VILNIAUS UNIVERSITETAS

OLEGAS NIAKŠU

DUOMENŲ TYRYBOS METODŲ, SKIRTŲ MEDICININEI DIAGNOSTIKAI IR SVEIKATOS APSAUGOS VADYBAI, VYSTYMAS IR TAIKYMAS

Daktaro disertacijos santrauka

Technologijos mokslai, informatikos inžinerija (07 T)

Vilnius, 2015

Disertacija rengta 2009–2014 metais Vilniaus universiteto Matematikos ir informatikos institute.

Mokslinė vadovė – doc. dr. Olga Kurasova (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Disertacija ginama Vilniaus universiteto Informatikos inžinerijos mokslo krypties taryboje:

Pirmininkas – prof. dr. Albertas Čaplinskas (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Nariai:

prof. dr. Rimantas Jankauskas (Vilniaus universitetas, biomedicinos mokslai, medicina – 06 B),

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

dr. Marta Sabou (Vienos technikos universitetas, technologijos mokslai, informatikos inžinerija – 07 T),

prof. habil. dr. Laimutis Telksnys (Vilniaus universitetas, technologijos mokslai, informatikos inžinerija – 07 T).

Disertacija bus ginama Vilniaus universiteto viešame Informatikos inžinerijos mokslo krypties tarybos posėdyje 2015 m. rugsėjo 21 d. 13 val. Vilniaus universiteto Matematikos ir informatikos instituto 203 auditorijoje.

Adresas: Akademijos g. 4, LT-08663 Vilnius, Lietuva.

Disertacijos santrauka išsiuntinėta 2015 m. rugpjūčio mėn. 20 d.

Disertaciją galima peržiūrėti Vilniaus universiteto bibliotekoje ir VU interneto svetainėje adresu: www.vu.lt/lt/naujienos/ivykiu-kalendorius.

VILNIUS UNIVERSITY

OLEGAS NIAKŠU

DEVELOPMENT AND APPLICATION OF DATA MINING METHODS IN MEDICAL DIAGNOSTICS AND HEALTHCARE MANAGEMENT

Summary of Doctoral Dissertation
Technological Sciences, Informatics Engineering (07 T)

Vilnius, 2015

Doctoral dissertation was prepared at the Institute of Mathematics and Informatics of Vilnius University in 2009–2014.

Scientific supervisor – assoc. prof. dr. Olga Kurasova (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

The dissertation will be defended at the Council of the Scientific Field of Informatics Engineering of Vilnius University:

Chairman – prof. Dr. Albertas Čaplinskas (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

Members:

Prof. Dr. Rimantas Jankauskas (Vilnius University, Biomedical Sciences, Medicine – 06 B),

Prof. Dr. Dalius Navakauskas (Vilniaus Gediminas Technical University, Technological Sciences, Informatics Engineering – 07 T),

Dr. Marta Sabou (Vienna Technical University, Technological Sciences, Informatics Engineering – 07 T),

Prof. Habil. Dr. Laimutis Telksnys (Vilnius University, Technological Sciences, Informatics Engineering – 07 T).

The dissertation will be defended at the public meeting of the Scientific Council of Science of Informatics Engineering in the auditorium number 203 at the Institute of Mathematics and Informatics of Vilnius University, at 1 p.m. on 21st of September 2015.

Address: Akademijos g. 4, LT–08663 Vilnius, Lithuania.

The summary of the doctoral dissertation was distributed on 20th of August 2015.

A copy of the doctoral dissertation is available for review at the Library of the Vilnius University or on this website: www.vu.lt/lt/naujienos/ivykiu-kalendorius.

Trumpiniai, naudojami šioje disertacijos santraukoje

AS	Aortos vožtuvo stenozė
AV	Aortos vožtuvas
BRCA	Krūties vėžio jautrumo genas (angl. *Breast Cancer Susceptibility Gene*)
CRISP-DM	Tarpdisciplininė duomenų tyrybos taikymo metodika (angl. *Cross-Industry Standard Process for the Data Mining*)
CRISP-MED-DM	Duomenų tyrybos taikymo medicinoje metodika (angl. *Cross-Industry Standard Process for the Medical Data Mining*)
DICOM	Standartas, skirtas struktūrizuoti, saugoti, perduoti ir atvaizduoti medicininius vaizdus ir susijusią informaciją (angl. *Digital Imaging and Communications in Medicine*)
DT	Duomenų tyryba
ER diagrama	Esybių-ryšių diagrama
KSIT	Kairiojo skilvelio išstūmimo traktas
MESH	Medicininių terminų antraštės, kurios tvarkomos žodyno, naudojamo indeksuoti, kataloguoti ir ieškoti biomedicininių bei su sveikata susijusių dokumentų ar informacijos (angl. *Medical Subject Headings*)
PACS	Radiologinių vaizdų archyvavimo ir perdavimo sistema (angl. *Picture Achiving and Communication System*)
PAM	Klasterizavimo metodas „grupavimas aplink medoidus“ (angl. *Partitining Around Medoids*).
PMML	Prognozavimo modelio aprašomoji kalba (angl. *Predictive Model Markup Language*)
ROC, ROC AUC	Grafiko kreivė, atvaizduojanti jautrumo ir specifiškumo sąryšį (angl. *Rceiver Operating Characteristic*) ir plotas po ROC kreive (angl. *Area Under Curve*)
TED	Medžio struktūros redagavimo atstumo metrika (angl. *Tree Edit Distance*)
XML	Bendros paskirties duomenų struktūrų bei jų turinio aprašomoji kalba (angl. *Extensible Markup Language*)
FURIA	Nesurūšiuotų *fuzzy* sprendimo taisyklių algoritmas, RIPPER algoritmo plėtinys (angl. *Fuzzy Unordered Rule Induction Algorithm*)
Penn II	*BRCA1* ir *BRCA2* genų mutacijų rizikos įvertinimo modelis

1 Įvadas

1.1 Tyrimų sritis

Duomenų analizė medicinoje pasižymi ontologiniu sudėtingumu, medicininių duomenų standartų įvairove ir kintama duomenų kokybe (Bodenreider, 2008; Cios ir Moore, 2002; Chen et al., 2006; Esfandiari et al., 2014). Visa tai, kartu su paciento duomenų privatumo problemomis, kelia veiksmingų ir praktiškai panaudojamų medicininių žinių gavybos klausimą, kuris pastaraisiais dešimtmečiais vis dar lieka atviras. Šiuolaikinėje medicinoje stebimi ne tik diagnostikos ir gydymo metodų, bet ir sveikatos bei ligos sampratų pokyčiai, pereinama nuo į ligą orientuoto problemų sprendimo prie į pacientą orientuoto požiūrio, kur svarbų vaidmenį atlieka automatizuoti žinių gavimo metodai (Rudnick, 2004).

Duomenų tyrybos metodai ir priemonės yra taikomi įvairiose srityse daugiau nei 40 metų. R. D. Wilson ir kt. (Wilson et al., 2004) surinko ir klasifikavo medicinines publikacijas nuo 1966 iki 2002 metų, kuriose buvo taikomi arba tiriami žinių gavybos ir duomenų tyrybos metodai. XX a. daugelis šalių numatė e. sveikatos sistemų sukūrimą kaip prioritetinę nacionalinę programą, kuri iš esmės siūlo pagerinti sveikatos apsaugos sistemą, standartizuojant sveikatos apsaugos paslaugas, kaupiant klinikinę pacientų informaciją ir suteikiant prieigą prie šios informacijos tiek sveikatos apsaugos specialistams, tiek ir patiems pacientams (Castro, 2009; Stroetmann, Artmann, Stroetmann, & Whitehouse, 2011). ES valstybių narių, JAV ir kitų šalių strateginiuose planuose numatytos gausios investicijos sukurti globalią kompiuterizuotą sveikatos apsaugos duomenų sistemą. Atsižvelgiant į pastarojo dešimtmečio skaitmenizuojamos medicininės informacijos augimo tempą (National Center for Biotechnology Information, 2009), galima teigti, kad išsivysčiusiose šalyse per ateinančius 10 metų visa pacientų medicininė informacija bus kaupiama elektroninėje terpėje. Pirmą kartą istorijoje mokslinių tyrimų bendruomenė gaus pilną asmens sveikatos istoriją. Šis scenarijus prognozuoja didžiulį duomenų tyrybos taikymo sveikatos apsaugos srityje potencialą.

