentelė

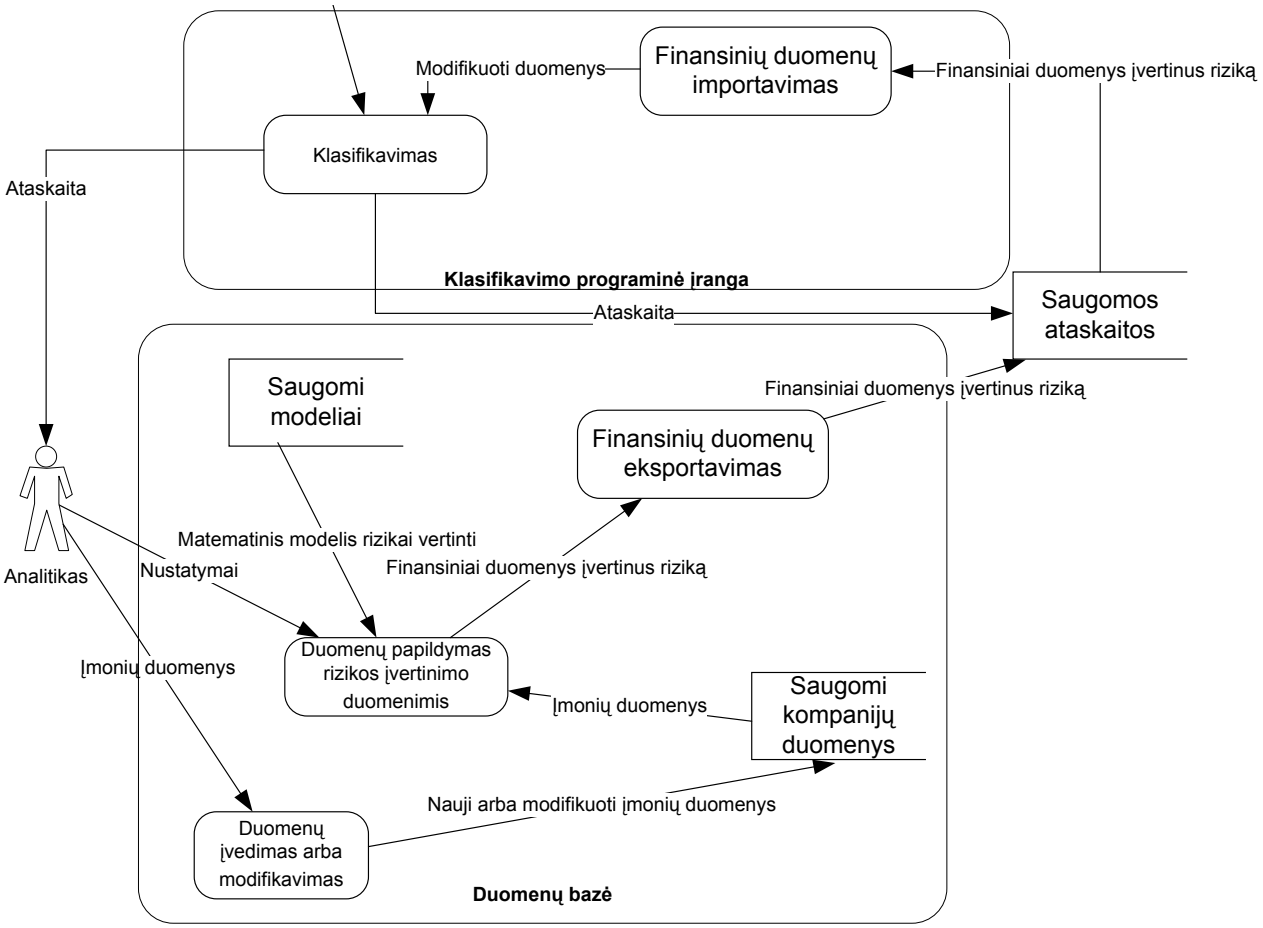
Kompanijų skirstymas į 9 sektorius

Sektoriaus kodas	Pavadinimas	Kompanijų skaičius
01-09	Agriculture, Forestry, And Fishing	33
10-14	Mining	469
15-17	Construction	83
20-39	Manufacturing	3027
40-49	Transportation, Communications, Electric, Gas, And Sanitary Services	786
50-51	Wholesale Trade	287
52-59	Retail Trade	405
60-67	Finance, Insurance, And Real Estate	1853
70-89	Services	1712

Iš viso naudoti 79 pirminiai ir išvestiniai rodikliai (1 priedas). Išeigos reikšmėms naudotas Altmano modelis, gamybinėms kompanijoms - vienas, kitoms - kitas (standartinis). Naudotas atributų (rodiklių) atrinkimas pagal reikšmingumą, naudojant genetinį algoritmą (feature selection),

taip sumažinant sugeneruoto modelio apimtį, tai atliko tyrimo kuratorius P. Danėnas, o atrinkti atributai pateikiami 3 priede. Modelio parametrai imti "pagal nutylėjimą“.

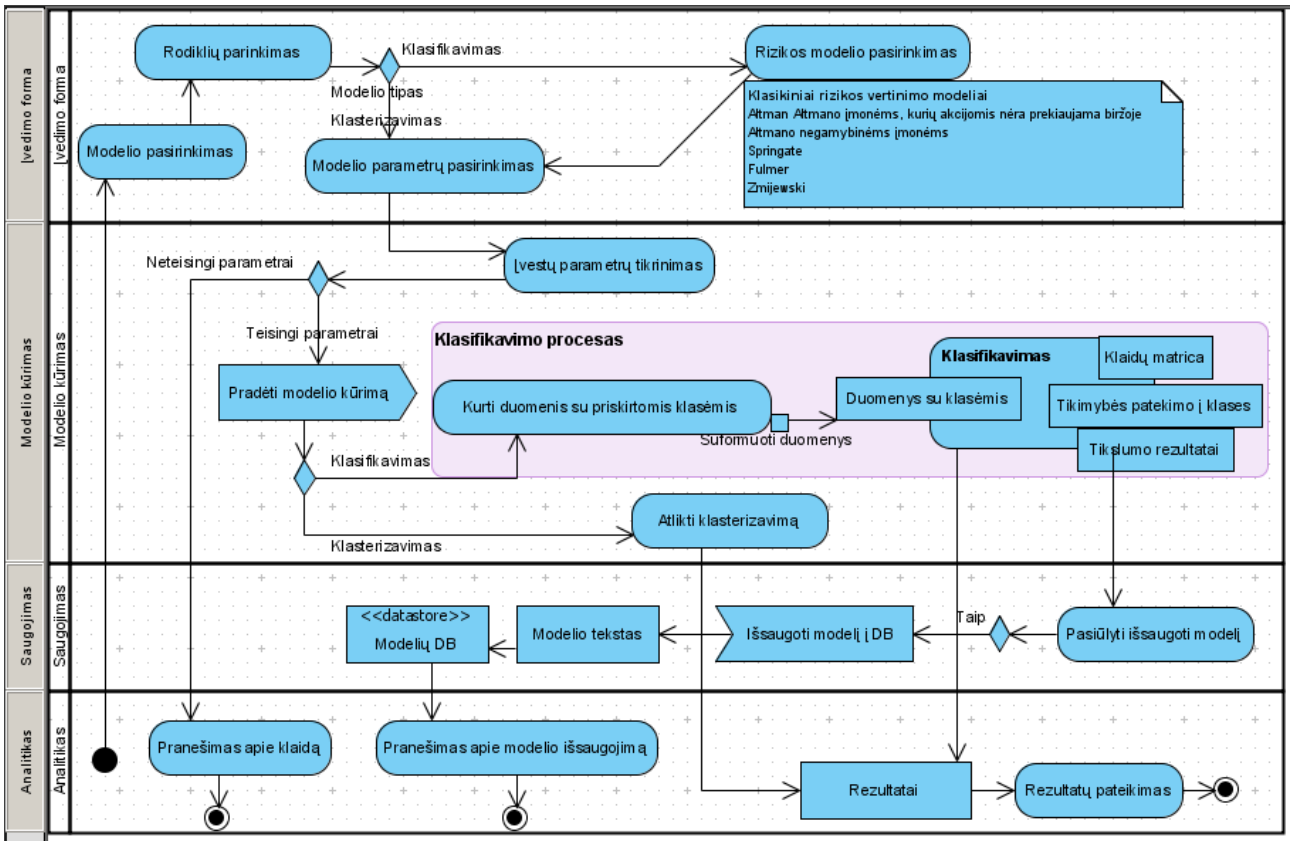
Informacijos srautai iliustruojami duomenų srautų diagrama (24 pav.). Duomenų srautų nėra itin daug, tačiau kai kurie jų apima didelį parametrų arba informacijos kiekį.



Šaltinis: sudaryta autoriaus

24 pav. Sistemos duomenų srautų diagrama

Sistemos veikimą galima iliustruoti sekos diagrama (25 pav.). Ši diagrama apima visą veiksmų seką nuo modelio pasirinkimo iki jo išsaugojimo. Galima pastebėti, kad į šią diagramą įtrauktas klasterizavimas; tai irgi viena iš galimybių plėsti sistemą, įdiegiant ir modernius klasterizavimo metodus bei juos pritaikant kredito rizikos vertinimo sričiai. Svarbu atsižvelgti į tai, kad šiuo atveju nebus sukuriamas ir išsaugojamas joks modelis; klasterizavimo proceso metu duomenys tik suskirstomi į klasterius pagal ryšius tarp jų [19].



Šaltinis: Danėnas P., Merkevičius E., Garšva G. Sistemos modulio, skirto intelektualių modelių kredito rizikos vertinimui kūrimui, koncepcinė struktūra [43]

25 pav. Galima modelio kūrimo veiksmų sekos diagrama

Sistema ateityje gali būti ir plečiama, integruojant naujus modelius bei metodus. Tam gali būti panaudotos įvairios bibliotekos bei programiniai komponentai, realizuojantys vieną ar daugiau mašininio mokymo algoritmų. Kadangi Weka mašininio mokymo sistemoje realizuoti beveik visi populiariausi modernūs mašininio mokymo algoritmai, tokie, kaip Bajeso tinklai, neuroniniai tinklai, sprendimų medžio ir kiti metodai, integravus šiuos algoritmus, sistema gali būti išplėsta iki sistemos, kuri galėtų netgi palyginti dviejų ar daugiau skirtingų algoritmų rezultatus (Danėnas, Merkevičius, Garšva, 2008).

Lygiai taip pat galimas ir kitų atvirojo kodo bibliotekų integravimas (Danėnas, Merkevičius, Garšva, 2008):

- daug mašininio kodo algoritmų realizacijų Java kalba galima rasti SourceForge portale, platinančiame atvirojo kodo programas bei bibliotekas, be to, daug bibliotekų ir realizacijų Java ir kitomis programavimo kalbomis galima rasti ir kituose Interneto tinklalapiuose;
- yra ir galimybė naudoti MATLAB skaičiavimų variklį, tam naudojant JMatLink biblioteką – galimas sistemos komunikavimas su MATLAB sistema bei tikėtina (bet dar netirta) galimybė modelių rezultatus išsaugoti sistemos DB;
- galimybė integruoti kitos atvirojo kodo sistemų algoritmus, juos adaptuojant sprendžiamai

problemai. Iš jų galima išskirti kompanijos Rapid-I YALE[1] duomenų gavybos sistemą bei Java duomenų gavybos standartą ir jo realizaciją JDM[2];

- integravus ekspertinį posistemį (naudojant jFuzzyLogic[3], Jess[4] ar kitas JAVA ekspertinių sistemų kūrimo priemones), galima sukurti ir galimybę klasifikavimo ir klasterizavimo procesuose naudoti ekspertines žinias bei kurti ir neraiškių aibių modelius.

2.2. Eksperimente naudojami metodai

Eksperimente naudojamas vienas iš populiariausių ir plačiausiai taikomų duomenų gavybos ir statistikos metodų – klasifikavimas, kurio pagrindinė užduotis yra suskirstyti atributų vertes pagal galimas klases. Šis metodas svarbus įvertinant tiek dabartines, tiek ir prognozuojamas vertes; prognozavimas gali būti traktuojamas kaip atributo vertės klasifikavimas į vieną iš galimų klasių (Dunham, 2003, cit. pagal šaltinį [19]).

Pati klasifikavimo problema formaliai apibrėžiama taip (Dunham, 2003, cit. pagal šaltinį [19]): turint duomenų bazę su $D = \{t_1, t_2, ..., t_n\}$ įrašų (vienetų) ir klasių aibę $C = \{C_1, C_2, ..., C_n\}$, klasifikavimo problema yra susiejimo $f : D \rightarrow$, kurioje kiekvienas t_i yra priskirtas kuriai nors klasei. Klasėje C_j yra tiksliai tik tie įrašai, kurie su ja susieti, t.y., $C_j = \{t_i \mid f(t_i) = C_j, 1 \leq i \leq n, t_i \in D\}$.

Šios problemos sprendimui gali būti naudojami trys pagrindiniai metodai (Dunham, 2003):

1. **Ribų apibrėžimas** – klasifikacija atliekama padalijant galimų įeigos įrašų erdvę į regionus, kur kiekvienas regionas susietas su viena klase;

2. **Tikimybių pasiskirstymo naudojimas** – kiekvienai duotai klasei C_j, $P(t_i \mid C_j)$ yra tikimybių pasiskirstymo funkcija (*probability distribution function*, sutr. PDF) klasei, vertinamai taške t_i;

3. **Vėlesnių (*posterior*) tikimybių panaudojimas** – turint duomenų reikšmę t_i, reikia apibrėžti tikimybę, kad t_i yra klasėje C_j. Tai žymima $P(C_j|t_i)$ ir vadinama vėlesne tikimybe. Vienas iš tokios klasifikacijos požiūrių būtų apibrėžti vėlesnę tikimybę kiekvienai klasei ir tada priskirti t_i klasei su didžiausia tikimybe.

Pastaruoju metodu remiamasi ir šiame darbe.

---

[1] Rapid-I [interaktyvus]. Adresas Internete: http://rapid-i.com/content/blogcategory/10/69/lang,en/.
[2] JSR-000073 Data Mining API (Final Release). Adresas Internete: http://jcp.org/aboutJava/communityprocess/final/jsr073/index.html
[3] jFuzzyLogic: Open Source Fuzzy Logic (Java). Adresas Internete: http://jfuzzylogic.sourceforge.net/html/index.html
[4] Jess, the Rule Engine for the Java Platform[interaktyvus]. Adresas Internete: http://herzberg.ca.sandia.gov/

2.3. Eksperimentinio tyrimo metodikos aprašymas

Siūlomas sprendimas apimtų klasikinių diskriminantinių modelių ir Bajeso metodo integravimą. Šiuo atveju diskriminantinis modelis būtų naudojamas suskaičiuoti išeigos reikšmes, t.y., įvertinant įmonės rizikingumą, priskirti ją prie vienos iš trijų klasių (rizikinga, vidutinio rizikingumo ar nerizikinga). Bajeso metodas naudojamas modelio gavimui, t.y., apskaičiuojami jo koeficientai, kuriuos naudojant, būtų gaunama duomenų be išeigos reikšmių tiksliausia klasifikacija. Šios problemos sprendimas apima ne binarinį klasifikavimą (į dvi klases – „rizikinga“ ir „nerizikinga“), o skirsto jas į tris klases. Kiti tyrimo žingsniai apimtų šiam tinkamiausių diskriminantinių modelių nustatymą bei geriausių Bajeso metodo parametrų ir paieškos algoritmų parinkimą.