1.2 Problemos aktualumas

Spartus sveikatos apsaugos industrijos kompiuterizavimas pateikė didžiulį kiekį tiek struktūrizuotų, tiek ir nestruktūrizuotų heterogeninių duomenų, prieinamų tyrimams ir antriniam naudojimui. Yra įgyvendinta šimtai algoritmų, skirtų klasifikuoti, klasterizuoti ir rasti duomenyse paslėptus dėsningumus.

Tačiau, kaip teigiama darbuose (Cios ir Moore, 2002; Bellazzi ir Zupan, 2008; Špečkauskienė ir Lukoševičius, 2009), tam, kad būtų sėkmingai taikomi duomenų tyrybos metodai, turi būti išspręstos specifinės sveikatos apsaugos sričiai būdingos problemos. Pasak minėtų autorių, siekiant pritaikyti duomenų tyrybos metodus klinikiniams duomenims, turi būti sprendžiamos problemos, susijusios su paciento privatumu, semantine tarpusavio sąveika, įvairialyčiais duomenų šaltiniais ir nestruktūrizuotais duomenimis, pateiktais teksto, vaizdo ar garso formatais. Trūksta metodikos, kuri padėtų išspręsti šias medicinos sričiai būdingas problemas. Duomenų tyrybos specialistų bendruomenės *KDNuggets* tyrimai (Piatetsky-Shapiro, 2014), atlikti 2009 ir 2014 metais, atskleidžia, kad dažniausiai naudojama duomenų tyrybos metodika yra *Cross-Industry Standard Process for the Data Mining* (CRISP-DM). Tačiau CRISP-DM metodika dėl savo ribotumo nėra pritaikyta naudoti medicinoje (Azevedo ir Lourenco, 2008). Be to, Lietuvos, Šveicarijos, Vokietijos, Pietų Afrikos Respublikos ir Albanijos universitetinėse

ligoninėse atlikta apklausa (Niakšu ir Kurasova, 2012) atskleidė, kad dažnai duomenų tyrybos taikymo projektai lieka teoriniai, nėra tęsiami atliekant klinikinius tyrimus ir retai kada peržengia tiesiogiai su jais susijusios institucijos ribas.

Todėl reikalinga nauja metodika, apibrėžianti duomenų tyrybos proceso modelį, apimantį medicinos srities problemų sprendimą.

1.3 Darbo tikslas ir uždaviniai

Disertacijos tikslas – sukurti ir įvertinti duomenų tyrybos taikymo metodiką, skirtą medicinos ir sveikatos apsaugos sritims.

Siekiant šio tikslo, buvo iškelti ir sprendžiami šie uždaviniai:

1. Ištirti egzistuojančias duomenų tyrybos taikymo metodikas, traktuojant duomenų tyrybą kaip žinių išgavimo proceso sudėtinę dalį.
2. Pasiūlyti medicinos dalykinei sričiai skirtą taikymo metodiką, kuri pašalintų kliūtis naudoti esamas duomenų tyrybos metodikas.
3. Įvertinti pasiūlytą metodiką keliose medicinos srityse, sukuriant diagnostinius modelius bei reikalingus medicininių duomenų (medicininių vaizdų, multireliacinių duomenų) apdorojimo metodus.
4. Pasiūlyti ir įgyvendinti klasterizavimo būdą, tinkantį multireliaciniams duomenims klasterizuoti.

1.4 Darbo rezultatų praktinė reikšmė

Atlikto tiriamojo darbo rezultatų praktinė reikšmė yra ši:

1. Sukurta duomenų tyrybos taikymo metodika CRISP-MED-DM palengvina duomenų tyrybos vykdymą medicinos srityje, pasiūlydama pavyzdinį procesų modelį ir metodikos atitikties įvertinimo metodą.
2. Sukurtas *BRCA1* geno mutacijos prognostinis modelis gali būti naudojamas kaip sprendimo priėmimo paramos įrankis, identifikuojant *BRCA1* mutacijos tikimybę prieš atliekant brangų genetinį tyrimą.
3. Sukurta kardiologinių echokardiogramų vaizdų apdorojimo ir savybių išgavimo metodika ir ją realizuojanti programinė įranga leidžia automatizuoti darbui imlią sistolės ciklo trasavimo procedūrą, kurią rankiniu būdu atlieka kardiologai. Pasiūlytas aortos vožtuvo stenozės laipsnio nustatymo prognostinis modelis gali būti naudojamas kaip sprendimo priėmimo paramos įrankis.
4. Pasiūlyta ir įgyvendinta atstumo metrika gali būti taikoma multireliacinių duomenų klasterizavimo uždaviniams spręsti, kai duomenų struktūros supaprastinimas lemtų informacijos praradimą. Sukurta programinė įranga, kuri atlieka multireliacinių duomenų atstumų matricos skaičiavimus.

1.5 Tyrimo metodai

Siekiant apžvelgti ir sintezuoti kitų tyrimų rezultatus, buvo naudojami apžvalginio tyrimo bei sisteminės analizės metodai. Disertacijos iškeltiems uždaviniams spręsti buvo taikomi statistinės analizės, operacijų tyrimo, duomenų tyrybos ir vaizdų apdorojimo metodai. Pasiūlytiems vaizdų apdorojimo metodams ir duomenų apdorojimo algoritmams įvertinti ir palyginti su alternatyviais sprendimais buvo naudojami eksperimentinio tyrimo metodai.

1.6 Ginamieji teiginiai

1. Duomenų tyrybos taikymo metodika CRISP-DM gali būti specializuota ir praplėsta tokiu būdu, kad pagerintų duomenų tyrybos rezultatus medicinoje.
2. Pasiūlytos CRISP-MED-DM metodikos taikymas pagerina *BRCA1* geno mutacijos prognostinio modelio tikslumą.
3. CRISP-MED-DM metodika taikant echokardiogramų vaizdų apdorojimo metodus leidžia sukurti aukšto tikslumo aortos vožtuvo stenozės prognostinį modelį, kuris gali būti naudojamas stenozės laipsniui prognozuoti.
4. Skaidymo klasterizavimo metodai su multireliacine atstumo metrika yra tikslesni multireliacinėje aplinkoje, kai duomenų struktūros denormalizavimas lemia informacijos praradimą.

1.7 Pasiūlyti sprendimai ir mokslinis naujumas

- Sukurta duomenų tyrybos taikymo medicinoje metodika CRISP-MED-DM, kurioje numatytos veiklos specifiniams medicininės srities uždaviniams spręsti. CRISP-MED-DM taikymas pagerino *BRCA1* geno mutacijos rizikos modelio bendrą tikslumą nuo 0,88 iki 0,94, jautrumą nuo 0,67 iki 0,83.
- Sukurta kardiologinių echokardiogramų vaizdų apdorojimo metodika, leidžianti automatizuotai nustatyti sistolės ciklus ir aortos vožtuvo stenozei diagnozuoti būtinus parametrus. Pasiūlytas metodika siūlo pusiau automatizuotą sistolės ciklų trasavimą, kas leidžia sutaupyti iki 2 minučių kardiologo laiko sąnaudų vieno paciento vaizdų analizei. Gauto aortos vožtuvo stenozės prognostinio modelio jautrumas ir specifiškumas siekia 100 %.
- Sukurta atstumo metrika, skirta multireliaciniam klasterizavimui. Pasiūlyta metrika, palyginti su denormalizuoto duomenų rinkinio klasterizavimu ir su *RTED* multireliacine atstumo metrika, parodė aukštesnį klasterizavimo tikslumą: silueto reikšmes 0,21–0,31, palyginti su denormalizuotų duomenų atveju silueto reikšmėmis 0–0,16 ir RTED silueto reikšmėmis 0,15–0,23.

1.8 Darbo rezultatų aprobavimas

Tyrimų rezultatai publikuoti 7 recenzuojamuose periodiniuose moksliniuose leidiniuose ir 2 kituose moksliniuose leidiniuose. Autorius dalyvavo ir pristatė rezultatus 5 mokslinėse konferencijose Lietuvoje bei 4 tarptautinėse mokslinėse konferencijose, vykusiose užsienyje.

Straipsnių ir konferencijose skaitytų pranešimų sąrašas pateikiamas disertacijos darbo rezultatų aprobavimo skyriuje.

1.9 Disertacijos struktūra

Disertaciją sudaro įvadas, tyrimų srities aprašymas, trys skyriai, skirti rezultatams pristatyti, išvados, cituotos literatūros sąrašas. Kiekvieno skyriaus, išskyrus įvadą ir išvadas, pabaigoje yra pateikiamas apibendrinimas. Bendra disertacijos apimtis – 154 puslapiai (be priedų), 44 paveikslėliai ir 22 lentelės.

Įvade aprašoma tyrimų sritis ir aktualumas, pateikiamas problemos apibrėžimas, aptariami darbo tikslai ir uždaviniai, tyrimo metodai, pateikiama darbo rezultatų praktinė reikšmė, mokslinis naujumas, ginamieji teiginiai ir darbo rezultatų aprobavimas.

Pirmajame disertacijos skyriuje pristatoma duomenų tyrybos metodų panaudojimo medicinoje problema ir jos aktualumas. Papildomai pristatomi apklausos, atliktos universitetinėse ligoninėse, rezultatai. Antrajame skyriuje aprašoma pasiūlyta duomenų tyrybos taikymo metodika CRISP-MED-DM. Taip pat aprašoma pasiūlytos metodikos taikymų onkologijoje ir kardiologijoje teorinė dalis, pasiūlytas multireliacinio klasterizavimo metodas. Trečiajame skyriuje pateikiami eksperimentiniai pasiūlytos CRISP-MED-DM metodikos ir medicininių duomenų apdorojimo metodų įvertinimai. Išvadų skyriuje pateikiamos bendros tiriamojo darbo išvados.

2 Duomenų tyryba medicinoje ir sveikatos apsaugoje: apžvalga ir analizė

Pastarųjų dešimtmečių tendencija kompiuterizuoti gydymo procesą užtikrina vis spartesnį medicininės informacijos kaupimą. Šių duomenų analizė ir tyryba turi strateginę reikšmę sveikatos sektoriui ir yra svarbi kiekvienam pacientui. Sukauptų duomenų intelektuali analizė, kuri apima duomenų tyrybą (DT) ir platesnę sąvoką – žinių gavybą, siūlo naujas priemones greičiau diagnozuoti ligas, parinkti optimalaus gydymo algoritmą, prognozuoti gydymo trukmę ir rezultatus, minimizuoti komplikacijų riziką, optimizuoti sveikatos priežiūros įstaigos resursus.

Pirmajame disertacijos skyriuje apžvelgiamos DT ir žinių gavybos sąvokos, taikymo tikslai bei uždaviniai. Didžiausias dėmesys skirtas literatūros, nagrinėjančios DT medicinoje taikymus, analizei ir apibendrinimui. Pradinėse skyriaus dalyse apžvelgiami literatūroje nagrinėjami DT taikymai medicinoje, toliau analizuojami jų ypatumai ir apribojimai. Taip pat aprašytos pažangios DT technologijos, skirtos multireliacinių duomenų, duomenų srautų, daugialypės terpės duomenų ir tekstinės informacijos DT.

Apibendrinant DT taikymo medicinoje publikuotų tyrimų sprendžiamus uždavinius, galima teigti, kad jie priskiriami prie gydymo resursų optimizavimo arba gydymo kokybės gerinimo tikslų.

Toliau pirmajame skyriuje aprašomos specifinės medicinos taikomosios srities DT problemos. DT taikymas medicinoje skiriasi nuo kitų sričių tuo, kad pradiniai duomenys heterogeniški, klinikinių duomenų panaudojimas DT gali būti ribojamas etikos, socialinių bei teisinių aspektų. Apribojimai DT medicinoje taip pat susiję su privatumo ir duomenų saugumo grėsmėmis, pacientų teisinių ieškinių galimybe bei būtinybe įvertinti DT pranašumus ir potencialią klaidingų sprendimų riziką.

Pirmajame skyriuje aprašytos DT taikymo medicinoje problemos:

- Pirminių duomenų problema:
    - medicininių informacinių sistemų sąveikumas;
    - semantinis pirminių duomenų sąveikumas (bendros ontologijos parinkimas).
- Pirminių duomenų kokybės užtikrinimas:
    - duomenų pilnumo analizė;
    - duomenų korektiškumo analizė.
- Pacientų duomenų saugos ir etikos problematika:
    - pacientų duomenų apsaugos užtikrinimas (kriptografijos metodų naudojimas);
    - pacientų duomenų depersonalizavimas.

2.1 Universitetinių ligoninių apklausa

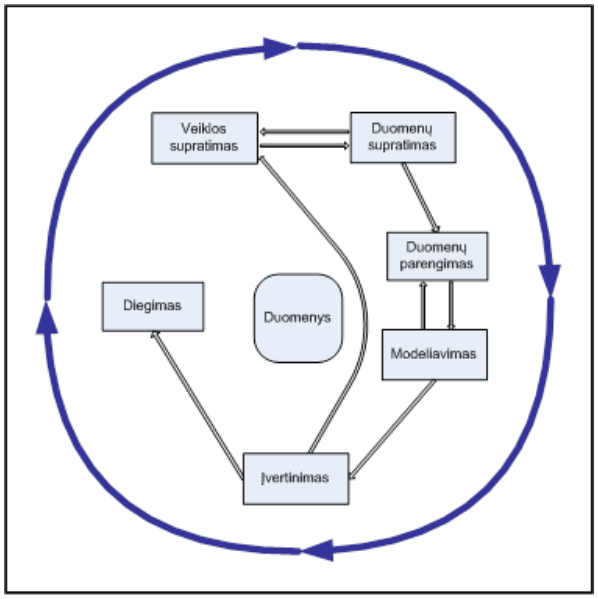
Siekiant patikrinti ir nustatyti praktinio DT metodų taikymo apribojimus medicinoje, buvo atlikta apklausa universitetinėse ligoninėse. Joje dalyvavo 8 ligoninių iš Lietuvos, Šveicarijos, Vokietijos, Albanijos ir Pietų Afrikos Respublikos atstovai. Apklausos dalyviams buvo patiekta įvadinė informacija ir 15 klausimų klausimynas, kurį galima buvo užpildyti internetu arba pateiktoje popierinėje formoje. Kiekviena ligoninė pateikė ne mažiau dviejų respondentų atsakymus.

Apibendrinus rezultatus, galima teigti, kad nors DT metodai naudojami dažniau, tik 29 % respondentų galėjo pateikti DT taikymo pavyzdį. Be to, medicinos profesijos atstovai plačiai taiko statistinės analizės metodus, tačiau yra mažai susipažinę su DT metodais ir jų taikymo galimybėmis. Net 86 % respondentų išreiškė norą gauti papildomą informaciją ir dalyvauti tarpdisciplininiuose projektuose, taikant DT medicinoje.

3 Sisteminis duomenų tyrybos ir duomenų analizės metodų taikymas medicinos srityje

Literatūroje apie DT taikymą medicinoje siūloma remtis bendrai priimtomis DT metodikomis. Siūlomi įvairūs DT proceso modeliai: CRISP-DM (Chapman, et al., 2000), SEMMA (Azevedo & Lourenco, 2008), Fayyad ir kt. 1996 m. proceso modelis (Fayyad, Piatetsky-Shapiro, & Smyth, 1996), Cabena 1998 m. proceso modelis (Cabena, Hadjinian, Stadler, & Verhees, 1998), Cios 2000 m. proceso modelis (Cios ir Moore, 2002).