Nustatant tinkamiausius diskriminantinius modelius, atliekamas apmokymas, naudojant visus rodiklių (iš viso 79). Šiuo atveju kuriami modeliai, naudojant kurį nors iš taikomų diskriminantinių modelių (Altman, Springate, Zmijewski) bei vieną iš Bajeso metodų. Vėliau naudojamas indukcinis principas, kai laikoma, kad modelis, su standartinėmis metodo reikšmėmis parodęs geriausius rezultatus, yra geriausias ir ieškoma būdų, kaip su geriausia kombinacija pasiekti dar geresnių rezultatų.

Šiame etape bus tiriami du Bajeso tinklų klasifikatoriai, tai Naivaus Bajeso ir BayesNet metodai (po 2 abiejų), o vėliau atrenkami du geriau pasirodę ir tiriami kartu su analogišku SVM metodu. Kaip jau buvo minėta, Bajeso tinklai yra tikimybiniai kryptiniai necikliniai grafai, sudaryti iš kintamųjų (mazgų, angl., nodes) ir kryptinių briaunų (angl., arcs), jungiančių kintamuosius. Kiekvienas kintamasis turi baigtinį skaičių jiems priskirtų būsenų ir kiekvienas kintamasis A_k su tėviniais mazgais B_1; ...;B_n turi priskirtą tikimybių pasiskirstymo lentelę $P(A_k| B_1; ...;B_n)$. Šiame tinkle kiekvienas mazgas reiškia atributą, o kiekviena briauna reiškia priežastinius ryšius tarp jų. Kiekvienas mazgas turi savo tikimybių lentelę, kuri kaupia informaciją apie visų įmanomų atributo būsenų jungtinį tikimybių pasiskirstymą, kai žinomi visi jo tėviniai mazgai. Šios lentelės vėliau naudojamos nuspėti klasės tikimybę bet kokiam duotam atvejui. Tikimybių pasiskirstymas skaičiuojamas pagal formulę:

$$P(X_1, ..., X_n) = \prod_{i=1}^{n} P(X_i|\text{parent}(X_i)) \quad (9),$$

kur n yra bendras atributo būsenų skaičius, X_i yra X atributo būsena, o $parent(X_i)$ reiškia būsenos X_i tėvinius mazgus.

Minėtiems Bajeso tinklų sudėtingumas didėja, kai kintamasis turi kuo daugiau atributų ir klasių, nes susidaro didžiuliai kiekiai būsenų, kurias reikia apskaičiuoti.

Naivaus Bajeso metodas, kaip jau minėta anksčiau, pasižymi tuo, kad jis kiekvieną atributą laiko vienodai reikšmingu priskiriant vienai ar kitai klasei, o taip pat daroma prielaida, jog jie tarpusavyje yra nekoreliuoti. Šio metodo principu tikimybių pasiskirstymas skaičiuojamas šia formule:

$$P(X, y) = P(y)P(X|y) = P(y)\prod_{d=1}^{n} P(X_d|\text{parent}(X_d)) \quad (10)$$

Iš gautų reikšmių tinkamiausio sprendinio parinkimui naudojama 8 formulė.

Bajeso tinklų klasifikatoriai (šiuo atveju BayesNet) naudoja skirtingus paieškos algoritmus, jog surastų optimaliausią duomenų pateikimą. Skirtingiems paieškos algoritmams gali skirtis klasifikavimo tikslumas ir greitis. Populiariausi paieškos algoritmai yra šie [42]:

- **K2** – tai įvertinimais paremtos godžios paieškos (angl., score-based greedy search) algoritmas, skirtas Bajeso tinklams apsimokyti iš duotų duomenų. Šis algoritmas maksimizuoja optimalaus grafo topologijos tikimybę naudojant Bajeso įverčius, jog išdėstytų pagal eiliškumą (angl., ranking) skirtingus grafus. Algoritmas yra apribotas kintamųjų išdėstymo tvarka.
- **Genetinė paieška** – remiasi natūralioje gamtoje būdingu populiacijos principu, kur vyrauja natūrali atranka ir laimi stipriausieji. Etapais atrenkamas geriausias galimas sprendimas atrenkant geriausiai pasirodžiusius sprendinius, proporcinės atrankos arba selekcijos taikymas reiškia, kad individai tolesnėms generacijoms parenkami proporcingai jų kokybės funkcijai. Genetiniuose algoritmuose pagrindinis reprodukcijos operatorius yra bitų eilutės kryžminimas. Jo metu imamos dvi bitų eilutės kurios reprezentuoja „tėvus“ ir sukuriami du nauji individai. Bitų eilutės pjūvio taškas pasirenkamas atsitiktinai. Kryžminimo metu pjūvio taške atskirtos tėvų dalys sukeičiamos vietomis ir suformuojamos naujos bitų eilutės. Kryžminimas taip pat atliekamas naudojant kelis pjūvio taškus vienu metu. Kitas populiarus genetinių algoritmų operatorius yra mutacija. Mutacijos metu tėvinėje bitų eilutėje yra pakeičiamas (paprastai invertuojamas) vienas atsitiktinai pasirinktas bitas ir tokiu būdu gaunamas naujas palikuonis. Skirtumas tas, kad kryžminimas, neįneša jokios naujos informacijos, kai tuo tarpu mutacijos operatorius įneša į populiaciją naują informaciją. [4, 42]
- **Dirbtinis atkaitinimas** – procedūra yra atsitiktiniu ieškojimu grindžiamas procesas, efektyvus ieškant globalinio funkcijos ekstremumo (minimumo ar maksimumo). Idėja paimta iš metalurginio šaldymo ar skysčių užšalimo ir kristalizavimosi procesų. Atomai skysčiuose juda

labai greitai, esant aukštoms temperatūroms, ir lėčiau, temperatūrai krintant. Lėto aušinimo metu išnyksta gardelių dislokacijos ir įtempimai. Globalinė sistemos energija tokiu atveju pasiekia absoliutų minimumą. Greito aušinimo metu atomams nelieka laiko išsirikiuoti, ir sistema lieka aukštoje energetinėje būsenoje. Per greitai mažinant temperatūrą įtempimai esantys metale ir užšąlančiame ar besikristalizuojančiame skystyje "įšąla". Matematiniai įrodymai, kad algoritmas randa globalinį funkcijos minimumą remiasi tuo, kad patikrinama be galo daug w pakeitimų. Jei po tam tikro laiko nebus priiminėjami blogi sprendimai, atsitiktinai ieškant kada nors bus surastas geriausias. Tačiau praktiniai rezultatai rodo, kad jau per baigtinį iteracijų skaičių pasiekiami neblogi rezultatai. Labai dažnai į globalinį minimumą papuolama gana greitai. [4, 42]

- **Greedy** (godus) **Hill Climber** (HC) – šiuo atveju reikia įsivaizduoti, kad visi galimi klasifikavimo problemos sprendimai yra atvaizduojami „trimatėje plokštumoje" (angl., three-dimensional landscape). HC eis grafu iš mazgo įmazgą didinant sprendimo įvertinimą tol, kol nepasieks lokalaus minimumo. Šis Bajeso tinklų apsimokymo algoritmas naudoja HC pridedant, išmetant ar perrikiuojant grafo briaunas. Skirtingai nuo K2 algoritmo, paieška nėra apribojama kintamųjų išsidėstymu.
- **Repeated** (pakartotinis) **Hill Climber** (RHC) – šis algoritmas tinkamiausios Bajeso tinklo struktūros ieško pakartotinai generuojant atsitiktinius tinklus ir pritaikant aukščiau minėta HC algoritmą. Šio algoritmo privalumas yra tas, jog kai HC užstringa kuriame nors mazge, pasirenkamas naujas atsitiktinis mazgas ir HC algoritmas pradedamas iš naujo. Ši procedūra kartojama *k* kartų, o sprendiniu pateikiamas geriausias rastas maksimumas.

Kai kurie galimi naudojami parametrai pateikti žemiau:

UseKernelEstimator (k). Kai šis parametras nustatytas, tinklas labiau tinkamas skaitiniams atributams nei normaliniam išsidėstymui, t.y., efektyvumas turėtų matytis kintant skaitinių reikšmių kiekiui tarp atributų.

UseSupervisedDiscretisation (Sd). Naudojama skaitines reikšmes konvertuoti į nominalias.

initAsNaiveBayes (iNb). Pagal nutylėjimą ši reikšmė būna parinkta, ji reiškia, kad pradinė struktūra apsimokymui yra naivus Bajeso tinklas, kitu atveju pradedama nuo tuščio tinklo.

markovBlanketClassifier (Mb). Kai tinklo struktūra yra išmokta, korekcijos atliekamos pagal Markov blanket kiekvienam mazgui, t.y., kiekvienas mazgas kaupia savyje informaciją ir yra susietas ne tik su savo tėviniais ir vaikiniai mazgais, bet ir su tėvų tėvais ir vaikų vaikais.

RandomOrder (R). Atsitiktinai sukeičia kintamųjų atributus, pakeičiant juos vietomis nuo pradinių įkeltų vietų importuojant duomenis.

useArcReversal (Ar). Kai ši reikšmė pasirenkama, briauna tarp dviejų mazgų yra pakeičiama atitinkmai priskiriant tėvinę briauną.

useTournamentSelection (Ts). Šis parametras nusako populiacijos parinkimą genetiniame paieškos algoritme. Kai ši reikšmė teigiama, parenkami du skirtingi tinklai ir geresnis iš jų praleidžiamas toliau, priešingu atveju tikrinami visi tinklai ir geriausią įvertinimą turintys parenkami tolesniam žingsniui.

Šiame darbe paieškos algoritmu buvo pasirinktas K2, tačiau ateityje galima eksperimentą praplėsti ir panaudoti kitus paminėtus algoritmus. Naiviam Bajesui buvo panaudoti šie parametrai: *UseKernelEstimator, UseSupervisedDiscretisation, initAsNaiveBayes.* Ateityje galimi ir kiti parametrų išbandymai.

2.4. Eksperimentas ir jo rezultatas

Siekiant įsitikinti šiame darbe atliktu analizės teisingumu, buvo atliktas tyrimas, kurio tikslas – išsiaiškinti Bajeso metodo tinkamumą kredito rizikos vertinimui. Kaip jau buvo minėta, pasirinkta mašinininio mokymo sistema Weka, kuri integruota su Firebird duomenų bazėje saugomais duomenimis. Pagal analogišką tyrimą SVM metodui testuoti [19] iš trijų modelių Altman, Zmijewski bei Springate pasirinktas Altman: gamybinėms kompanijoms - vienas, kitoms - kitas (standartinis). Eksperimento metu buvo nustatyta kuris iš keturių Bajeso metodų variantų teikia geriausius rezultatus (BayesNet ir Naive Bayes su reikšmėmis pagal nutylėjimą, su parametru *KernelEstimator* ar su *SupervisedDiscretization*). Eksperimentui naudojami visi galimi rodikliai (iš viso 79 atribudai; jų sąrašas pateikiamas 1 priede).

Įvertinimui naudojama standartinė įvertinimo metodika, dar žinoma kaip maišos matrica[5] (angl., confusion matrix), bei pagrindiniai rodikliai, apskaičiuojami iš jos verčių.

Esant atvejams, kai galimi tik du sprendiniai, o ne trys, kaip šio darbo atveju, maišos matrica apibrėžiama kaip matrica, turinti tokias reikšmes [19, 45]:

- a yra **teisingų** spėjimų skaičius, kad atvejis yra **neigiamas**;
- b yra **neteisingų** spėjimų skaičius, kad atvejis yra **teigiamas**;
- c yra **neteisingų** spėjimų skaičius kad atvejis yra **neigiamas**;
- d yra **teisingų** spėjimų skaičius, kad atvejis yra **teigiamas**.

[5] Maišos matrica gali būti apibrėžta bet kuriam klasių skaičiui; čia ji apibrėžiama tik sprendžiamos problemos atveju (binariniam kalsifikavimui)

		Prognozė	
		Neigiamas	**Teigiamas**
Sistema	**Neigiamas**	**a**	**b**
	Teigiamas	**c**	**d**

Iš šios jos gaunami parametrai, kurie gali būti naudojami tikslumo ir efektyvumo vertinimui:

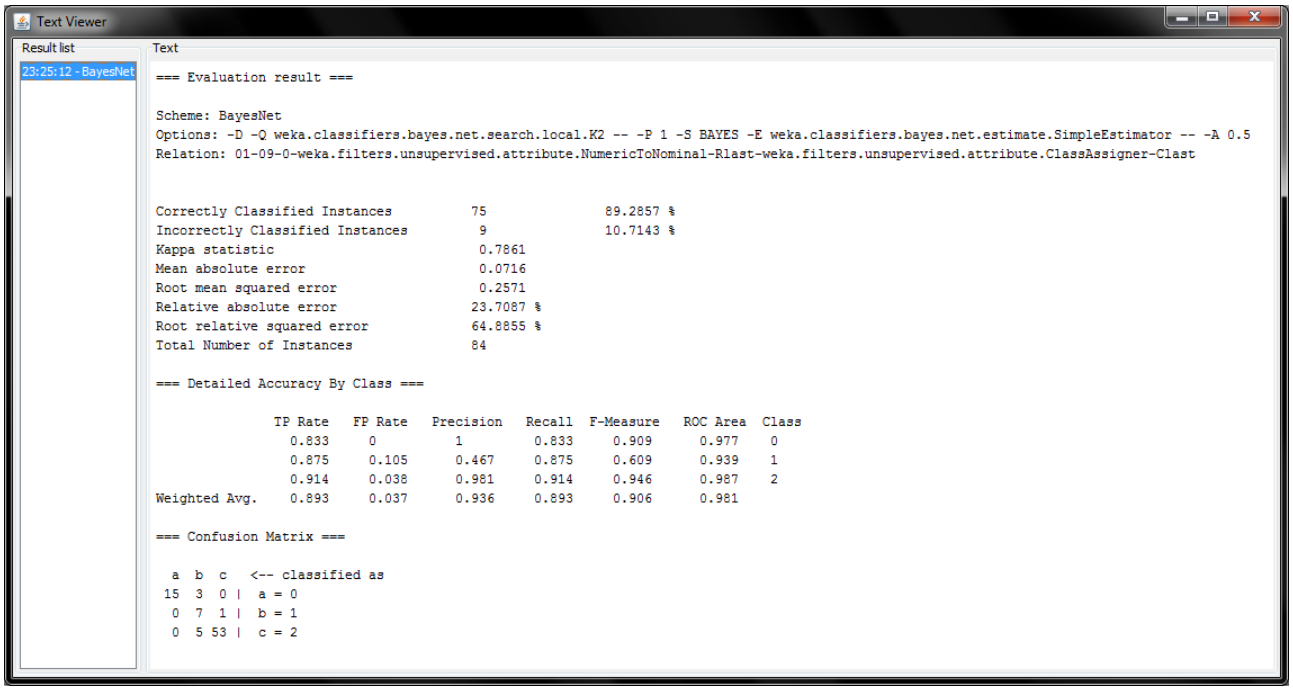
- Teisingumas (accuracy) AC yra teisingų spėjimų skaičiaus ir bendro spėjimų skaičiaus proporcija. Ji apskaičiuojama taip: $AC =$
- *Recall* arba teisingų teigiamų rodiklis (true positive rate) arba TP yra teigiamų atvejų, kurie buvo teisingai identifikuoti, proporcija, apskaičiuojama taip: $TP =$
- Neteisingų teigiamų rodiklis (the false positive rate) arba FP yra neigiamų atvejų, kurie buvo neteisingai suklasifikuoti kaip teigiami, proporcija, apskaičiuojama taip: $FP =$
- Teisingų neigiamų rodiklis (the true negative rate) arba TN apibrėžiamas kaip neigiamų atvejų, kurie buvo teisingai suklasifikuoti, proporcija, apskaičiuojamam taip: $TN =$
- Neteisingų neigiamų rodiklis (the false negative rate) arba FN apibrėžiamas kaip teigiamų atvejų, kurie buvo neteisingai suklasifikuoti kaip neigiami, propocija: $FN =$
- Tikslumas (precision) P yra prognozuojamų teigiamų teisingų atvejų propocija, apskaičiuojama taip: $P =$
- Teisingumo įvertis gali būti neadekvatus našumo matavimas, kai neigiamų atvejų skaičius yra daug didesnis nei teigiamų atvejų skaičius. Šiuo atveju apbrėžiamas geometrinis vidurkis (g-mean) ir F-Measure rodiklis:

$$g - \quad = \quad ; \; g - \quad =$$

- F_1 rodiklis (F-Measure) yra harmoninis tikslumo ir recall įverčiųvidurkis:

$$F_1 = \frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} = \frac{2}{\frac{1}{\text{Precision}} + \frac{1}{\text{Recall}}}$$