Pagal DT specialistų bendruomenės *KDNuggets* 2014 m. vykdytos apklausos rezultatus (Piatetsky-Shapiro, 2014), plačiausiai paplitusi DT proceso metodika yra *Cross–Industry Standard Process for the Data Mining* (CRISP-DM). CRISP-DM apibrėžia procesų modelį, kuris DT išskaido į šešis etapus: Veiklos supratimą, Duomenų supratimą, Duomenų parengimą, Modeliavimą, Įvertinimą ir Diegimą. CRISP-DM apibrėžtas iteracinis procesų modelis pavaizduotas 1 paveiksle. Metodikos pavyzdiniame modelyje (angl. *reference model*) numatyta kiekvieno etapo įvestis, išvestis bei vykdymo strategija.



1 pav. CRISP-DM procesų modelis

CRISP-DM traktuoja DT procesą kaip klasikinį projektą, kuris turi apibrėžtą tikslą bei pagrindinius projekto apribojimus – laiką, išteklius ir apimtį. Kaip pabrėžia P. Baylis (Baylis, 1999), DT medicinoje prasideda nuo teisingo užduoties suformulavimo, kai gydytojai kartu su duomenų analizės specialistais suformuluoja probleminę sritį ir, analizuodami veiklos sritį bei medicininėse informacinėse sistemose prieinamus duomenis, suformuluoja problemą bei techninę užduotį.

Literatūroje pabrėžiamas DT sveikatos apsaugos srityje unikalumas (Bellazzi ir Zupan, 2008; Canlas Jr, 2009; Koh ir Tan, 2005; Cios ir Moore, 2002). Straipsnių autoriai siūlo atsižvelgti į papildomus uždavinius išgaunant žinias iš medicininių duomenų, tačiau DT taikymams medicinoje trūksta specifinės ir detalizuotos metodikos. Siekiant išspręsti DT taikymo medicinoje problemas, sukurtas CRISP-DM metodikos plėtinys CRISP-MED-DM, kuriame atsižvelgiama į medicinai būdingas problemas:

- tyryba nestandartinių duomenų pateikčių: multireliacinių duomenų, laiko ir erdvinių duomenų;
- heterogeninės informacinės sistemos;
- semantinis duomenų sąveikumas;
- etiniai, socialiniai ir teisiniai apribojimai.

Vienas iš CRISP-DM trūkumų yra metrikų arba kriterijų, kurie leistų įvertinti DT projekto atitiktį metodikos reikalavimams, stoka. Siekiant išspręsti šias problemas, disertacijoje pasiūlytas atitikties įvertinimo metodas. Šis metodas leidžia įvertinti DT projektą, nustatyti, kokia apimtimi buvo įvykdytos metodikoje apibrėžtos veiklos.

3.1 CRISP-DM metodikos plėtinys medicinai ir sveikatos apsaugai

Siekiant pašalinti CRISP-DM trūkumus, sukurtas naujas specializuotas CRISP-MED-DM pavyzdinis modelis (angl. *reference model*) ir atitikties vertinimo metodas. CRISP-DM metodikos pavyzdinis modelis įgyvendina „iš viršaus į apačią“ (angl. *„top-bottom approach“*) principą. Pirmieji du modelio abstrakcijos lygiai apibrėžia bendrus proceso etapus, jų užduotis ir pateiktis, kurios nėra specializuotos ir turi tikti visoms taikymo sritims. Trečiasis ir ketvirtasis pavyzdinio modelio abstrakcijos lygiai yra skirti specializacijai. Siūlomo pavyzdinio modelio plėtinio santrauka yra pateikta žemiau.

Disertacijoje pateiktas išsamus CRISP-MED-DM užduočių, veiklų ir pateikčių sąrašas.

3.1.1 Projekto apimtį apibrėžiantys 1 ir 2 etapai

Pirmasis etapas „Veiklos supratimas“ buvo pervadintas į „Problemos supratimą“ siekiant išvengti dviprasmiškumo klinikinio taikymo ir sveikatos priežiūros vadybos taikymo srityse. Be to, užduotis „Apibrėžti objektus/tikslus“ buvo suskaidyta į „Apibrėžti klinikinius objektus/tikslus“ ir „Apibrėžti sveikatos apsaugos vadybos objektus/tikslus“. Sprendžiant pacientų duomenų privatumo klausimą, užduotis „Situacijos vertinimas“ buvo papildytas veikla „Įvertinti paciento duomenų privatumą ir teisinius apribojimus“. Sprendžiant heterogeninių informacinių sistemų klausimą, buvo pridėta veikla „Įvertinti duomenų šaltinius ir vientisumą“.

Antrajame etape „Duomenų supratimas“ buvo įvesta nauja bendra užduotis „Pasirengimas duomenų rinkimui“. Buvo atsižvelgta į transporto, semantinio ir funkcinio sąveikumo problemas. Medicininių duomenų formatų gausa suvokiama per naujai įvestą užduočių rinkinį: nestandartinių duomenų išankstinio apdorojimo projektavimas apima

multireliacinių duomenų, nestruktūrizuotos tekstinės informacijos, daugialypės terpės duomenų apdorojimą. Medicininės nomenklatūros, klasifikatorių ir ontologijų apibrėžimas yra svarbus tolesniam išankstiniam duomenų apdorojimui. Etapo pabaigoje apibrėžiami ir analizuojami klinikinių duomenų modeliai bei duomenų šaltinių sistemose naudojami klinikiniai protokolai.

3.1.2 Įgyvendinimo 3 ir 4 etapai

Pasak Q. Yang ir S. Wu (Yang ir Wu, 2006), iki 90 % DT sąnaudų sudaro išankstinis duomenų apdorojimas: duomenų integravimas, transformacija, valymas, t. t. Tos pačios tendencijos pastebimos ir medicinos srityje. Pirminės CRISP-DM užduoties „Surinkti duomenis“ praktinis taikymas medicinoje yra ribotas, todėl yra įvedama užduotis „Parengti duomenis“, kurią sudaro šios veiklos:

- įdiegti nesusietų informacinių sistemų sąsajas;
- parengti medicinos terminų sąsajumo lenteles;
- analizuoti ir atlikti išankstinį duomenų apdorojimą, remiantis patvirtintais klinikinių duomenų modeliais ir protokolais.

Be to, į proceso modelį buvo įtraukta nauja užduotis „Duomenų išskyrimas“. Ji apima veiklas, susijusias su išankstiniu nestruktūrizuotų duomenų apdorojimu siekiant palengvinti požymių išskyrimą ir pasirengti DT modeliavimui. Užduoties veiklos yra šios:

- tekstinių duomenų apdorojimas;
- garsinių ir vaizdinių duomenų apdorojimas:
  - vaizdinė informacija;
  - vaizdo įrašo informacija;
  - garsinė informacija.
- signalų duomenų apdorojimas.

Pirminė CRISP-DM užduotis „Surinkti duomenis“ buvo papildyta požymių išskyrimu (angl. *feature extraction*), naudojant statistinius ir DT metodus. Veikla nustato požymių išskyrimo ir dimensijos mažinimo metodų panaudojimą, apibrėžiant modeliavimo veiklos galimus atributų rinkinius. Prognostiniai DT metodai reikalauja atskirti mokymo, testavimo ir validavimo duomenų rinkinius, todėl buvo įvesta duomenų atrankos užduotis.

Klinikiniuose duomenyse dažnai susiduriama su trūkstamų duomenų problema (angl. *missing data*). Tuo tarpu duomenų semantinė analizė leidžia identifikuoti klaidas duomenyse, paklaidas, atsiradusias dėl medicininės įrangos matavimo klaidų arba spragų laboratorijos ir stebėsenos įrangos sąsajose. Automatizuota semantinių klaidų paieška yra grindžiama veiklos taisyklėmis, įgyvendinančiomis matavimų reikšmių minimumo/maksimumo, paciento lyties ir amžiaus patikrinimus. Šios veiklos atspindi „Valyti duomenis“ pagrindinį uždavinį.

„Duomenų integravimo“ užduotyje buvo pridėta duomenų abstrakcijos lygio pakeitimo veikla. Ši veikla reikalinga agreguoti duomenų srautų duomenis. Pavyzdžiui, intensyvios terapijos skyriaus įranga gali generuoti tūkstančius duomenų elementų per sekundę. Todėl duomenų agregavimo metodai turi būti taikomi anksčiau nei DT modeliavimo veiklos.

Multireliaciniai duomenys reikalauja arba duomenų denormalizavimo į „vienos lentelės“ formatą, arba numatyti multireliacinių DT metodų naudojimą, tokių kaip indukcinės logikos programavimą (ILP).

Pasak C. Catley ir kt. (Catley, Smith, McGregor, & Tracy, 2009), DT bendradarbiavimo metodai (pavyzdžiui, metodų rinkiniai, metodų grandinės) gali būti pranašesni už atominius klasifikavimo DT metodus. Atsižvelgiant į tai, buvo pridėta nauja veikla „Optimalių modelių arba modelių rinkinių nustatymas“.

Galiausiai, besiruošiant Diegimo etapui, gauti modeliai turi būti parengti naudoti sprendimo priėmimo informacinėse sistemose. Šiam tikslui pasiūlyta veikla „Eksportuoti gautą modelį arba modelių rinkinį PMML formatu“.

3.1.3 Įvertinimo ir Diegimo etapai

Originalios CRISP-DM Įvertinimo ir Diegimo etapų veiklos atitinka medicinos srities poreikius ir gali būti naudojamos DT projektuose medicinoje ir sveikatos apsaugoje. Todėl šie etapai palikti be reikšmingų pokyčių.

3.1.4 CRISP-MED-DM vertinimo modelis

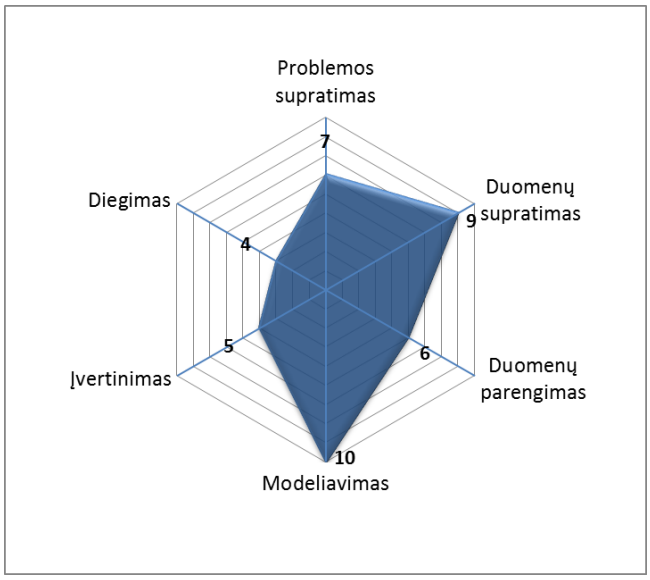
DT procesų nustatymas, stebėsena ir kokybės gerinimas reikalauja ne tik detalaus proceso modelio, bet ir patikimų bei pagrįstų vertinimo modelių. Atsižvelgiant į DT tikslus ir metodus, duomenų struktūros sudėtingumą ir duomenų apimtį, DT projektai yra labai skirtingi, todėl neįmanoma nubrėžti griežtą metodikos taikymo vertinimo standartą. Atsižvelgiant į tai, siūlomi vertinimo modeliai pasižymi lankstumu.

Disertacijoje pasiūlytos dvi atitikties CRISP-MED-DM metodikos vertinimo strategijos. Pirmoji grindžiama prielaida, kad kiekvienas proceso modelio etapas yra vienodai svarbus. Išimtis yra daroma paskutiniam Diegimo etapui, kurio veiklos yra faktinis DT proceso rezultatų panaudojimas. Antroji strategija grindžiama nuosekliu CRISP-DM proceso modeliu. Kiekvieno etapo rezultatai apibrėžia ir formuoja tolesnį etapą. Pritaikant šią strategiją vertinimo modelyje, priskiriami palaipsniui mažėjantys svoriai nuo pirmojo iki paskutiniojo etapo.

Pirmojoje vertinimo strategijoje kiekvienam etapui, išskyrus Diegimo etapą, priskiriama 10 balų, kurie parodo maksimalų etapo veiklų įgyvendinimo įvertinimą, kai visos CRISP-MED-DM metodikoje apibrėžtos privalomos veiklos yra įvykdytos.

Antrojoje vertinimo strategijoje kiekvienas etapas yra vertinamas skirtingai. Atitinkamai kiekvienas privalomos veiklos etapas įvertinamas taškais, gautais iš kaupiamojo balo. Disertacijoje yra pateikiamos CRISP-MED-DM užduočių ir veiklų vertinimo metrikos.

DT projekto atitikties vertinimo rezultatai, gauti naudojant siūlomus vertinimo modelius, gali būti vizualizuojamos *Radar* diagrama, kaip pavaizduota 2 paveiksle.



2 pav. DT projekto vertinimo pavyzdinė radaro (*Radar*) principo diagrama.

Kitose skyriaus dalyse pateiktas DT metodų taikymo medicinoje teorinės dalies aprašymas. Eksperimentinių tyrimų, panaudojant CRISP-MED-DM metodiką, aprašymai ir rezultatai pateikti ketvirtajame skyriuje.

3.2 BRCA1 geno mutacijos prognozavimas

Krūties vėžys yra dažniausiai moterims diagnozuojama vėžio rūšis ir viena iš dažniausiai pasitaikančių moterų mirties priežasčių visame pasaulyje. Pacientėms, turinčioms mutavusį *BRCA* geną, tikimybė susirgti krūties vėžiu siekia net 65 %. Mutavusiam *BRCA* genui nustatyti naudojami įvairūs rizikos modeliai. *BRCAPRO*, *Penn II*, *Myriad II*, *FHAT* ir *BOADICEA* modeliai apskaičiuoja riziką, remdamiesi skirtingų vėžio rūšių diagnozėmis šeimoje (Panchal, Ennis, Canon, & Bordeleau, 2008). Tarp visų minėtų rizikos modelių *Penn II* modelis, kurį pateikė Pensilvanijos universiteto Abramson vėžio centras, turi geriausią jautrumo reikšmę 0,93 (Panchal, Ennis, Canon, & Bordeleau, 2008).

Disertacijoje pasiūlytas naujas mutavusio *BRCA1* geno nešiotojų identifikavimo būdas, metodiškai taikant CRISP-MED-DM proceso modelio žingsnius ir naudojant DT metodus.

Atlikto tyrimo tikslas – pasiūlyti alternatyvų *BRCA1* geno mutacijos rizikos nustatymo modelį. CRISP-MED-DM taikymo rezultatai aprašyti ketvirtajame skyriuje.