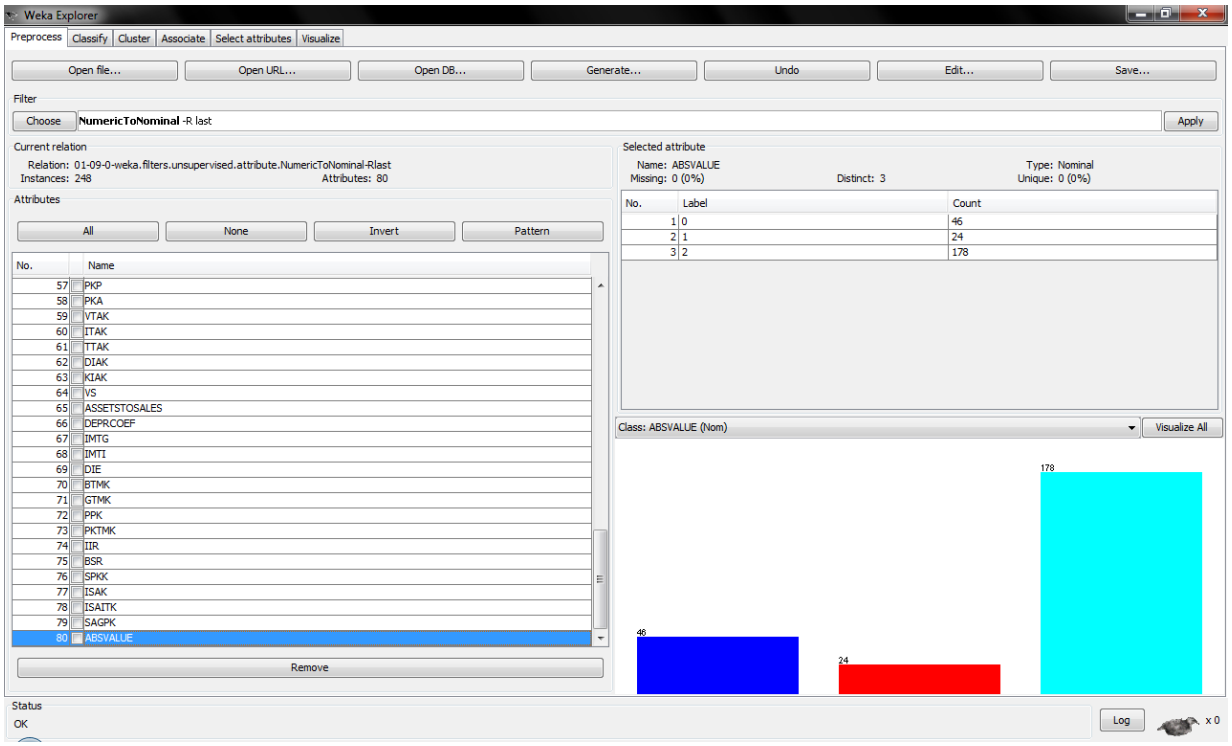
Šiame darbe naudojami tik teisingumo, tikslumo ir *recall* parametrai, tačiau realizacijoje taip pat skaičiuojama ir visa maišos matrica bei kiti rodikliai (26 pav.).



Šaltinis: sukurta autoriaus

26 pav. Sistemos išvedami rezultatai

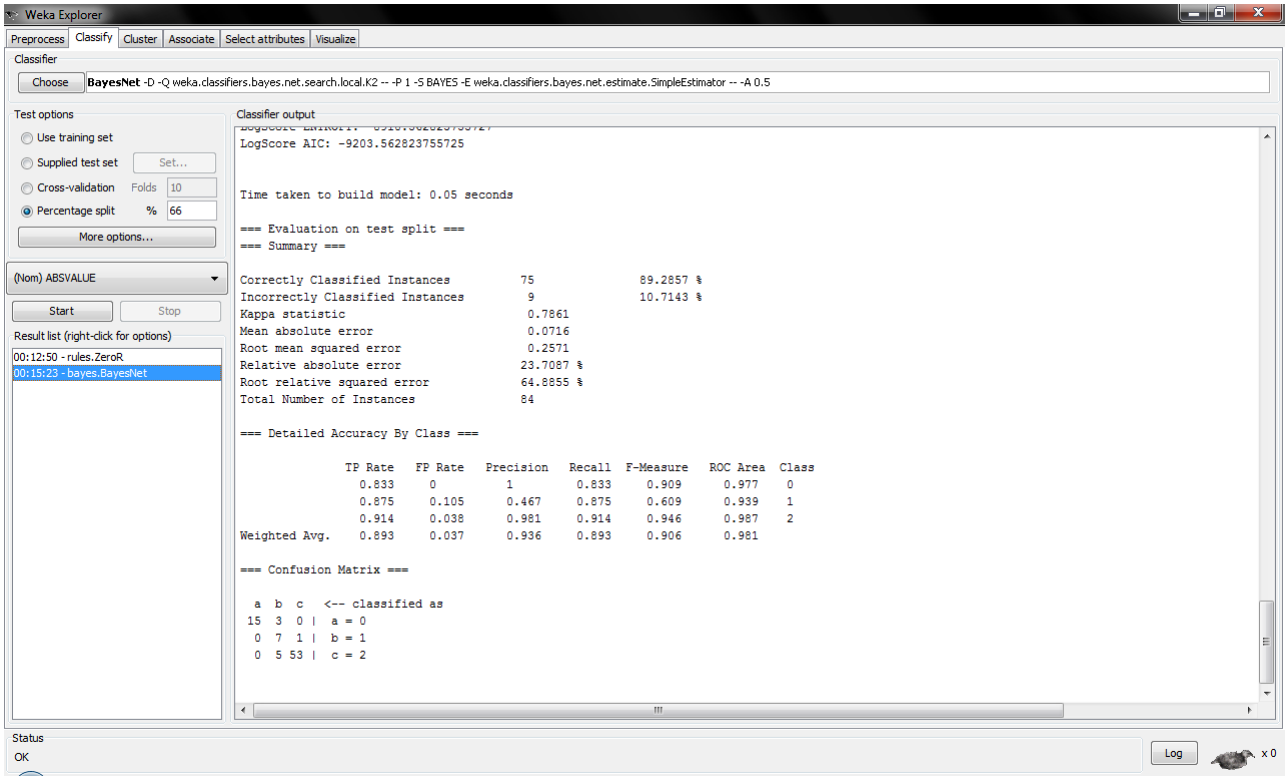
Visų pirma, paprastumo ir patogumo dėlei iš duomenų bazės buvo išeksportuoti visi duomenys į CSV formato failus, iš viso jų išeksportuota 18. Po to kiekvienas iš jų importuojamas į Weka aplinką bei pritaikomas filtras *NumericToNominal* klasės atributui, kuris būtinas naudojant bet kurį iš Bajeso metodų (27 pav.).



Šaltinis: sukurta autoriaus

27 pav. Duomenų užkrovimo ir filtravimo langas sistemoje

Suimportavus duomenis ir pritaikius filtrą, pasirenkamas klasifikatorius bei nustatomi atitinkami parametrai pasirinktam Bajeso metodui (28 pav.), t.y., kiekvienam iš 18 duomenų rinkinių išbandomi keturi minėti Bajeso metodai: BayesNet, Naive Bayes su reikšmėmis pagal nutylėjimą, su parametru *KernelEstimator* ir su *SupervisedDiscretization*. Taigi, iš viso buvo atlikti 72 bandymai, o jų rezultatas matomas 6 lentelėje.



Šaltinis: sukurta autoriaus

28 pav. Sistemos langas, kur parenkamas klasifikatorius, nurodomi parametrai bei pateikiami gauti rezultatai

Kiekvienam eksperimentiniam bandymui duomenys pagal Weka sistemos rekomendacijas buvo skaldyti santykiu 1:2, t.y., 2/3 duomenų skirta apsimokymui, o 1/3 testavimui.

6 lentelė

Eksperimento metu gautų rezultatų suvestinė lentelė

Duomenys	Visų duom. % mokymui	Mokymo įrašų skaičius	Testav.įrašų skaičius	Baysnet		Naïve Bayes		Naïve Bayes Kernel		Naïve Bayes discret.	
				Tiksluma s	Type1 klaida	Tiksluma s	Type1 klaida	Tiksluma s	Type1 klaida	Tiksluma s	Type1 klaida
Ketvirčių duomenys											
01-09	59.055	150	104	89.2875	0.037	70.2381	0.104	73.8095	0.075	89.2857	0.037
10-14	59.744	2054	1384	74.8828	0.063	39.5501	0.114	36.5511	0.115	74.8828	0.063
15-17	61.180	394	250	77.7778	0.107	60.6481	0.117	66.6667	0.115	77.3148	0.113
20-39	60.203	13972	9236	72.5337	0.153	31.6980	0.107	35.1494	0.115	72.4412	0.151
40-49	60.048	3526	2346	71.0779	0.126	40.6083	0.202	45.7844	0.193	70.8111	0.126
50-51	59.764	1319	888	80.8067	0.073	74.2698	0.210	74.4089	0.254	80.6676	0.075
52-59	60.185	1888	1249	81.7739	0.074	36.0624	0.145	38.3041	0.129	81.6764	0.072
60-67	61.003	7784	4976	80.7081	0.070	26.9509	0.148	26.5173	0.135	80.3468	0.071
70-89	60.306	7777	5119	77.1723	0.095	42.3428	0.265	45.6313	0.247	77.1723	0.096
Visi duomenys	**63.115**	**44179**	**25819**	**71.848**	**0.15 1**	**65.688**	**0.011**	**61.822**	**0.006**	**71.843**	**0.151**
Metiniai duomenys											
01-09	59.556	134	91	81.9444	0.126	66.6667	0.191	68.0556	0.158	81.9444	0.126
10-14	59.777	1660	1117	71.8104	0.077	32.3208	0.164	34.3864	0.163	71.9320	0.077
15-17	60.674	324	210	71.7791	0.147	51.5337	0.176	61.3497	0.166	71.7791	0.147
20-39	60.104	11641	7727	68.5059	0.155	30.6964	0.170	24.7299	0.173	68.4394	0.155
40-49	59.634	2838	1921	68.5956	0.133	55.8222	0.358	55.6810	0.371	68.4545	0.133
50-51	59.803	1095	736	77.1020	0.101	65.4741	0.294	64.4007	0.351	76.3864	0.107
52-59	60.199	1570	1038	81.7043	0.056	45.3634	0.121	71.9298	0.147	81.7043	0.056
60-67	61.181	6413	4069	78.5940	0.086	27.0741	0.139	28.5624	0.128	78.5307	0.086
70-89	60.224	6241	4122	74.0741	0.121	38.4665	0.309	39.8635	0.299	74.0741	0.122
Visi duomenys	**63.529**	**36674**	**21054**	**70.909**	**0.169**	**57.148**	**0.011**	**55.472**	**0.002**	**70.904**	**0.169**
Svorinis ketvirčių duomenų vidurkis (testavimo įrašų aspektu):				**75.907**	**0.110**	**36.282**	**0.161**	**38.605**	**0.158**	**75.765**	**0.110**
Svorinis metinių duomenų vidurkis (testavimo įrašų aspektu):				**72.776**	**0.122**	**36.204**	**0.210**	**36.049**	**0.211**	**72.708**	**0.122**

2.5. Gautų rezultatų apibendrinimas

Taigi, visa atlikto eksperimento suvestinė matosi 6 lentelėje. Duomenys, kaip jau buvo aptarta, suskaidyti į 9 sektorius, o taip pat į dvi grupes – pagal metinius ir ketvirčio duomenis, todėl atitinkamai skiriasi ir duomenų (angl., instances) kiekiai tam pačiam sektoriui, ketvirčio duomenų yra daugiau. Iš viso eksperimento metu atlikti ir pateikti 72 bandymai.

Pagal gautus rezultatus galima pastebėti, kad panašiausiai, atsižvelgiant tiek į tikslumą, tiek į *false positive* (FP) klaidas, klasifikuoja Naivus Bajesas su reikšmėm pagal nutylėjimą ir Naivus Bajesas su pasirinktu parametru *KernelEstimator*, ir analogiškai BayesNet su Naiviu Bajesu, kai parinktas parametras *SupervisedDiscretization*. False positive išskirtas todėl, kad jis svarbesnis kredito rizikos atveju, nes jis reiškia, kad kompanija, kuri turėtų būti priskirta rizikos grupei, bus priskirta prie nerizikingų arba vidutiniškai rizikingų, kas itin aktualu.

Atlikus tyrimą paaiškėjo, kad savo klasifikavimo tikslumu BayesNet ir Naivus bajesas su parinktu parametru *SupervisedDiscretization* klasifikuoja žymiai tiksliau, tai matoma tiek po ketvirčių, tiek po metinių duomenų klasifikavimo bei jų svorinio vidurkio, o be to, jų FP yra kur kas mažesni nei likusių dviejų klasifikatorių. Iš šių dviejų šiek tiek tiksliau klasifikuoja BayesNet klasifikatorius.

Tyrimo metu pastebėta, kad duomenų klasifikavimo efektyvumas kinta nuo duomenų kiekio: esant didesniems kiekiams duomenų BayesNet ir Naivus bajesas su parinktu parametru *SupervisedDiscretization* klasifikuoja tiksliau ir daro mažesnes FP klaidas, kai tuo tarpu likę du klasifikatoriai veikia atvirkščiai – pasirodė geriausiai tuomet, kai tų duomenų mažai ir itin mažai, pvz, kai jų buvo tik 10, jų tikslumas pašoko iki 100%, o kituose bandymuose kiti du klasifikatoriai pasirodė blogiau. Analogiškai visi iki vieno klasifikatoriai ketvirčio duomenis klasifikavo geriau ir klydo mažiau. Kadangi Naivus Bajesas su nutylėtom reikšmėm ir Naivus Bajesas su *KernelEstimator* geriau klasifikuoja prie mažesnių duomenų kiekių, buvo išbandyta ir pakeisti treniravimo ir testavimo duomenų santykį ir rezultatas toks, kad ženkliai sumažinus apmokymo duomenis šių klasifikatorių tikslumas pakilo apie 10 procentų.

Kadangi klasifikuojant geriausiai tikslumo atžvilgiu pasirodė BayesNet metodas, galima daryti prielaidas, jog su kitais paieškos algoritmais ir kitais parametrais, kaip *estimator* būtų galima išgauti geresnį klasifikavimo tikslumą.