3.3 Kraujo tėkmės echokardiogramų vaizdų analizė

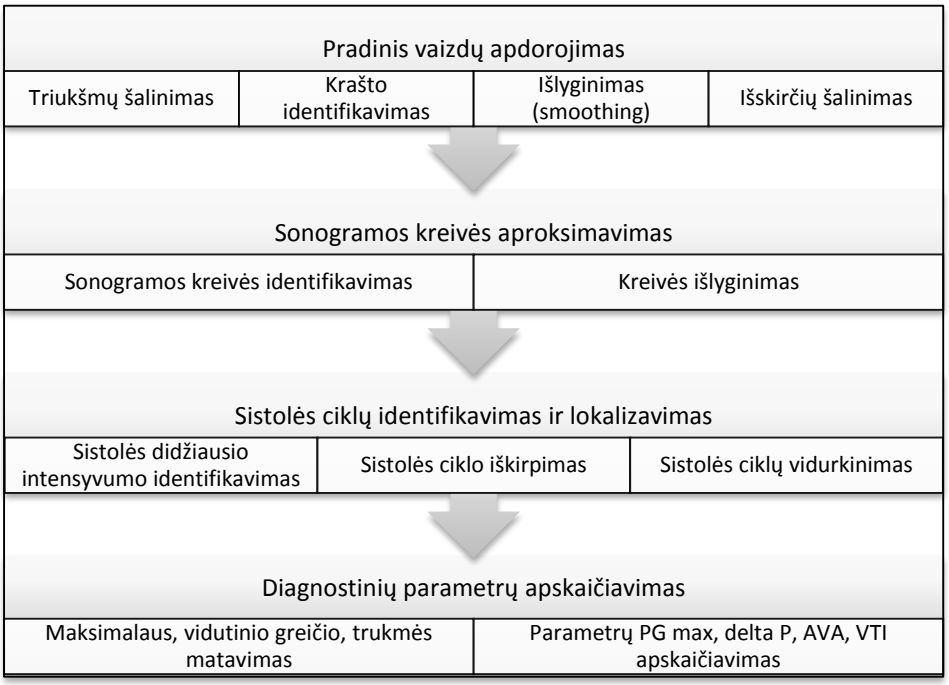
Aorta yra pagrindinė arterija, kuri perduoda kraują iš širdies į likusias kūno dalis. Aortos vožtuvas pasitarnauja kaip vartai tarp širdies ir aortos. Dėl kalkėjimo ar kitų procesų aortos vožtuvo angai susiaurėjus, kairysis skilvelis privalo dirbti stipriau, kad užtikrintų didesnį spaudimą, išpumpuotų kraują per vožtuvą. Aprašyta situacija vadinama aortos vožtuvo stenoze (AS).

Tam, kad diagnozuotų AS, kardiologas turi atlikti echokardiogramose pavaizduotų sistolių ciklų trasavimą. Šis procesas reikalauja kruopštaus rutininio darbo ir gali įvelti su žmogiškuoju faktoriumi susijusių klaidų. Sprendžiant šią problemą, buvo sukurtas pusiau automatinis echokardiogramų vaizdų apdorojimo įrankis, kuris padėtų kardiologams išvengti klaidų ar minimizuotų laiko sąnaudas, apdorojant echokardiogramų vaizdus.

Taikant CRISP-MED-DM metodiką, gautas aukšto tikslumo AS laipsnio prognostinis modelis.

3.3.1 Kraujo tėkmės echokardiogramų vaizdų apdorojimo metodika

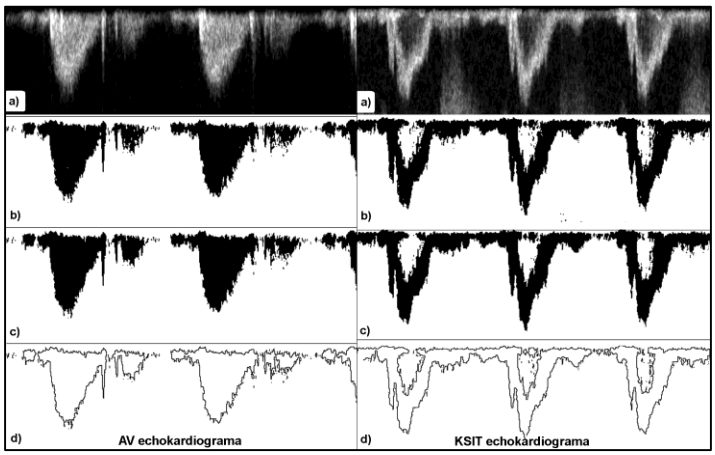
Siekiant automatizuoti echokardiogramų vaizdų apdorojimą, buvo pasiūlyta vaizdų apdorojimo metodika, pavaizduota 3 paveiksle, ir jos įgyvendinimas. Pasiūlyti vaizdų apdorojimo metodai buvo įgyvendinti naudojant R funkcinę programavimo kalbą (R Core Team, 2014), naudojant *ImageJ* (Abramoff, Magalhaes, & Ram, 2004; Schneider, et al., 2012) bibliotekos funkcijas.



3 pav. Aortos vožtuvo ir kairiojo skilvelio išvarymo trakto echokardiogramų apdorojimo metodika.

3.3.2 1 žingsnis – išankstinis echokardiogramų vaizdų apdorojimas

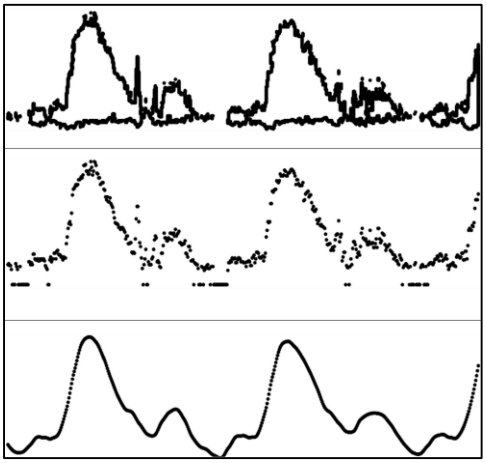
Pirmiausia vaizdai buvo konvertuojami į juodai baltus, buvo pašalintos išskirtys, triukšmai ir nepageidaujami artefaktai, lokalizuotas sistolės ciklus apibrėžiantis kontūras. Šio žingsnio rezultatai pateikti 4 paveiksle.



4 pav. Pradinio kraujo tėkmės echokardiogramos išankstinio apdorojimo žingsniai. Aortos vožtuvo (AV) kraujo tėkmės vaizdas kairėje. Kairiojo skilvelio išstūmimo trakto (KSIT) kraujo tėkmės vaizdas dešinėje. Išankstinio apdorojimo žingsniai išdėstyti horizontaliai nuo viršaus iki apačios: a) originalus paveikslas, b) juodai baltas vaizdas, c) vaizdai su užpildytomis skylėmis, d) gautas kontūras.

*3.3.3 **2 žingsnis – kraujo tėkmės echokardiogramos aproksimacijos kreivė***

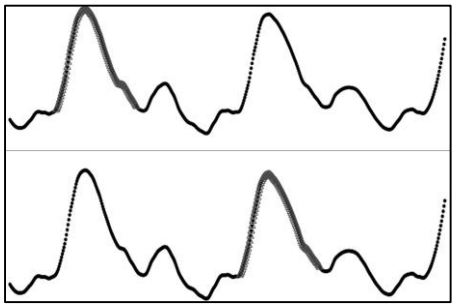
Gautas vaizdas išankstinio apdorojimo žingsnyje pateikia kontūrą, atitinkantį ultragarsinės įrangos pateiktą echokardiogramos vaizdą. Kairiojo skilvelio išstūmimo trakto (KSIT) echokardiogramose stebimi dvigubi kontūrai (4 pav. dešinėje). Kita svarbi problema – atsitiktiniai grioveliai, kurie yra Doplerio signalų matavimo triukšmo rezultatai. Šios problemos sprendžiamos pradžioje išfiltruojant duomenis iki 95 % procentilio, o toliau glotninant kreivę, pritaikant lokalią polinominę regresiją (Cleveland ir Loader, 1996). Šių žingsnių iliustracija pateikiama 5 paveiksle. Galutinis šių žingsnių rezultatas naudojamas matavimams ir skaičiavimams.



5 pav. Kraujo tėkmės echokardiogramos kreivės glotninimo žingsniai. Viršutinis paveikslas – išankstinio vaizdo apdorojimo rezultatas; vidurinis – taikant 95% procentilį; apatinis – taikant *Loess* kreivę.

3.3.4 3 žingsnis – sistolės ciklų atpažinimas

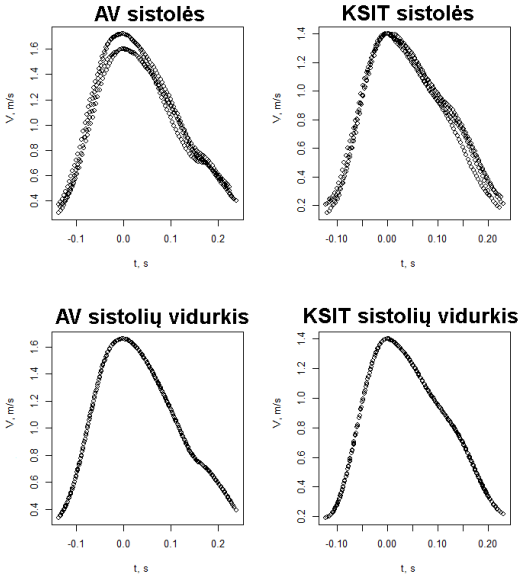
Kadangi echokardiogramose gali būti skirtingas sistolių ciklų skaičius, buvo pasiūlytas algoritmas, nustatantis jų skaičių ir identifikuojantis kiekvieną pilną sistolės ciklą. Atpažinti sistolės ciklai pavaizduoti 6 paveiksle.



6 pav. Atpažinti sistolės ciklai. Kraujo tėkmės kreivės vaizde yra atpažinti du pilni sistolės ciklai. Viršutinis paveikslas – pirmasis pilnas sistolės ciklas. Apatinis paveikslas – antrasis pilnas sistolės ciklas.

3.3.5 4 žingsnis – diagnostinių parametrų apskaičiavimas

Prieš atliekant matavimų skaičiavimus, vaizdų mastelis buvo keičiamas pagal echokardiogramose užfiksuotą pacientų širdies susitraukimo dažnį. Toliau buvo gaunama vidutinė sistolės ciklo kreivė, pritaikant lokalią polinominę regresiją (7 pav.). Greičio ir laiko integralo (angl. *velocity time integral*) parametrui apskaičiuoti gauta sistolės ciklo kreivė buvo aproksimuojama antrojo laipsnio polinomu ir apskaičiuotas jos apibrėžtinis integralas. Kiti diagnostiniai parametrai buvo apskaičiuoti naudojant Europos echokardiografijos asociacijos, Amerikos echokardiografijos visuomenės ir Europos kardiologų draugijų gairėse rekomenduojamas formules (Otto, 2012).



7 pav. Aortos vožtuvo ir kairiojo skilvelio išstūmimo trakto sistolių ciklų vidurkiai

Toliau, pagal CRISP-MED-DM numatytą eigą, trečiajame etape gauti duomenys buvo pritaikyti prognostiniam modeliui sudaryti ir įvertinti. Eksperimentinio tyrimo rezultatai pateikti ketvirtajame skyriuje.

3.4 Multireliacinių duomenų klasterizavimas

Dauguma DT metodų skirti struktūrizuotiems duomenims, kurie išreikšti pirmoje normalinėje formoje („vienos lentelės“ pavidalu). Duomenys medicinoje dažnai pateikiami reliaciniu pavidalu ir, normalizuojant juos į pirmą normalinę formą, galimai prarandama vertinga informacija.

Klasterizuojant multireliacinius duomenis, naudotas grupavimo algoritmas *Partitioning Around Medoids* (PAM), pasiūlytas 1987 metais L. Kaufman ir P. J. Rousseeuw (Kaufman ir Rousseeuw, 1987). PAM algoritmas buvo pasirinktas dėl išplečiamumo galimybių, išskirčių toleravimo ir galimybės pritaikyti neeuklidinės erdvės atstumo (panašumo) metrikas.

Disertacijoje pasiūlyta nauja atstumo metrika (panašumo matas), skirta multireliaciniams duomenims. Metodas skaičiuoja reliacinių objektų panašumų matricą, kuri naudojama kaip įvestis PAM klasterizavimo algoritme. Panašumo matas remiasi Gower ir Ochiai-Barkman panašumo metrikomis. Pasiūlyto multireliacinio klasterizavimo būdo taikymas medicininių leidinių metaanalizei aprašytas ketvirtajame skyriuje.

3.4.1 Panašumo matas multireliacinėje aplinkoje

Reliacinės duomenų struktūros, kurios turi skaitines ir nominalias vertes, neturi tiesioginės atitikties Euklidinėje erdvėje. Šiuo atveju klasikiniai atstumų matai, kurie naudojami klasterizavimo metoduose, kaip Manhattano, Minkovskio ar Euklidinio atstumo, nėra tinkami. Mišriems duomenų tipams gali būti naudojamas Gowerio bendrasis panašumo koeficientas (Gower, 1971). Gowerio panašumo koeficientas $s_{i,j}$ užrašomas taip:

$$s_{i,j} = \frac{\sum_k w_k s_{ijk}}{\sum_k w_k}, \quad (1)$$

čia s_{ijk} žymi k-tojo kintamojo panašumo matą, kuris skaičiuojamas atsižvelgiant į jo duomenų tipą, w_k – priskirta svorio funkcija.

Kitaip sakant, dviejų objektų i ir j panašumo matas yra normalizuotų pagal svorio funkciją kintamųjų k panašumų matų suma. s_{ijk} apskaičiavimas priklauso nuo duomenų tipo.

Nominalių reikšmių sąrašų palyginimui siūloma naudoti Ochiai–Barkman koeficientą.

$$s_{l1,l2} = \frac{n(l_1 \cap l_2)}{\sqrt{n(l_1) \times n(l_2)}}, \quad (2)$$

čia l_1, l_2 – sąrašai su lyginamomis nominaliomis reikšmėmis, $n(l)$ – elementų skaičius l sąraše.

Multireliacinėje aplinkoje lyginami objektai yra išreikšti esybėmis, jų sąryšiais bei atributais. Kiekvienam esybės atributui, kuris įtraukiamas į paieškos erdvę, turi būti paskaičiuojamas kiekvienas panašumo matas s_{ijk}, naudojant Gower koeficiento formulę, atitinkančią atributo duomenų tipą, ir Ochiai–Barkman koeficientą nominalinių kintamųjų sąrašų sulyginimui. Galiausiai, bendras dviejų objektų panašumo matas skaičiuojamas

kaip svertinė suma s_{ijk}, remiantis (1) formule.

4 CRISP-MED-DM metodikos ir duomenų analizės metodų aprobavimas

Šiame disertacijos skyriuje aprašyti teorinėje dalyje pasiūlytų metodikų ir metodų eksperimentiniai tyrimai.

4.1 BRCA1 geno mutacijos prognostinis modelis

4.1.1 Tyrimo duomenys

Medicininis tyrimas atliktas Lietuvos sveikatos mokslų universiteto Onkologijos institute nuo 2010 iki 2013 metų. Buvo tiriamos 83 moterys, kurioms diagnozuotas I ir II stadijos krūties vėžys su naviko morfologija: T1 N0, T2 N0, T3 N0, T1 N1, T2 N1. Pateikti duomenys buvo aukštos kokybės, be išskirčių ar trūkstamų reikšmių.

Atsižvelgiant į paciento duomenų privatumo teisinį reguliavimą, tyrimo duomenys buvo nuasmeninti ir panaikinta pacientų identifikaciją įgalinanti informacija.

4.1.2 BRCA1 geno mutacijos modeliavimas

Remiantis CRISP-MED-DM 4 etapo „Modeliavimas“ veiklų iteratyvumu, DT klasifikavimo metodai buvo taikyti keletą kartų naudojant sprendimų medžius, sprendimų taisykles, daugiasluoksnį perceptroną, loginę regresiją, *Naive Bajeso* klasifikatorius, modelių ansamblių (angl. *adaptive boosting)* ir modelių vidurkinimo (angl. *bagging*) klasifikavimo metodus.

Pirmojoje duomenų parengimo ir modeliavimo iteracijoje buvo įvertinti pasirinkti klasifikavimo algoritmai. Be to, klasifikavimo rezultatai buvo pagerinti pakeitus numatytuosius algoritmo parametrus. Antrojoje iteracijoje buvo subalansuotas duomenų rinkinis, palaipsniui sulyginant objektų proporciją pagal prognozuojamo atributo reikšmes.