2.6. Tolimesnis eksperimentinis tyrimas (straipsnis)

Atlikus minėtą eksperimentinį tyrimą buvo atkitas jo tęsinys, lyginant du geriau pasirodžiusius metodus, t.y., BayesNet ir Naive Bayes discr. su analogiškais dvejais SVM metodais: LibSVM ir LIBLINEAR [45]. Šio tyrimo smulkus aprašymas pateiktas 6 priede, kuris yra straipsnio pavidalu, pristatytu konferencijoje IVUS15 2010 m. gegužės 13 d.

Eksperimentas, kaip ir ankstesnis atliktas su tais pačiais duomenimis Weka aplinkoje, tačiau skiriasi duomenų paruošimas ir padalinimas apmokymui bei testavimui. Glaustai modelio eiga pateikiama taip:

1. Duomenys vertinami diskriminantiniu Altmano metodu priskiriant jiems rizikingumo klases;
2. Pašalinami "tušti" įrašai, kurie negali būti įvertinti 1 žingsnyje (pvz., dalyba iš 0);
3. Tušti atributai pakeičiami tos kompanijos atributo vidurkio reikšme;
4. Padalinamos kompanijos į dvi dalis – apmokymo ir testavimo duomenis: (kompanijų sąrašas $C = C\_train \cup C\_test$ bei $|C\_train| > |C\_test|$);
5. Paskaičiuojama kaip procentiškai dalinami duomenys į treniravimo ir testavimo (svarbu dėl skirtingų pagal diskriminantinį modelį įvertintų kompanijų įrašų skaičių);
6. Sukuriami treniravimo ir apmokymo duomenys pagal 5 žingsnį ($C_D = C_D\_train \cup C_D\_test$ ir $|C_D\_train| > |C_D\_test|$);
7. Naudojant genetinį algoritmą, atrenkami reikšmingiausi atributai;
8. Naudojant vieną iš klasifikavimo metodų, sukuriamas modelis, kuris yra testuojamas ir įvertinamas.

Gauti metiniai rezultatai pateikti žemiau esančiame paveiksle (29 pav.)

Sector	Training percentage, %	No of training instances	No of testing instances	SVM (LIBLINEAR)		SVM (LibSVM)		Naïve Bayes		BayesNet	
				Accuracy	Type1 Error	Accuracy	Type1 Error	Accuracy	Type1 Error	Accuracy	Type1 Error
01-09	59,556	134	91	85,71	0,3	75,82	0,561	83,52	0,105	84,62	0,079
10-14	59,777	1660	1117	71,17	0,612	77,17	0,408	77,35	0,174	77,17	0,174
15-17	60,674	324	210	84,76	0,534	72,38	0,596	77,14	0,091	75,24	0,109
20-39	60,104	11641	7727	88,92	0,523	90,86	0,336	68,10	0,097	68,12	0,098
40-49	59,634	2838	1921	70,85	0,666	81,05	0,294	69,91	0,123	69,91	0,124
50-51	59,803	1095	736	91,58	0,555	90,08	0,363	75,00	0,127	75,14	0,127
52-59	60,199	1570	1038	95,38	0,775	95,09	0,586	79,87	0,082	79,77	0,080
60-67	61,181	6413	4069	68,86	0,512	82,13	0,206	77,02	0,088	77,17	0,088
70-89	60,224	6241	4122	80,81	0,524	85,30	0,286	74,31	0,173	74,36	0,173
All data	**63,529**	**36674**	**21054**	**70,45**	**0,667**	**71,67**	**0,635**	**72,26**	**0,174**	**72,29**	**0,174**
Weighted mean				**81,21**	**0,552**	**86,39**	**0,318**	**72,68**	**0,117**	**72,70**	**0,117**

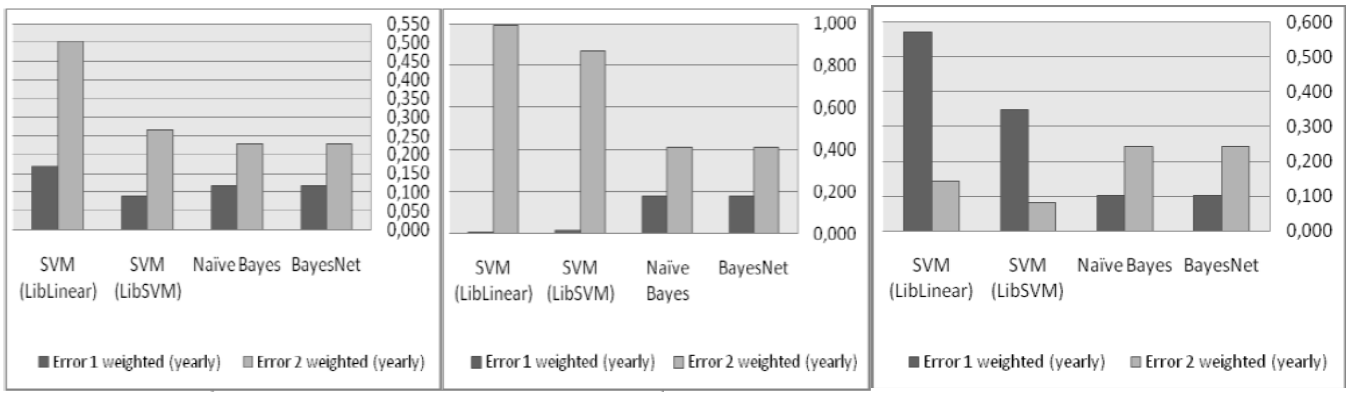
Šaltinis: sukurta autoriaus

29 pav. Eksperimento rezultatų suvestinė

Iš pateiktų rezultatų suvestinės galima pastebėti, kad, nors ir pasikeitė duomenų paruošimo principas, rezultatai labai panašūs į pirminio eksperimento gautus rezultatus.

Lyginant su SVM metodais, Bajeso metodai lyginant svertinius tikslumo vidurkius klasifikavo prasčiau ir geriau pasirodė tik ten, kur apmokymo ir testavimo duomenų buvo gerokai mažiau nei kitose pozicijose (Agriculture, Forestry, And Fishing; Mining; Construction) ir Finance, Insurance, And Real Estate, kur duomenų kiekis nebuvo sąlyginai mažas. Klasifikuojant visus duomenis, neskirstant juos į veiklos sektorius, Bajeso metodai klasifikavo iki 1.84% tiksliau.

Šiame darbe taip pat atkreiptas didesnis dėmesys į klaidas Type 1 ir Type 2, kurių rezultatai pateikti žemiau esančioje iliustracijoje (30 pav.), atitinkmai klasėms rizikinga, vidutinio rizikingumo ir nerizikinga.



Šaltinis: sukurta autoriaus

30 pav. Klaidos Type 1 ir Type 2

Pagal pateiktus rezultatus galima pastebėti, kad Bajeso metodai pasižymėjo stabiliais klaidų dydžiais, kai tuo tarpu SVM metodai turėjo didesnes klaidų reikšmes, tačiau jos pasižymėjo tuo, kad 2 atvejais (nerizikinga ir rizikinga klasėms) nerizikingą kompaniją priskyrė prie didesnės rizikos kompanijų, t.y., darė didesnę Type 2 klaidą, kuri yra mažiau svarbi nei Type 1, nes ji blogiausiu atveju įtakoja galimas prarastas pajamas, kai tuo tarpu Type 1 siejama su tiesioginiais nuostoliais prarandant skolininkui išduoto kredito dalį. Platesnė visų gautų rezultatų analizė ir aprašymas pateiktas priede.

IŠVADOS IR PERSPEKTYVOS

Darbo eigoje buvo atlikta:

1. Surinkus eilę literatūros šaltinių, susijusių su kreditų rizikos valdymu, dirbtiniu intelektu, įvairiais mašininio mokymo metodais, tinkančiais klasifikavimui, ši literatūra buvo susisteminta ir apžvelgta. Pateikti svarbiausi kreditų rizikos valdymo aspektai, siejami su šio darbo tematika, o taip pat pateikti metodų naudojamų kreditų rizikai valdyti aprašymai, aprašyti įvairūs lyginamieji metodų straipsniai, kur parodyta, jog Bajeso metodas tinka įvairiems klasifikavimo uždaviniams taikyti. Apibrėžtas Bajeso metodo veikimo principas ir jo panaudojimo atvejai kreditų rizikai vertinti.
2. Vertinant juridinio asmens finansinį pajėgumą, svarbu atsižvelgti ne tik į finansinius rodiklius, bet ir kitus aspektus, t.y., vertinant kredito riziką, svarbu įvertinti ir kitas rizikas. Kai kurios iš šių rizikų gali būti vertinamos ekspertiniu būdu, t.y., jų įvertinimui gali būti naudojami kokybiniai įverčiai; taip pat daugumoje jų naudojama techninė analizė bei matematiniai/ekonometriniai metodai. Tai sudaro prielaidas taikyti DI modelius ir ekspertines sistemas.
3. Apibrėžtas eksperimentinis modelis, naudojami įrankiai, metodai ir vykdoma eiga; Paruošti duomenys bandymams atlikti, o vėliau su Weka 3.6.2 sistema atlikti tyrimai su keturiais Bajeso klasifikatoriais;
4. Atlikus eksperimentą, gauti rezultatai rodo, kad galimas gana sėkmingas klasifikavimas, kur tam tikriems veiklos sektoriams pasiektas didesnis nei 89% tikslumas, tuo tarpu imant svorinius vidurkius du metodai rodė gana neblogus rezultatus, siekiančius beveik 76% tikslumą ketvirčių duomenims ir beveik 73% metiniams duomenims. Analitinės dalies išvadose iškelta tikslumo kartelė metodo tinkamumui vertinti (lyginant su kitų metodų rodomais rezultatais įvairiuose šaltiniuose), 70% šiuo atveju, yra pasiekta ir viršyta. Klaidų dydis neviršijo 0.25 ir vidurkis daugmaž 0.2, o tai lyginant su įvairiais metodais yra gana mažas klaidų dydis.
5. Geriausias Bajeso klasifikatorius nagrinėjamai problemai – BayesNet su iki 76% svoriniu vidurkiu, nuo kurio nežymiai (dešimtosiomis ar net šimtosiomis procento skirtingiems duomenims) atsiliko Naivus Bajesas su parinktu parametru *SupervisedDiscretization.*
6. Atliktas eksperimentinis tyrimas lyginant Bajeso metodus su SVM metodais nežymiai pasikeitusiom sąlygom, kaip duomenų paruošimas apmokymams ir testavimams, pasirinktas kitas paieškos algoritmas BayesNet metodui (vietoje K2 pasirinktas

HillClimbing). Tyrimo rezultatai parodė, kad Bajeso metodo klasifikavimo tikslumas ir klaidų dydis pakito labai nežymiai, tikslumas sumažėjo tik šimtosiomis procento dalimis.

7. Lyginant Bajeso ir SVM metodus, pastarasis parodė geresnius rezultatus 7 iš 9 veiklos sektorių (visus, išskyrus 10-14 ir 60-67 sektorius) ir todėl rodė klasifikavimo tikslesnius svorinius vidurkius. Neskirstant duomenų į veiklos sektorius, tikslesni buvo Bajeso metodai, rodantys iki 1.84% didesnį tikslumą.
8. Vertinant klaidų dydžius Type 1 ir Type 2, Bajeso metodai rodė panašius ir lyginant su SVM gana mažus klaidos dydžius, o taip pat jie visų klasių atžvilgiu rodė stabilesnį klaidų dydį.
9. Atlikus įvairių literatūros šaltinių analizę ir eksperimentinius tyrimus, galima teigti, kad Bajeso metodas (BaysNet ir Naivus Bajesas su pasirinktu parametru *SupervisedDiscretization*) gali būti sėkmingai naudojamas kreditų rizikos valdyme, kadangi rodo sąlyginai aukštus tikslumo ir gana mažus klaidos rodiklius lyginant su kitais šiuo metu naudojamais metodais (tiek matematiniais, tiek mašininio mokymo).
10. Ateities perspektyvoje geresniems klasifikavimo rezultatams išgauti būtų galima paieškoti įvairių parametrų optimizavimo (pvz.: apmokymo/testavimo duomenų santykis, vietoje procentinio dalinimo naudoti žingsninį, angl. fold), duomenų apdorojimo metodų (vietoje genetinės paieškos naudoti kitas reikšmingiems atributams atrinkti tinkančias sistemas), pritaikant įvairius duomenų filtrus duomenų įvedimo į sistemą etapuose (angl.: normalize, remove, discretize, cernelfilter ir t.t.) ar apjungiant keletą klasifikavimo metodų (pvz., Bajeso su ANN, Bajeso su SVM, kt.).

LITERATŪRA

1. Edward I. Altman. Financial Ratios, *Discriminant Analysis And The Prediction of Corporate Bancruptcy*. The Journal of Finance, Vol. 23, No. 4 (Sep., 1968), 589-609. Adresas Internete: http://www.defaultrisk.com/_pdf6j4/Financial_Ratios_Discriminant_Anlss_n_Prdctn_o_Crprt_Bnkrptc.pdf, Prieiga 2008.11.19
2. Kin Keung Lai, Lean Yu, Shouyang Wang, and Ligang Zhou. *Credit Risk Analysis Using a Reliability-Based Neural Network Ensemble Model*. 2006.09.16. Adresas Internete: http://madis1.iss.ac.cn/madis.files/pub-papers/2006/41320682.pdf. Prieiga 2009.01.15
3. Graham Kendall, *History of AI, Introduction to Artificial Intelligence*, The University of Nottingham within the School of Computer Science & IT, Adresas Internete: http://www.cs.nott.ac.uk/~gxk/courses/g5aiai/ Prieiga 2008.11.19
4. Rimvydas Simutis, *Dirbtiniai neuroniniai tinklai, Intelektinės sistemos finansų rinkose*, Vilniaus Universitetas, Kauno humanitarinis fakultetas, Informatikos katedra. Adresas Internete: ftp.vukhf.lt/aulay/Informatikos_katedra/Simutis(/IF_2008/ ir /KHF_ANN2009)(jungiantis iš išorės, reikia vartotojo vardo ir slaptažodžio), Prieiga 2009.06.03
5. E. Charniak, *Bayesian Networks without Tears*, AI magazine 1991. Adresas Internete: http://www.cs.ubc.ca/~murphyk/Bayes/Charniak_91.pdf. Prieiga 2008.11.19
6. D. Heckerman, *A tutorial on learning with Bayesian networks*, Microsoft Research tech. report, MSR-TR-95-06, 1996 Adresas Internete: ftp://ftp.research.microsoft.com/pub/tr/TR-95-06.PS Prieiga 2008.11.19
7. Ingrida Šarkiūnaitė, Dalia Krikščiūnienė, Rimvydas Simutis, *Magistro baigiamojo darbo rengimo tvarka, Metodiniai nurodymai* (VU KHF informatikos katedros verslo informatikos (62109P101) ir verslo informacijos sistemų (62103S138) studijų programų studentams), Kaunas, 2007
8. Kin Keung Lai, Lean Yu, Shouyang Wang, and Ligang Zhou. *Credit Risk Analysis Using a Reliability-Based Neural Network Ensemble Model*, 2006.09.16. Adresas Internete: http://madis1.iss.ac.cn/madis.files/pub-papers/2006/41320682.pdf. Prieiga prie Interneto 2008-12-01
9. *An Introduction to Bayesian Networks and their Contemporary Applications* [interaktyvus] Prieiga internetu: http://www.niedermayer.ca/papers/bayesian/bayes.html#table , [žiūrėta 2009.01.18]
10. *A Brief Introduction to Graphical Models and Bayesian Networks* [interaktyvus] Prieiga internetu: http://www.cs.ubc.ca/~murphyk/Bayes/bayes.html, [žiūrėta 2009.01.18]
11. Vytautas Valvonis, *Kredito rizikos valdymas banke*, Lietuvos bankas, Pinigų studijos, Apžvalginiai straipsniai, 2004, p. 57-82
12. Vytautas Valvonis, *Kreditų koncentracijos rizikos vertinimas ir valdymas*, ISSN 1392–1258. EKONOMIKA, 2007, p. 94-113
13. Edward I. Altman, Anthony Saunders, Cre*dit risk measurement: Developments over the last 20 years*, Journal of Banking & Finance 21, 1998, p. 1721-1742

14. Nijolė Žaltauskienė, Gražina Masionytė, *Finansinių institucijų paskolų kitimo tendencijos Lietuvoje*, atitinkamai Vytauto Didžiojo universitetas, Lietuvos žemės ūkio universitetas, 2008, p. 183-191, [interaktyvus] Adresas Internete: http://www.lzuu.lt/vadyb/lt/18196, Prieiga 2009.12.17

15. Saulius Nainys, *Kreditų monitoringas – kredito rizikos valdymo priemonė*, Lietuvos žemės ūkio universitetas, 2004, [interaktyvus] Adresas Internete: http://www.lzuu.lt/jaunasis_mokslininkas/smk_2006/menu_finansai.html, Prieiga 2009.12.17

16. Oesterreichische Nationalbank (OeNB), Austrian Financial Market Authority (FMA), *Credit Approval Process and Credit Risk Management*, Guidelines on Credit Risk Management, DVR 0031577, 2004

17. Virgilijus Sakalauskas, *Trumpalaikių investicijų rizikos tyrimas finansinėse rinkose*, ISSN 1392–0561. Informacijos mokslai, 2005

18. Virgilijus Sakalauskas, Dalia Kriksciuniene, Sho*rt-Term Investment Risk Measurement using VaR and CVaR*, Lecture notes in computer science, 2006, vol. 3994. p. 316-323. ISSN 0302-9743

19. Paulius Danėnas, *Dirbtinio intelekto metodų taikymas kredito rizikos vertinime*, magistrinis darbas, Vilniaus universiteto Kauno humanitarinis fakultetas, Informatikos katedra, 2008

20. Tom Mitchell, Bruce Buchanan, Gerald DeJong, Thomas Dietterich, Paul Rosenbloom, Alex Waibel, *Machine Learning*, Annual Reviews Inc., Computer Science, 1990.4:417-433, 1990, p. 417-433

21. Chih-Wei Hsu, Chih-Chung Chang, Chih-Jen Lin, *A Practical Guide to Support Vector Classification*, Department of Computer Science National Taiwan University, Taipei 106, Taiwan, 2009

22. Laisvoji internetinė enciklopedija Wikipedia, raktiniai žodžiai: „SVM“ [interaktyvus] Adresas Internete: http://en.wikipedia.org/wiki/SVM, Prieiga 2009.06.13

23. Tjen-Sien Lim, Wei-Yin Loh, Yu-Shan Shih, *A Comparison of Prediction Accuracy, Complexity, and Training Time of Thirty-Three Old and New Classification Algorithms*, Machine Learning, 40, 2000, p. 203–228

24. Rich Caruana, Alexandru Niculescu-Mizil, *An Empirical Comparison of Supervised Learning Algorithms*, Department of Computer Science, Cornell University, Ithaca, NY 14853 USA, Proceedings of the 23 rd International Conference on Machine Learning, Pittsburgh, PA, 2006

25. Philippe H. Gosselin, Matthieu Cord*, A Comparison of Active Classification Methods for Content-Based Image Retrieval*, France, ACM 1581139179/04/06, 2004

26. Nigel Williams, Sebastian Zander, Grenville Armitage, *A Preliminary Performance Comparison of Five Machine Learning Algorithms for Practical IP Traffic Flow Classification*, ACM SIGCOMM Computer Communication Review, Volume 36, Number 5, October 2006

27. Faming Liang, *An Effective Bayesian Neural Network Classifier with a Comparison Study to Support Vector Machine*, Neural Computation 15, 2003, p. 1959–1989

28. M. Kudra, H. Bohlig, E. Geide1, *Fuzzy and Probabilistic Interpretation of Spectral Information,* Proceedings of ISUMA-NAFIPS '95, 0-8186-7126-2/95, 1995, p. 662-667

29. Tom Mitchel, McGraw Hill, *Machine learning, Bayesian Learning, Chapter 6*, ISBN 0070428077, 1997, p. 154-211

30. David Heckermen, *A Tutorial on Learning With Bayesian Networks*, Microsoft Research, Advanced Technology Division, Microsoft Corporation, Technical Report, MSR-TR-95-06, 1996

31. Pagrindiniai KNN metodo principai ir realūs pavyzdžiai, [interaktyvus] Adresas Internete: http://www.cra.org/Activities/craw_archive/dmp/awards/2003/Mower/KNN.html, Prieiga 2009.06.22

32. Laisvoji internetinė enciklopedija Wikipedia, raktiniai žodžiai: „KNN“ [interaktyvus] Adresas Internete: http://en.wikipedia.org/wiki/K-nearest_neighbor_algorithm, Prieiga 2009.06.22

33. Laisvoji internetinė enciklopedija Wikipedia, raktiniai žodžiai: „Decision tree“ [interaktyvus] Adresas Internete: http://en.wikipedia.org/wiki/Decision_tree, Prieiga 2009.06.22

34. Weka: Practical Machine Learning Tools and Techniques with Java Implementations Ian H. Witten, Eibe Frank, Len Trigg, Mark Hall, Geoffrey Holmes, and Sally Jo Cunningham, Department of Computer Science, University of Waikato, New Zealand. 2000

35. Remco R. Bouckaert, *Bayesian Network Classifiers in Weka for Version 3-5-8*, University of Waikato, 2008

36. Ashraf M. Kibriya and Eibe Frank, *An Empirical Comparison of Exact Nearest Neighbour Algorithms*, Department of Computer Science, University of Waikato, 2008

37. Luc De Raedt, *Integrating Naive Bayes and FOIL*, Journal of Machine Learning Research 8 (2007) 481-507

38. John Wiley & Sons, Inc., *Naive Bayes Estimation And Bayesian Networks,* Data Mining Methods and Models By Daniel T. Larose, 2006.

39. Jorge Jambeiro Filho, Jacques Wainer, *HPB: A Model for Handling BN Nodes with High Cardinality Parents*, Journal of Machine Learning Research 9 (2008) 2141-2170

40. Jiang Su, Harry Zhang, *Full Bayesian Network Classifiers*, Faculty of Computer Science, University of New Brunswick, Canada, 2008

41. Remco R. Bouckaert, *Practical Bias Variance Decomposition*, University of Waikato, 2008

42. Stuart Moran, Yulan Hey, Kecheng Liu, *Choosing the Best Bayesian Classifier: An Empirical Study,* IAENG International Journal of Computer Science, 36:4, IJCS_36_4_09, 19 November 2009

43. E. Merkevicius, G. Garsva, and S. Girdzijauskas, "A hybrid SOM-Altman model for bankruptcy prediction," Lecture Notes in Computer Science, vol. 3994, 2006, pp. 364-371.

44. *Evaluation of models* (discovered knowledge) [interaktyvus] Adresas Internete: http://dms.irb.hr/tutorial/tut_mod_eval_1.php, Prieiga 2010.01.18

45. Gediminas Būzius, Paulius Danėnas, Gintautas Garšva*, Credit risk evaluation using SVM and Bayesian classifiers*, IVUS 15 conference, 2010

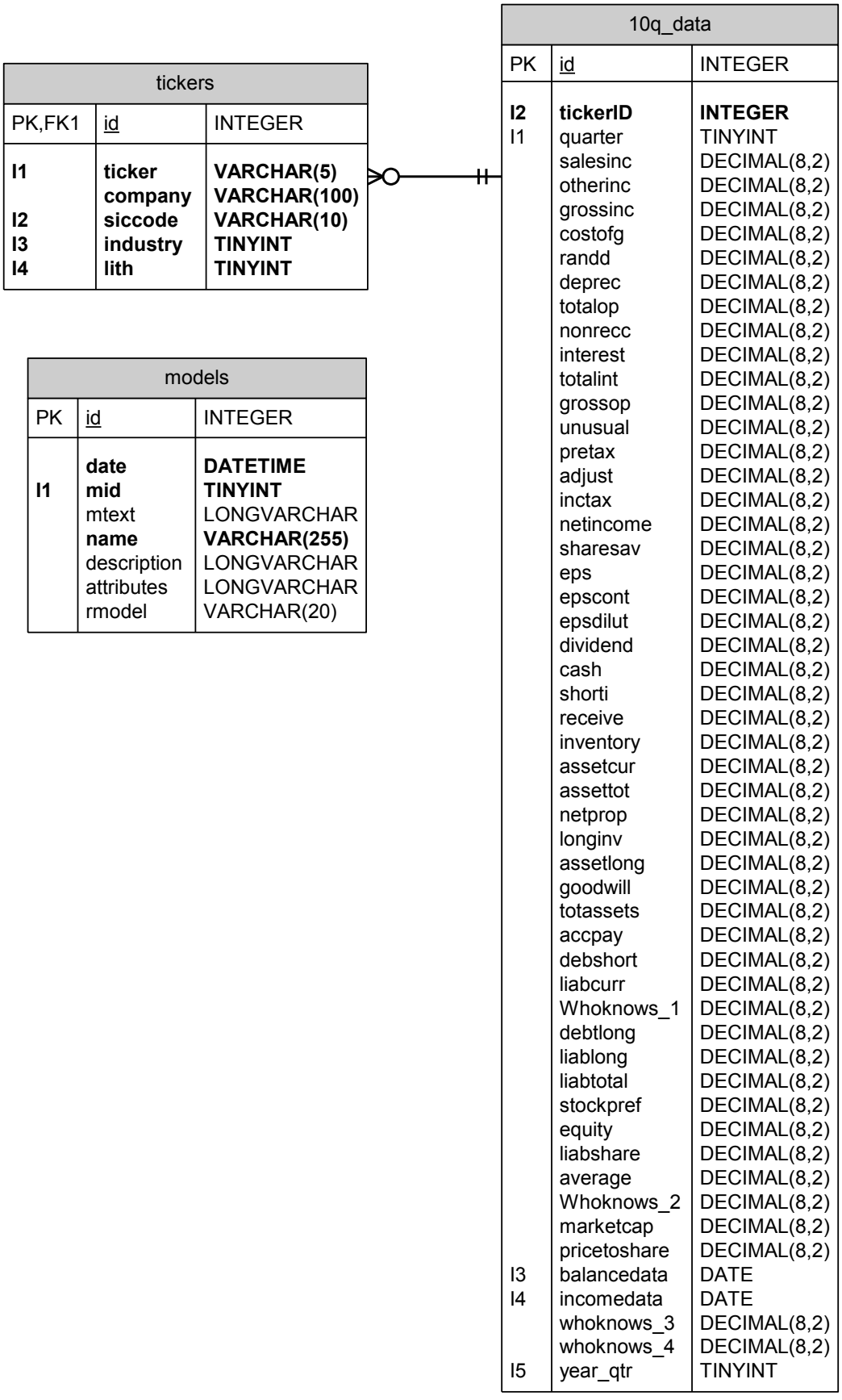
PRIEDAI

1 priedas. Sistemoje naudojamų duomenų aprašymas ir duomenų bazės schema

Ataskaita	**Santrumpa**	**Aprašymas**
Pelno ataskaita	salesinc	Pajamos iš pardavimų Sales income
	otherinc	Kitos pajamos Other income
	grossinc	Bendrosios įplaukos Gross income
	costofg	Parduotų prekių ar paslaugų kaina Cost of goods sold
	randd	Tyrimai ir plėtra Research and development
	deprec	Nuvertėjimas Depreciation
	totalop	Bendros veiklos išlaidos Total operation expenses
	nonrecc	Neperiodinės lėšos balanse Nonrecurring items
	interest	Išlaidos palūkanoms Interest expenses
	totalint	Bendros išlaidos palūkanoms Total interest expenses
	grossop	Bendrosios veiklos pajamos Gross operating expenses
	unusual	Netikėtos pajamos Unusual income
	pretax	Pajamos prieš mokesčius Pre-tax income
	adjust	Pajamų patikslinimai Adjustments to income
	inctax	Pajamų mokestis Ingome tax
	netincome	Grynasis pelnas Net income
	sharesav	Vidutinė akcijos kaina Shares average
	eps	EPS EPS
	epscont	Dabartinis (continued) EPS Continued EPS
	epsdilut	EPS Diluted DilutedEPS

Ataskaita	Santrumpa	Aprašymas
	dividend	Dividendai Dividend
Balansas	cash	Grynieji pinigai Cash
	shorti	Trumpalaikės investicijos Short-term investments
	receive	Debitorinės sąskaitos Receivables
	inventory	Prekių atsargos Inventory
	assetcur	Trumpalaikis turtas Other Current Assets
	assettot	Bendras dabartinis turtas Total Current Assets
	netprop	Grynoji gamybinių fondų nuosavybė Net property plant and equipment
	longinv	Ilgalaikės investicijos Long-term investments
	assetlong	Kiti ilgalaikiai aktyvai Other long-term assets
	goodwill	Prestižo vertė ir nematerialiosios vertybės Goodwill and intangibles
	totassets	Bendrieji aktyvai Total assets
	accpay	Tiekėjų įsiskolinimas Accounts payable
	debshort	Trumpalaikiai įsiskolinimai Short-term debt
	liabcurr	Kiti dabartiniai įsipareigojimai Other current liabilities
	debtlong	Ilgalaikiai įsiskolinimai Long-term debt
	liablong	Kiti ilgalaikiai įsipareigojimai Other long-term liabilities
	liabtotal	Bendri įsipareigojimai Total liabilities
	stockpref	Privilegijuotosios akcijos Stocks preferred
	equity	Bendrasis kapitalas Common equity
	liabshare	Bendri įsipareigojimai ir akcininkų kapitalas Total liabilities and shareholder's equity
	average	Vidutiniškai emituotų akcijų Average shares outstanding
	marketcap	Rinkos kapitalizacija Market capitalization
	pricetoshare	Kainos ir akcijų skaičiaus santykis periodo gale Price to share (end of period)
Nežinomi parametrai (tyrimuose	Whoknows_1	Whoknows_1
	Whoknows_2	Whoknows_2
	whoknows_3	Whoknows_3

Ataskaita	Santrumpa	Aprašymas
nenaudojami)	whoknows_4	Whoknows_4
Išvestiniai rodikliai	likvid	Trumpalaikis/likvidumo koeficientas Current ratio Trumpalaikis turtas / Trumpalaikiai įsipareigojimai assetcur/debshort
	liabinventory	Trumpalaikių įsipareigojimų ir inventoriaus santykis Current liabilities to inventory ratio Trumpalaikės skolos/inventorius liabcurr/inventory
	totalliab	Visi įsiskolinimai grynai vertei Total liabilities to net worth ratio Visi įsiskolinimai / gryna vertė Liabtotal/assetlong
	collection	Surenkimo periodo koeficientas Collection period ratio Gaunamos pajamos / pardavimai x 365 dienos salesinc/ costofg*365
	salesinventory	Pardavimų ir inventoriaus koeficientas Sales to inventory ratio Grynieji metiniai pardavimai /inventorius Sales/inventory
	assetssales	Turto ir pardavimų koeficientas Assets to sales ratio Bendras turtas / Grynieji pardavimai assettot/salesinc
	salescapital	Pardavimai ir grynasis įstatinis kapitalas Sales to net working capital Pardavimai / grynasis įstatinis kapitalas salesinc/equity
	accountssales	Mokėjimų ir pardavimų koeficientas Accounts payable to sales ratio Mokėjimai / Grynieji pardavimai accpay/salesinc
	quickratio	Greitas koeficientas Quick ratio (Gryni pinigai + gautinos_pajamos) / trumpalaikiai įsipareigojimai (cash+salesinc)/debshort
	ros	Pardavimų grąža Return on sales ratio Grynasis pelnas atskaičius mokesčius / Grynieji pardavimai netincome/salesinc
	roa	Turto grąža Return on assets (ROA) ratio Grynasis pelnas atskaičius mokesčius / Visas turtas netincome/assettot
	ronw	Grynosios vertės grąža Return on net worth ratio Grynasis pelnas atskaičius mokesčius /Grynoji vertė netincome/equity


tickers
PK,FK1 id INTEGER
I1 ticker VARCHAR(5)
company VARCHAR(100)
I2 siccode VARCHAR(10)
I3 industry TINYINT
I4 lith TINYINT
10q_data
PK id INTEGER
I2 tickerID INTEGER
I1 quarter TINYINT
salesinc DECIMAL(8,2)
otherinc DECIMAL(8,2)
grossinc DECIMAL(8,2)
costofg DECIMAL(8,2)
randd DECIMAL(8,2)
deprec DECIMAL(8,2)
totalop DECIMAL(8,2)
nonrecc DECIMAL(8,2)
interest DECIMAL(8,2)
totalint DECIMAL(8,2)
grossop DECIMAL(8,2)
unusual DECIMAL(8,2)
pretax DECIMAL(8,2)
adjust DECIMAL(8,2)
inctax DECIMAL(8,2)
netincome DECIMAL(8,2)
sharesav DECIMAL(8,2)
eps DECIMAL(8,2)
epscont DECIMAL(8,2)
epsdilut DECIMAL(8,2)
dividend DECIMAL(8,2)
cash DECIMAL(8,2)
shorti DECIMAL(8,2)
receive DECIMAL(8,2)
inventory DECIMAL(8,2)
assetcur DECIMAL(8,2)
assettot DECIMAL(8,2)
netprop DECIMAL(8,2)
longinv DECIMAL(8,2)
assetlong DECIMAL(8,2)
goodwill DECIMAL(8,2)
totassets DECIMAL(8,2)
accpay DECIMAL(8,2)
debshort DECIMAL(8,2)
liabcurr DECIMAL(8,2)
Whoknows_1 DECIMAL(8,2)
debtlong DECIMAL(8,2)
liablong DECIMAL(8,2)
liabtotal DECIMAL(8,2)
stockpref DECIMAL(8,2)
equity DECIMAL(8,2)
liabshare DECIMAL(8,2)
average DECIMAL(8,2)
Whoknows_2 DECIMAL(8,2)
marketcap DECIMAL(8,2)
pricetoshare DECIMAL(8,2)
I3 balancedata DATE
I4 incomedata DATE
whoknows_3 DECIMAL(8,2)
whoknows_4 DECIMAL(8,2)
I5 year_qtr TINYINT
models
PK id INTEGER
date DATETIME
I1 mid TINYINT
mtext LONGVARCHAR
name VARCHAR(255)
description LONGVARCHAR
attributes LONGVARCHAR
rmodel VARCHAR(20)

2 priedas. Sektoriui priklausančios rinkos

CODE	TITLE	DIVISION
01	Agricultural Production - Crops	01-09
02	Agricultural Production - Livestock and Animal Specialties	01-09
07	Agricultural Services	01-09
08	Forestry	01-09
09	Fishing, Hunting and Trapping	01-09
10	Metal Mining	10-14
12	Coal Mining	10-14
13	Oil and Gas Extraction	10-14
14	Mining and Quarrying of Nonmetallic Minerals, Except Fuels	10-14
15	Building Construction - General Contractors & Operative Builders	15-17
16	Heavy Construction, Except Building Construction - Contractors	15-17
17	Construction - Special Trade Contractors	15-17
20	Food and Kindred Products	20-39
21	Tobacco Products	20-39
22	Textile Mill Products	20-39
23	Apparel, Finished Products from Fabrics & Similar Materials	20-39
24	Lumber and Wood Products, Except Furniture	20-39
25	Furniture and Fixtures	20-39
26	Paper and Allied Products	20-39
27	Printing, Publishing and Allied Industries	20-39
28	Chemicals and Allied Products	20-39
29	Petroleum Refining and Related Industries	20-39
30	Rubber and Miscellaneous Plastic Products	20-39
31	Leather and Leather Products	20-39
32	Stone, Clay, Glass, and Concrete Products	20-39
33	Primary Metal Industries	20-39
34	Fabricated Metal Products, Except Machinery & Transport Equipment	20-39
35	Industrial and Commercial Machinery and Computer Equipment	20-39
36	Electronic, Electrical Equipment & Components, Except Computer Equipment	20-39
37	Transportation Equipment	20-39
38	Measure/Analyze/Control Instruments; Photo/Med/Opt Gds; Watches/Clocks	20-39
39	Miscellaneous Manufacturing Industries	20-39
40	Railroad Transportation	40-49
41	Local, Suburban Transit & Interurban Highway Passenger Transport	40-49
42	Motor Freight Transportation	40-49
43	United States Postal Service	40-49
44	Water Transportation	40-49
45	Transportation by Air	40-49
46	Pipelines, Except Natural Gas	40-49
47	Transportation Services	40-49
48	Communications	40-49
49	Electric, Gas and Sanitary Services	40-49
50	Wholesale Trade - Durable Goods	50-51
51	Wholesale Trade - Nondurable Goods	50-51
52	Building Materials, Hardware, Garden Supply & Mobile Home Dealers	52-59
53	General Merchandise Stores	52-59
54	Food Stores	52-59
55	Automotive Dealers and Gasoline Service Stations	52-59

56	Apparel and Accessory Stores	52-59
57	Home Furniture, Furnishings and Equipment Stores	52-59
58	Eating and Drinking Places	52-59
59	Miscellaneous Retail	52-59
60	Depository Institutions	60-67
61	Non-depository Credit Institutions	60-67
62	Security & Commodity Brokers, Dealers, Exchanges & Services	60-67
63	Insurance Carriers	60-67
64	Insurance Agents, Brokers and Service	60-67
65	Real Estate	60-67
67	Holding and Other Investment Offices	60-67
70	Hotels, Rooming Houses, Camps, and Other Lodging Places	70-89
72	Personal Services	70-89
73	Business Services	70-89
75	Automotive Repair, Services and Parking	70-89
76	Miscellaneous Repair Services	70-89
78	Motion Pictures	70-89
79	Amusement and Recreation Services	70-89
80	Health Services	70-89
81	Legal Services	70-89
83	Social Services	70-89
84	Museums, Art Galleries and Botanical and Zoological Gardens	70-89
86	Membership Organizations	70-89
87	Engineering, Accounting, Research, Management & Related Services	70-89
88	Private Households	70-89
89	Services, Not Elsewhere Classified	70-89
91	Executive, Legislative & General Government, Except Finance	91-99
92	Justice, Public Order and Safety	91-99
93	Public Finance, Taxation and Monetary Policy	91-99
94	Administration of Human Resource Programs	91-99
95	Administration of Environmental Quality and Housing Programs	91-99
96	Administration of Economic Programs	91-99
97	National Security and International Affairs	91-99
99	Nonclassifiable Establishments	91-99

3 priedas. Klasių priskyrimui atrinkti rodikliai naudojant genetinę paiešką

Sektorius	**Atrinkti rodikliai**
Ketvirčių duomenys	
Agriculture, Forestry, And Fishing	EPSCONT, SHORTI, LONGINV, GOODWILL, TOTASSETS, LIABCURR, EQUITY, PRICETOSHARE, CAPITALCONST, ACCOUNTSSALES, TA, VTAK, ITAK, IMTG, BTMK, PPK, PKTMK, BSR, SPKK, ISAK
Mining	INTEREST, ASSETTOT, STOCKPREF, EQUITY, LIABSHARE, PRICETOSHARE, CAPITALCONST, PP, ACCOUNTSSALES, SALESTOINVENTORY, PKA, ITAK, GTMK, PKTMK, IIR, BSR, SPKK, ISAK, ISAITK
Construction	SHARESAV, RECEIVE, INVENTORY, LONGINV, TOTASSETS, DEBTLONG, LIABLONG, EQUITY, PRICETOSHARE, PPS, TOTALLIAB, SALESTOCAPITAL, PKP, DIAK, DIE, PKTMK, IIR, BSR, SPKK, ISAK, SAGPK
Manufacturing	DEPREC, INTEREST, INCTAX, EPS, EPSCONT, EPSDILUT, DIVIDEND, ACCOUNTSSALES, LIABTOINVENTORY, QUICKRATIO, ITAK, TTAK, KIAK, ASSETSTOSALES, DIE, GTMK, PPK, IIR, BSR, SPKK, ISAK, ISAITK
Transportation, Communications, Electric, Gas, And Sanitary Services	SALESINC, NONRECC, UNUSUAL, INCTAX, EPSCONT, EPSDILUT, NETPROP, GOODWILL, LIABLONG, STOCKPREF, EQUITY, LIABSHARE, PRICETOSHARE, ASSETFIXED, TOTALLIAB, QUICKRATIO, SALESTOCAPITAL, PKP, PKA, ITAK, TTAK, DEPRCOEF, DIE, PKTMK, IIR, BSR, SPKK, ISAK
Wholesale Trade	OTHERINC, INTEREST, EPS, EPSCONT, INVENTORY, LONGINV, STOCKPREF, EQUITY, LIABSHARE, MARKETCAP, PP, ACCOUNTSSALES, LIABTOINVENTORY, PKP, PKA, VTAK, PPK, IIR, BSR, SPKK, ISAK
Retail Trade	GROSSINC, DEPREC, NONRECC, GROSSOP, ADJUST, INCTAX, SHARESAV, EPS, SHORTI, RECEIVE, INVENTORY, LIABCURR, EQUITY, LIABSHARE, PPS, PP, ACCOUNTSSALES, TOTALLIAB, QUICKRATIO, SALESTOCAPITAL, PKP, PKA, VTAK, ITAK, IMTG, PPK, PKTMK, IIR, BSR, SPKK, ISAK, SAGPK
Finance, Insurance, And Real Estate	OTHERINC, GROSSINC, INTEREST, INCTAX, DIVIDEND, CASH, SHORTI, RECEIVE, ASSETCUR, ASSETTOT, NETPROP, ASSETLONG, LIABTOTAL, MARKETCAP, LIQUIDITY, PPS, PP, LIABTOINVENTORY, QUICKRATIO, SALESTOCAPITAL, TA, PKP, PKA, ITAK, DIAK, KIAK, ASSETSTOSALES, DEPRCOEF, DIE, PKTMK, IIR, BSR
Services	OTHERINC, RANDD, INTEREST, EPS, LONGINV, TOTASSETS, DEBTLONG, STOCKPREF, EQUITY, MARKETCAP, PRICETOSHARE, ACCOUNTSSALES, LIABTOINVENTORY, TOTALLIAB, QUICKRATIO, SALESTOCAPITAL, PKP, VTAK, ITAK, DEPRCOEF, BTMK, GTMK, PPK, PKTMK, IIR, BSR, SPKK, ISAK, SAGPK
Metiniai duomenys	
Agriculture, Forestry, And Fishing	SALESINC, GROSSINC, TOTALOP, PRETAX, EPS, DIVIDEND, RECEIVE, NETPROP, TOTASSETS, ACCPAY, LIABSHARE, PRICETOSHARE, CAPITALCONST, PP, SALESTOCAPITAL, DIAK, VS, PPK, PKTMK, IIR, BSR
Mining	INTEREST, GROSSOP, ADJUST, INCTAX, EPS, EPSDILUT, LONGINV, EQUITY, PP, QUICKRATIO, SALESTOCAPITAL, PKP, PKA, VTAK, BTMK, PPK, PKTMK, IIR, BSR, SPKK, ISAK
Construction	COSTOFG, UNUSUAL, NETINCOME, EPS, EPSCONT, EPSDILUT, CASH, RECEIVE, NETPROP, TOTASSETS, DEBTLONG, PRICETOSHARE, PPS, PP, SALESTOCAPITAL, PKP, DIAK, VS, IMTG, PKTMK, BSR, SPKK, ISAK, SAGPK
Manufacturing	NONRECC, INTEREST, TOTALINT, PRETAX, INCTAX, EPS, EPSCONT, EPSDILUT, EQUITY, LIABSHARE, PRICETOSHARE, PP, ACCOUNTSSALES, QUICKRATIO, SALESTOCAPITAL, TA, PKP, VTAK, ITAK, TTAK, KIAK, PPK, SPKK, ISAK, ISAITK
Transportation, Communications, Electric,	RANDD, TOTALINT, GROSSOP, PRETAX, INCTAX, EPS, EQUITY, PRICETOSHARE, ASSETFIXED, PPS, PP, ACCOUNTSSALES, LIABTOINVENTORY,

Gas, And Sanitary Services	SALESTOINVENTORY, SALESTOCAPITAL, ITAK, PPK, PKTMK, IIR, BSR, SPKK, ISAK, ISAITK
Wholesale Trade	SALESINC, TOTALINT, GROSSOP, EPS, EPSCONT, EPSDILUT, DIVIDEND, ASSETCUR, STOCKPREF, EQUITY, AVERAGE, PRICETOSHARE, CAPITALCONST, PPS, PP, ACCOUNTSSALES, SALESTOCAPITAL, TA, PKP, GTMK, PKTMK, IIR, BSR, SPKK, ISAK, ISAITK, SAGPK
Retail Trade	COSTOFG, RANDD, PRETAX, INCTAX, EPS, EPSCONT, EPSDILUT, SHORTI, NETPROP, LONGINV, ASSETLONG, TOTASSETS, LIABSHARE, PRICETOSHARE, PP, ACCOUNTSSALES, PKP, PKA, VTAK, ITAK, DIAK, VS, DEPRCOEF, IMTG, PPK, IIR, BSR, SPKK, ISAK, ISAITK, SAGPK
Finance, Insurance, And Real Estate	OTHERINC, DEPREC, TOTALOP, NONRECC, DIVIDEND, CASH, RECEIVE, LONGINV, ACCPAY, LIABTOTAL, LIABSHARE, PRICETOSHARE, CAPITALCONST, PPS, PP, ACCOUNTSSALES, LIABTOINVENTORY, SALESTOCAPITAL, TA, PKP, PKA, ITAK, KIAK, DEPRCOEF, DIE, PKTMK, BSR, ISAK, SAGPK
Services	RANDD, TOTALINT, GROSSOP, PRETAX, INCTAX, EPS, EQUITY, PRICETOSHARE, ASSETFIXED, PPS, PP, ACCOUNTSSALES, LIABTOINVENTORY, SALESTOINVENTORY, SALESTOCAPITAL, ITAK, PPK, PKTMK, IIR, BSR, SPKK, ISAK, ISAITK

Šaltinis: P. Danėnas, VU KHF.

4 priedas. Mokslinis tiriamasis straipsnis „Credit risk evaluation using SVM and Bayesian Classifiers“ pristatytas IVUS15 konferencijoje

Credit risk evaluation using SVM and Bayesian classifiers

Gediminas Buzius, MSc Student
Kaunas Faculty of Humanities
Vilnius University
Muitines St. 8, Kaunas, Lithuania
gedbuz@gmail.com

Paulius Danenas, PhD Student
Kaunas Faculty of Humanities
Vilnius University
Muitines St. 8, Kaunas, Lithuania
paulius.danenas@vukhf.lt

Gintautas Garsva, Assoc. Prof.
Kaunas Faculty of Humanities
Vilnius University
Muitines St. 8, Kaunas, Lithuania
gintautas.garsva@vukhf.lt

Abstract— **This article presents a method combining popular machine learning technique for classification, genetic search as a feature selection method for relevant attribute selection and Altman Z-Score discriminant technique for credit risk evaluation. Support Vector Machines implementations (LIBLINEAR, LibSVM) and Bayesian method based classifiers (Naïve Bayes, Bayesian Networks) were explored and used in this article to train classifiers. This method was applied to different sectors in service and industry. Its performance was evaluated using weighted mean accuracy and weighted mean error techniques.**

Keywords- Support Vector Machines, SVM, Bayes, Naïve Bayes, Bayesian Networks, machine learning, credit risk, evaluation, bankruptcy, forecasting

I. INTRODUCTION

Credit risk evaluation and bankruptcy prediction are one of the most important problems in finance because they are directly related to possibility of losing money given as a credit to a customer. Many researchers are trying to construct methods that are based on various data mining and machine learning techniques such as Neural Networks (NN), Genetic Algorithms (GA), Swarm Intelligence (SI) and others as they show promising results. Support Vector Machines (SVM) is widely researched and applied in various industrial solutions and can be used as an effective solution to many various classification, regression and forecasting problems. Bayesian classifiers are one of oldest classification techniques based on Bayesian inference; they are widely applied for solution to many problems, including credit risk. These methods are often applied as standalone or complementary techniques.

II. RELATED WORK

Many statistical and data mining techniques were used for classifier creation in various studies. Early researches were based on discriminant analysis which was used by Altman [1], Springate [2]. Novel techniques such as Neural Networks (NN) were applied later for this problem together with other natural computing methods, mostly evolutionary techniques for their optimization [3][4][5] to solve classification or clustering tasks with self-organizing maps commonly referred to as Kohonen maps [6][7][8].

There are numerous examples where Naïve Bayes or Bayesian Networks (further referred as BayesNet) classifiers are used to evaluate credit risk. It was shown in [9] that Bayesian classifier may successfully increase prediction from 2 to 9% including the cases where some data are missing compared to traditional discriminant or logit analysis. Ranjan et. al. [10] used Bayesian method to estimate probability of default while [11] used Bayesian method for credit risk scoring combining it together with genetic algorithms. Antonakis et. al. [12] compared Naïve Bayes method with five other classification methods using real banks' data. Baesens et. al. [12] integrated Bayesian network together with Markov Chain Monte Carlo search. Gossl [14] used Bayesian approach to the modeling of credit risky portfolios by forming Bayesian portfolio model which allows describing default frequencies and intra-portfolio correlations.

Support Vector Machines (abbr. as SVM) has been proved as an efficient technique possible to obtain results comparable to Neural Networks and is often used to solve classification or forecasting tasks in finance and other fields. They were also successfully applied for company failure prediction [15], bankruptcy analysis [16], to estimate probability of default [17], to study credit rating systems [18], capital risk assessment [19]. Lai et. al. applied SVM using ensemble learning approach to identify high-risk customers in customer relationship management [20]. Van Gestel et. al. used SVM and Bayesian combination to develop Bayesian evidence framework [21] [22]. Many previous researches based on SVM and related to credit risk evaluation were analyzed in [23]; it showed that best results were achieved by using optimization techniques such as genetic algorithms or swarm intelligence and fuzzy logic together with SVM.

III. RESEARCH METHOD

A. *Description of classification algorithms used in this experiment*

Naïve Bayes and Bayesian Networks. These classifiers perform the classification in a way that minimizes decision risk and therefore some optimum decision variables are derived by which the input-signal space is mapped into the decision space with the most distant classes. Probability distribution for Naïve Bayes is calculated using

with the most distant classes. Probability distribution for Naïve Bayes is calculated using

$$P(X,y) = \text{?}(y)P(X \mid y) = \text{?}(y)\prod_{d=} X_d \mid parent(X_d)) \quad (1)$$

where X_d is the same as X_i above [24].

The main difference between these two methods is that Naïve Bayes method assumes attributes of case being unrelated among them and so is called "naive" while Bayesian Networks are directional acyclic networks (graphs) that contain nodes representing attributes and directional arcs (causal relations) between nodes where nodes are variables which have a finite number of states. Every node A_k given parent nodes $B_1,...,B_n$ has a distribution table of probabilities $P(A_k| B_1; ...;B_n)$ which contains information about joint probability distribution of all possible attribute's states when all its parent nodes are known. These tables are used to predict the probability of class to any given case. Probability distribution is calculated using this formula:

$$P(X_1,..., X_n) = \prod_{i=1}^{n} P(X_i \mid parent(X_i)) \quad (2)$$

where n the number of states of the attribute, X_i is X state of attribute and $parent(X_i)$ is state's X_i parent node [25]. The complexity increases with the number of classes and attributes. BayesNet also has an ability to use different search algorithms for finding the best decision as K2, Hill Climbing, Genetic search, etc. [32]

C-SVC (LibSVM). The main idea in SVM is identification of special data points (*support vectors*) that are used to separate the provided cases by defining the binary class boundaries. Generally SVM uses a linear model to implement nonlinear class boundaries by nonlinear mapping of the input vectors x_i into the high-dimensional feature space by using the kernel function $K(x_i, x_j) \equiv$ .A linear model giving maximum separation, maximum margin hyperplane, is constructed in the new space. Support vectors are the training vectors closest to this hyperplane. C-SVC is a classical implementation of SVM first described in [26]. It is formulated as follows [27]: given training vectors $X_i \in$ = , two classes and a vector $y \in$ such as $y_i \in -$ C-SVC solves the following primal problem

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

subject to?

$$y_i(w^T \phi\ \text{?}_i) + \text{'}) \geq -\zeta, \zeta \geq \text{'}), i = \text{,..,} N \quad (3)$$

A detailed description of this algorithm can be found in [27][28].

LIBLINEAR. LIBLINEAR is a family of linear SVM classifiers for large sparse data with a huge number of instances and features. It supports logistic regression and linear support vector machines and can be very efficient for training large-scale problems. In our experiment we apply L2-loss linear support vector machines for classification task. It solves the following optimization problem; a detailed description can be found in [29]:

$$\min_{w} \frac{1}{2} w^T w + \text{?}\sum_{i=}^{} \text{ix}(0,1 - \text{'}_i w^T x_i))^2 \quad (4)$$

where training vectors x_i ? R^n, $i = 1,..,l$ in two class, vector **y** ?

R^l such that $y_i = \{1,? 1\}$, **w** is a weight vector that a linear classifier generates as the model. The decision function is $\text{sgn}(\mathbf{w}^T\mathbf{x})$.

Genetic algorithm (GA). It is the most widely used evolutionary computing technique which is often applied to solve various optimization problems. Application of GA leads to evolutionary process as the solutions generated from one population are used to form a new population hoping that this one would be better. The better solutions are obtained with this population the more chances they have for recombination. This method uses concepts of population initialization, crossover, mutation, fitness selection similar to these concepts in biology to search for the best solutions in the search space by taking best individuals and forming new populations until a particular limit is reached. There are many various sources of information which can give more information on this technique, i.e. [30]

A. Proposed method

This section describes a method based on genetic search, machine learning technique for classification and discriminant analysis. The main steps are as follows:

1. Results of every instance are calculated using discriminant analysis and are converted to bankruptcy classes;
2. Data preprocessing:
   a. Instances with empty outputs (records, which couldn't be evaluated in Step 1 because of lack of data or division by zero) are eliminated;
   b. Data imputation to eliminate empty values; here missing values are replaced with company's average of the attribute with missing data (for dataset D with attributes X and its subset D_C as financial records (instances) related to company C, if $D_{Cij} = \{\}$ then $D_{Cij} = average(X_i), i=1..,m, j=1,..,n;$ here m is the number of attributes, n – length of D_C);
   c. Companies are divided to the ones whose data will be used for training and whose data will be used for testing. These sets are disjoint (for company list $C =$

      C_train ? C_test, and $|C\_train| > |C\_test|$);

   d. Calculate training and testing data split percentage;
   e. Create training and testing data by splitting data of selected companies in the sector by a percentage

a. selected companies in the sector by a percentage calculated in Step 5. These sets are also disjoint (C_D $= C_{D\_train}$? C_{D_test} and $|C_{D\_train}| > |C_{D\_test}|$);

2. A genetic algorithm for feature selection is used to select the most relevant ratios. Evaluation uses correlation-based feature subset selection based on consideration of individual predictive ability of each feature along with the degree of redundancy between them [32] [33]
3. Model is trained using one of machine learning algorithms for classification (SVM and Bayesian classifiers are selected for this experiment in order to compare their performance);
4. The created model is tested using testing (holdout) data and results are evaluated.

A. Methods for evaluation of instances and results

Altman's Z-Score was chosen for evaluation and forming of dependant variable as a popular and widely used evaluation technique. It predicts whether or not a company is likely to enter into bankruptcy within one or two years. Companies from Manufacturing sector (20-39) were evaluated using original Altman Z-Score, other – by using Z-Score for non-manufacturing companies [31]. Z-Score for non-manufacturing companies was used for full dataset evaluation. These models are presented in Table I.

TABLE I. ALTMAN MODELS USED IN EXPERIMENTS

	Altman (original)	**Altman for non-manufacturing companies**
w_1	1.2	6.56
x_1	Working capital/Total assets	Working capital/Total assets
w_2	1.4	3.26
x_2	Retained earnings/Total assets	Retained earnings/Total assets
w_3	3.3	6.72
x_3	Earnings before interest and taxes/ Total assets	Earnings before interest and taxes/Total assets
w_4	0.6	1.05
x_4	Book value of Equity/ Book value of total liabilities	Book value of Equity/ Book value of total liabilities
w_5	0.999	-
x_5	Net sales/Total assets	-
Eval	Z>3 – healthy; 2.7<Z<2.99 – non-bankrupt; Z <1.79 - bankrupt	

Algorithms described in section 3.1 were used in this experiment to train models. The test results are evaluated by using accuracy together with Type 1 and Type 2 errors. If one would define define companies that have high or low value of risk evaluation in terms of "bad" and "good". Type I error will show the accuracy with which the model didn't identify "bad" debtors, and Type II error will show the accuracy with which the model identified "good" debtors as bad. Type 1 error is considered as more important in credit risk evaluation as the inability to identify a failing company will cost a lender far more than rejecting a healthy company.

Weighted mean was used as following to properly evaluate overall performance:

$$\bar{e} = \frac{\sum_{i=1}^{n} r_i e_i}{\sum_{i=1}^{n} r_i} \quad (5)$$

where n is the number of sectors, e_i is the value of error for sector i, r_i – the number of records in sector i used for testing. It should be noted that the here r is the number of test records; it was chosen because the test results represent the accuracy better. Here we evaluated only the weighted mean according to amounts of testing instances; however, if the proportion of training and testing instances varies significantly, it might be useful to evaluate weighted mean with both training and testing instances.

II. THE EXPERIMENT

A. Research data

The experiments were made by using data from EDGAR database of over 8600 companies from the year 1999-2006. It consists of yearly financial records with 79 financial ratios and indices used in financial analysis. This data was split into sectors according to SIC classification. We used 3:2 split in our experiment (60% percent of companies of each sector were selected for training). The sectors and company splits are given in Table II.

TABLE II. THE SECTORS USED IN EXPERIMENTS

Sector code	**Sector name**	**Total no. of companies**	**No. of companies for training**	**No. of companies for testing**
01-09	Agriculture, Forestry, And Fishing	33	20	13
10-14	Mining	469	281	188
15-17	Construction	83	50	33
20-39	Manufacturing	3027	1816	1211
40-49	Transportation, Communications, Electric, Gas, And Sanitary Services	786	472	314
50-51	Wholesale Trade	287	172	115
52-59	Retail Trade	405	243	162
60-67	Finance, Insurance, And Real Estate	1853	1112	741
70-89	Services	1712	1027	685
All data		8665	5199	3466

B. Experiment configuration

The experiment was run using implementations of SVM and Bayesian classifiers in described in Section 3.1 in Weka [32] framework. Weka integrates both LibSVM [27] and LIBLINEAR [29] as its SVM implementations. The algorithm in Section 3.2 was applied for every sector.

Genetic search used for attribute selection was run using the following parameters: crossover probability equal to 0.6, number of generations equal to 20, mutation probability of 0.033 and population size of 20. Polynomial kernel K(x,z) $=(?x? z? + \theta)^d$ was used with SMO and LibSVM (C-SVC) classifiers, using parameters C = 4 for SMO; C = 7 and $\gamma = 7$ for C-SVC. LIBLINEAR was run with C = 20. Naïve Bayes

for C-SVC. LIBLINEAR was run with C = 20. Naïve Bayes classifier was run using supervised discretization. BayesNet was used with the following parameters: estimator – SimpleEstimator algorithm for finding the conditional probability tables of the Bayesian Network with $\alpha = 0.5$. Hill climbing was used as search algorithm for adding, deleting and reversing arcs as the search is not restricted by an order on the variables. These parameters were selected experimentally.

A. Experiment results

The results of these experiments are presented in Table III, which includes weighted accuracies together with weighted Type 1 errors. LibSVM performed with highest accuracy as well as with highest weighted accuracy but it had higher weighted Type 1 error than Bayesian classifiers. Very similar results by both weighted accuracy and weighted error were obtained by using Naïve Bayes and BayesNet classifiers; although accuracy in most cases (especially in cases of higher number of instances for training) they obtained significantly lower results than SVM, but weighted errors and weighted mean errors were the smallest.

TABLE I. EXPERIMENT RESULTS

Sector	Training percentage, %	No of training instances	No of testing instances	SVM (LIBLINEAR)		SVM (LibSVM)		Naïve Bayes		BayesNet	
				Accuracy	Typel Error	Accuracy	Typel Error	Accuracy	Typel Error	Accuracy	Typel Error
01-09	59,556	134	91	85,71	0,3	75,82	0,561	83,52	0,105	84,62	0,079
10-14	59,777	1660	1117	71,17	0,612	77,17	0,408	77,35	0,174	77,17	0,174
15-17	60,674	324	210	84,76	0,534	72,38	0,596	77,14	0,091	75,24	0,109
20-39	60,104	11641	7727	88,92	0,523	90,86	0,336	68,10	0,097	68,12	0,098
40-49	59,634	2838	1921	70,85	0,666	81,05	0,294	69,91	0,123	69,91	0,124
50-51	59,803	1095	736	91,58	0,555	90,08	0,363	75,00	0,127	75,14	0,127
52-59	60,199	1570	1038	95,38	0,775	95,09	0,586	79,87	0,082	79,77	0,080
60-67	61,181	6413	4069	68,86	0,512	82,13	0,206	77,02	0,088	77,17	0,088
70-89	60,224	6241	4122	80,81	0,524	85,30	0,286	74,31	0,173	74,36	0,173
All data	**63,529**	**36674**	**21054**	**70,45**	**0,667**	**71,67**	**0,635**	**72,26**	**0,174**	**72,29**	**0,174**
Weighted mean				**81,21**	**0,552**	**86,39**	**0,318**	**72,68**	**0,117**	**72,70**	**0,117**

For more accurate evaluation Type 1 and Type 2 errors for separate classes (“bankrupt”, “average” and “healthy”) were also evaluated. Table IV and Table V represent these results; they show that Type 1 Error was relatively small for “bankrupt” companies in cases of SVM based classifiers, except Finance, Insurance and Real Estate sector where it was very high (0.811). Naïve Bayesian and BayesNet performed with errors under 0.1; similar result was also in the Finance, Insurance and Real Estate sector.

TABLE II. TYPE 1 ERROR VALUES FOR DIFFERENT CLASSES (“BANKRUPT” (B), ”AVERAGE” (A), ”HEALTHY” (H))

	LIBLINEAR			LibSVM			Naïve Bayes			BayesNet		
	B	A	H	B	A	H	B	A	H	B	A	H
01-09	0,026	0,036	0,381	0,064	0,024	0,714	0,105	0,036	0,182	0,079	0,048	0,182
10-14	0,012	0,000	0,889	0,050	0,006	0,580	0,174	0,081	0,048	0,174	0,081	0,054
15-17	0,010	0,051	0,625	0,021	0,164	0,688	0,091	0,155	0,090	0,109	0,155	0,100
20-39	0,015	0,000	0,618	0,033	0,001	0,394	0,097	0,235	0,136	0,098	0,234	0,136
40-49	0,008	0,002	0,948	0,032	0,051	0,401	0,123	0,215	0,092	0,124	0,214	0,091
50-51	0,009	0,003	0,628	0,031	0,024	0,407	0,127	0,160	0,050	0,127	0,158	0,050
52-59	0,003	0,000	0,818	0,007	0,010	0,618	0,082	0,136	0,052	0,080	0,137	0,055
60-67	0,811	0,000	0,012	0,303	0,030	0,052	0,088	0,148	0,092	0,088	0,147	0,090
70-89	0,014	0,000	0,702	0,059	0,005	0,369	0,173	0,118	0,095	0,173	0,118	0,094
All data	0,001	0,000	0,962	0,003	0,000	0,914	0,174	0,153	0,064	0,174	0,153	0,064

Table III shows that in most of the sectors LIBLINEAR and LibSVM performed with smaller Type 1 errors for “bankrupt” and “average” companies than Naïve Bayes and BayesNet; however, the latter two showed better results for “healthy” companies’ recognition. Usage of Bayesian method based resulted in much more balanced Type 1 and Type 2 errors. These results indicate that SVM usage might effectively predict “bankrupt” and “average” companies but Bayesian classifiers might be more suitable for general classification.

Table V shows that the possibility to identify “healthy” debtors as bankrupt is significantly high for SVM based classifiers compared to the possibility to select “bankrupt” company as “healthy”; conversely, the case of Finance, Insurance and Financial Services showed a high possibility to identify ”healthy” companies rather than “bankrupt”.

Respectively, Bayesian classifiers performed with this error respectively below 0.25 for both “bankrupt” and “healthy” company incorrect identification.

Insurance and Financial Services showed a high possibility to identify "healthy" companies rather than "bankrupt".

Respectively, Bayesian classifiers performed with this error respectively below 0.25 for both "bankrupt" and "healthy" company incorrect identification.

As in Type 1 error cases, their results were more even; the smallest errors were in cases of Mining and Transportation, Communications, Electric, Gas and Sanitary Services sectors.

For overall evaluation weighted mean errors of all these three classes were calculated; they are given in Figure 1. They illustrate that SVM based classifiers (especially LibSVM) might better identify "bankrupt" and "average" companies, but "healthy" companies might be better recognized by Bayesian based classifiers.

TABLE I. TYPE 2 ERROR VALUES FOR DIFFERENT CLASSES ("BANKRUPT" (B), "AVERAGE" (A), "HEALTHY" (H))

	LIBLINEAR			LibSVM			Naïve Bayes			BayesNet		
	B	A	H	B	A	H	B	A	H	B	A	H
01-09	0,154	0,875	0,057	0,615	1,000	0,086	0,000	0,857	0,130	0,000	0,857	0,116
10-14	0,822	1,000	0,010	0,376	0,966	0,047	0,175	0,474	0,217	0,179	0,487	0,217
15-17	0,647	0,800	0,051	0,882	0,800	0,174	0,422	0,255	0,136	0,422	0,309	0,145
20-39	0,530	1,000	0,012	0,290	1,000	0,024	0,269	0,430	0,247	0,269	0,429	0,248
40-49	0,899	0,991	0,011	0,335	0,594	0,065	0,164	0,295	0,377	0,164	0,298	0,376
50-51	0,517	1,000	0,008	0,400	0,962	0,037	0,249	0,308	0,235	0,243	0,308	0,235
52-59	0,756	1,000	0,003	0,561	1,000	0,014	0,288	0,258	0,166	0,283	0,266	0,168
60-67	0,005	1,000	0,695	0,041	0,606	0,262	0,208	0,360	0,204	0,206	0,360	0,203
70-89	0,606	0,993	0,012	0,252	0,934	0,046	0,195	0,517	0,244	0,195	0,511	0,245
All data	0,949	1,000	0,001	0,886	1,000	0,003	0,203	0,416	0,291	0,202	0,416	0,291

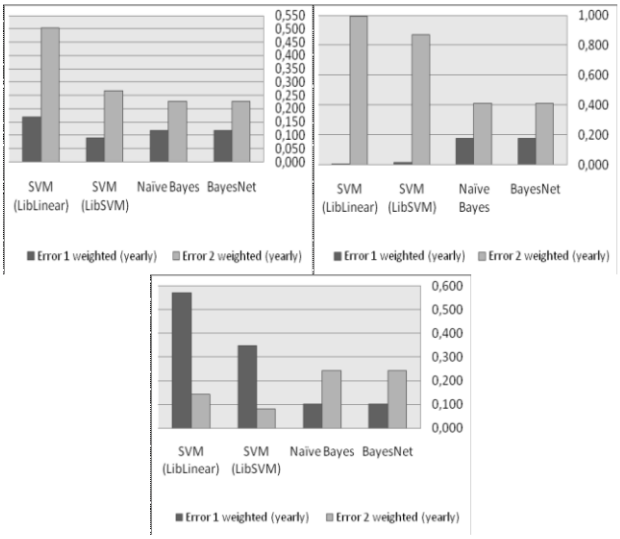


Figure 1. Weighted mean error values for all classes (bankrupt, average, healthy)

However, weighted mean does not precisely reflect the situation because of the "outstanding" Finance, Insurance and Real Estate sector thus weighted mean errors excluding this sector were also calculated. These values, together with "original" (including errors of Finance, Insurance and Real Estate sector) are given in Table VI. It shows that after excluding the "outlying" Finance sector weighed mean Type 1 error in case of "bankrupt" class is reduced from 0.167 to 0.013 for LIBLINEAR; LibSVM error is reduced from 0.09 to 0.039 which means that LIBLINEAR classifier might be a good choice for "bankrupt" company identification. However, weighted mean Type 1 error significantly increased for "healthy" class for the latter classifiers; this exclusion didn't affect Bayesian-based classifiers as their performance was almost even for all sectors. After excluding Finance sector results Weighted Type 1 Error increased for SVM based classifiers but these results didn't change significantly for Bayesian classifiers.

Proposed approach of "slicing" data to sectors and usual training using full dataset with all data was also compared to do overall evaluation; the results of classifiers' training using all data are presented in Table III and Table IV. They show that weighted accuracy achieved while training as separate classifier for each sector after division to sectors was significantly higher; Type 1 errors also differ. Table 4 shows that "bankrupt" companies were identified with Type 1 Error in range 0.001-0.003 by using SVM classifiers; Bayesian based methods performed with Type 1 error of 0.174 respectively.

TABLE II. WEIGHTED MEAN ERROR VALUES FOR DIFFERENT CLASSES ("BANKRUPT"(B), "AVERAGE"(A), "HEALTHY"(H))

	LIBLINEAR			LibSVM			Naïve Bayes			BayesNet		
	B	A	H	B	A	H	B	A	H	B	A	H
Including Finance, Insurance and Real Estate sector (60-67) results												
Error 1 weighted	0,167	0,001	0,571	0,090	0,015	0,349	0,117	0,176	0,103	0,117	0,176	0,103
Error 2 weighted	0,503	0,995	0,144	0,268	0,869	0,081	0,229	0,411	0,242	0,229	0,411	0,242
Excluding Finance, Insurance and Real Estate sector (60-67) results												
Error 1 weighted	0,013	0,001	0,705	0,039	0,012	0,421	0,124	0,183	0,106	0,124	0,182	0,106
Error 2 weighted	0,622	0,994	0,012	0,322	0,932	0,038	0,234	0,423	0,252	0,233	0,423	0,252

I. CONCLUSIONS AND FURTHER RESEARCH

This article presents a research on credit risk evaluation that uses evaluation by widely applied discriminant analysis models, such as Z-Score, genetic search for feature selection to select best financial ratios, and model training using popular machine learning techniques. SVM implementations (LIBLINEAR, C-SVC from LibSVM) and methods which use Bayesian inference (Naïve Bayes, BayesNet) were used to train classifiers. The experiment results show that LibSVM (C-SVC) in many cases outperformed other classifiers; however, LIBLINEAR performed best while identifying particularly "bankrupt" companies. The experiment results show that in most of the cases this method gives promising results for "bankrupt" and "average" company identification. Naïve Bayes and BayesNet performed with more balanced errors, together with the smallest weighted Type 1 errors. Further research might be concentrated on optimizing these classifiers by best parameter selection or integration of other machine learning techniques.

REFERENCES

[1] E. Altman, "Financial ratios, discriminant analysis and the prediction of corporate bankruptcy," The Journal of Finance, vol. 23, 1968, pp. 589–609.

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

[3] Y. Wu and Z. Si, "Application of RBF Neural Network Based on Ant Colony Algorithm in Credit Risk Evaluation of Construction Enterprises," 2008 International Conference on Risk Management & Engineering Management, 2008, pp. 653-658.

[4] F. Huang, "A Particle Swarm Optimized Fuzzy Neural Network for Credit Risk Evaluation," Genetic and Evolutionary Computing. 2008 Second International Conference on, 2008, pp. 153-157.

[5] E. Lacerda, A.C. Carvalho, A.n. Braga, and T.B. Ludermir, "Evolutionary Radial Basis Functions for Credit Assessment," Applied Intelligence, vol. 22, 2005, pp. 167-181.

[6] E. Merkevicius, G. Garsva, and R. Simutis, "Neuro-discriminate Model for the Forecasting of Changes of Companies Financial Standings on the Basis of Self-organizing Maps," Lecture Notes In Computer Science, Y. Shi, G. Albada, J. Dongarra, and P. Sloot, Berlin, Heidelberg: Springer-Verlag, 2007, pp. 439-446.

[7] E. Merkevicius, G. Garsva, and S. Girdzijauskas, "A hybrid SOM-Altman model for bankruptcy prediction," Lecture Notes in Computer Science, vol. 3994, 2006, pp. 364-371.

[8] Y. Xiao, "Research on comparison of credit risk evaluation models based on SOM and LVQ neural network," 2008 7th World Congress on Intelligent Control and Automation, 2008, pp. 2230-2235.

[9] P.N. Posch, G. Loffler, C. Schone, "Bayesian Methods for Improving Credit Scoring Models," Finance, 2005, pp. 1-26.

[10] S. Ranjan, R. Fan, G. Geng, "Bayesian Migration in Credit Ratings Based on Probabilities of Default". The Journal of Fixed Income, Vol. 12, No. 3, 2002, pp. 17-23.

[11] T.C. Fogarty, N.S. Ireson, "Evolving Bayesian classifiers for credit control-a comparison with other machine-learning methods", IMA J Management Math, vol. 5, no. 1, 1993, pp. 63-75.

[12] A. C. Antonakis, M. E. Sfakianakis, "Assessing naive Bayesian as a method for screening credit applicants". Journal of Applied Statistics, vol. 36, Issue 5, 2009, pp. 537 - 545

[13] B. Baesens, M. Egmont-Petersen, R. Castelo, J. Vanthienen, "Learning Bayesian Network Classifiers for Credit Scoring Using Markov Chain Monte Carlo Search," ICPR, 16th International Conference on Pattern Recognition, vol. 3, 2002, pp. 49-52

[14] C. Gössl, "Predictions Based on Certain Uncertainties—A Bayesian Credit Portfolio Approach," HypoVereinsbank AG, London, 2005, preprint.

[15] Z.R. Yang, "Support vector machines for company failure prediction," 2003 IEEE International Conference on Computational Intelligence for Financial Engineering Proceedings, 2003, pp. 47-54.

[16] W. Hardle, R. Moro, D. Schafer, "Bankruptcy Analysis with Support Vector Machines", European Finance Association, 33rd Annual Meeting, 2006, pp. 1-20.

[17] W. Hardle, R. Moro, D. Schafer, "Estimating probabilities of default with support vector machines", SFB 649 Discussion Papers SFB649DP2007-035, Sonderforschungsbereich 649, Humboldt University, Berlin, Germany, 2007.

[18] W. Chen and J. Shih, "A study of Taiwan's issuer credit rating systems using support vector machines," Expert Systems with Applications, vol. 30, 2006, pp. 427-435.

[19] W. Chong, G. Yingjian, and W. Dong, "Study on Capital Risk Assessment Model of Real Estate Enterprises Based on Support Vector Machines and Fuzzy Integral," Control and Decision Conference, 2008, pp. 2317 - 2320.

[20] K. Lai, L. Yu, S. Wang, and W. Huang, "An Intelligent CRM System for Identifying High-Risk Customers: An Ensemble Data Mining Approach," Lecture Notes in Computer Science, vol. 4488, 2007, pp. 486-489.

[21] T. Van Gestel, B. Baesens, J. Suykens, D. Van Den Poel, D. Baestaens, and M. Willekens, "Bayesian kernel based classification for financial distress detection," European Journal of Operational Research, vol. 172, 2006, pp. 979–1003.

[22] T. Van Gestel, J. Suykens, G. Lanckriet, A. Lambrechts, B. De Moor, and J. Vandewalle, "Bayesian framework for least-squares support vector machine classifiers, gaussian processes, and kernel Fisher discriminant analysis.," Neural computation, vol. 14, 2002, pp. 1115-1147.

[23] P. Danenas, G. Garsva, "Support Vector Machines and their Application in Credit Risk Evaluation", Transformations in Business & Economics, vol. 8, No. 3 (18), 2009, pp. 46-58

[24] I. Rish, "An empirical study of the naive Bayes classifier", IBM Research Report, IBM Research Division, 2001, pp. 1-2

[25] F. Ruggeri, F. Faltin, R. Kenett, "Encyclopedia of Statistics in Quality & Reliability", Wiley & Sons, 2007.

[26] C. Cortes and V. Vapnik, "Support-vector networks," Machine learning, vol. 20, 1995, pp. 273–297.

[27] C. Chang and C. Lin, "LIBSVM: a library for support vector machines", 2001.

[28] V. Vapnik, "The Nature of Statistical Learning Theory", Springer-Verlag, 2000

[29] R. Fan, K. Chang, C. Hsieh, X. Wang, and C. Lin, "LIBLINEAR: A library for large linear classification," The Journal of Machine Learning Research, vol. 9, 2008, pp. 1871–1874.

[30] S. E. Haupt, R. L. Haupt, "Practical genetic algorithms", Wiley-IEEE, 2004

[31] E.I. Altman, "Predicting financial distress of companies: Revisiting the Z-score and Zeta models," 2000, unpublished manuscript.

[32] Weka 3: Data Mining Software in Java, http://www.cs.waikato.ac.nz/ml/weka/

[33] M.A. Hall, "Correlation-based Feature Subset Selection for Machine Learning", Hamilton, New Zealand, 1998.