Klasifikavimo modeliai buvo vertinami pagal jautrumą, specifiškumą, bendrą tikslumą, ROC reikšmes. Galutiniame etape klasifikavimo modelis su aukščiausiu ROC įverčiu buvo eksportuotas į PMML (*Predictive Model Markup Language*) formatą.

4.1.3 Tyrimo rezultatai

BRCA1 mutacijos rizikos prognostinis modelis su aukščiausia ROC reikšme buvo sukurtas pagal daugiasluoksnio perceptrono algoritmą (*MultilayerPerceptronCS realizacija* WEKA pakete). Gauto modelio bendras tikslumas 0,92, jautrumas 0,67, specifiškumas 0,96 ir plotas po ROC kreive 0,87. Tačiau šio modelio menkos interpretavimo galimybės ir žemas jautrumas nėra tinkami pritaikyti praktinės diagnostikos srityje. Aukščiausią jautrumą parodė modelis, gautas pritaikius sprendimų medžių vidurkinimo algoritmą *Bagging*: bendras tikslumas 0,71, jautrumas 0,96, specifiškumas 0,62 ir plotas po ROC kreive 0,65. Šio modelio jautrumo reikšmė aukštesnė už *Penn II* modelio, kurio jautrumas yra 0,93.

Aukščiausias specifiškumas buvo gautas taikant sprendimų taisyklių algoritmą *FURIA*: bendras tikslumas 0,75, jautrumas 0,09, specifiškumas 0,98 ir plotas po ROC kreive 0,63.

Kuriant krūties vėžio remisijos prognostinį modelį, pirmojoje iteracijoje optimizuojant algoritmų parametrus, pasiektas bendras modelio tikslumas 0,73, jautrumas

0,59, specifiškumas 1,0, plotas po ROC kreive 0,63. Antrojoje iteracijoje, atlikus duomenų rinkinio balansavimą, pavyko padidinti modelio jautrumą iki 0,96.

4.1.4 Klinikinis rezultatų įvertinimas

Pažymėtina, kad šiuo metu naudojami ir viešai prieinami *BRCA* genų mutacijos rizikos vertinimo modeliai yra grindžiami vien tik paciento šeimos ligos istorija. Disertacijoje gauti modeliai papildomai įtraukia klinikines ir morfologines paciento savybes ir turi panašų, o tam tikrais atvejais ir aukštesnį rizikos modelio tikslumą.

Taikant DT metodus, gautas prognostinis modelis patvirtino žinomus ir praktikoje naudojamus kriterijus, nurodančius tikėtiną *BRCA1* mutacijų radimą. Stiprus prognostinis veiksnys yra šeiminė vėžio anamnezė, ypač jos deriniai su tokiais klinikiniais bei morfologiniais požymiais, kaip abiejų krūtų vėžys (angl. *bilateral breast cancer*), bloga naviko diferenciacija (angl. *high grade*), histologinis medulinės karcinomos tipas, trejopai neigiamas krūties vėžys (angl. *triple negative*). Pažymėtina, kad gautame modelyje išryškėjo neigiamos progesterono receptorių raiškos kaip savarankiško prognozinio *BRCA1* mutacijų veiksnio vaidmuo. Didesnė *BRCA1* mutacijų radimo tikimybė, kai navikas yra 1 cm ir didesnis, arba kai pažeista daugiau nei vienas sritinis pažasties limfmazgis. Tai galima paaiškinti *BRCA* asocijuotų navikų blogesne diferenciacija ir greitesniu augimu.

Analizuojant ligos progreso prognozavimo modelius, neišryškėjo *BRCA1* mutacijų kaip prognozinio veiksnio vaidmuo. Tai patvirtina ankstesnių klinikinių tyrimų rezultatus (Robson et al., 2004), kad *BRCA* asocijuoto krūties vėžio prognozė nesiskiria nuo sporadinio krūties vėžio prognozės, kurią lemia klinikomorfologinės navikų savybės. Mūsų modelis patvirtino ankstesniuose tyrimuose nustatytus krūties vėžio progresavimo rizikos veiksnius, tokius kaip bloga naviko diferenciacija, pirminio naviko dydis, neigiami progesterono receptoriai, jaunas pacientės amžius, kai neskiriama adjuvantinė chemoterapija.

Galima teigti, kad validavus gautus modelius su didesnės apimties duomenimis, jie gali būti taikomi medicininėje praktikoje.

4.1.5 Atitikimas CRISP-MED-DM metodikai

Etapai – Duomenų supratimas (10 balų), Duomenų parengimas (10 balų), Modeliavimas (8,9 balo) ir Įvertinimas (6,7 balo) – rodo gerą metodikos atitikimą. Problemos supratimo etapas vertinamas 5,6 balų. Pateiktas modelis nėra šiuo metu įdiegtas sveikatos apsaugos infrastruktūroje, todėl Diegimo etapas vertinamas 0 balų.

4.2 Aortos vožtuvo stenozės prognostinis modelis

4.2.1 Tyrimo duomenys

Mediciniuius duomenis pateikė Vilniaus universiteto Santariškių klinikos. Tyrime dalyvaujantys kardiologai atrinko 18 pacientų ligos istorijas su demografiniais, klinikiniais ir Doplerio echokardiografijų duomenimis. Atrankos kriterijus antrinio panaudojimo duomenims buvo aortos vožtuvo stenozės laipsnis. Penki iš šių pacientų neturėjo jokių klinikinių aortos vožtuvo stenozės (AS) ženklų, 5 pacientams pasireiškė nedidelio laipsnio AS, 4 pacientams – vidutinio laipsnio AS ir 4 pacientams – sunki AS ligos forma.

Atsižvelgiant į paciento duomenų privatumo teisinį reguliavimą, tyrimo duomenys

buvo nuasmeninti ir panaikinta pacientų identifikaciją įgalinanti informacija.

4.2.2 Eksperimentiniai tyrimo metodai

Kraujo tėkmės matavimai buvo atlikti ultragarsinės diagnostinės sistemos pagalba, naudojant pulsinių bangų doplerometriją kairiojo skilvelio išstūmimo trakto (KSIT) tėkmei ir tęstinę bangų doplerometriją aortos vožtuve (AV) tėkmei matuoti. Kraujo tėkmė atvaizduojama grafiškai pateikiant kraujo tėkmės ertmėje vizualizaciją – kraujo greičio pokyčio laike echokardiogramą. Gauti vaizdai buvo saugomi vaizdų archyve (PACS sistemoje), iš kurios jie buvo eksportuojami DICOM formatu tolesniam apdorojimui asmeniniame kompiuteryje.

Buvo pritaikyti 2 skyriuje aprašyti Doplerio spektro vaizdų pirminio apdorojimo metodai. Pradinis duomenų rinkinys sudarytas iš 18 pacientų su 71 echokardiogramos vaizdu. Taikant sukurtą vaizdų apdorojimo metodiką, iš pradinio vaizdų rinkinio identifikuotos 71 AV ir 68 KSIT sistolės ciklai.

Siekdami įvertinti siūlomos metodikos efektyvumą, palyginome kardiologų rankiniu būdu gautus matavimus (M) su automatizuotos vaizdų analizės rezultatais (A). Įvertinta A ir M metodų rezultatų koreliacija bei Bland–Altman susiderinamumo ribos (angl. *limits of agreement*) (Bland ir Altman, 1986).

4.2.3 Aortos vožtuvo stenozės modeliavimas

Vadovaujantis CRISP-MED-DM metodologija, išankstinės analizės etape gautų duomenų rinkiniams buvo iteratyviai taikomi DT klasifikavimo metodai. Atlikti eksperimentai ir palyginti klasifikavimo modeliai, gauti iš kardiologų rankiniu būdu gautų parametrų ir analizuojant echokardiogramų vaizdus gautų parametrų. Abiem atvejais gautų prognostinių modelių bendrasis tikslumas buvo didesnis nei 98 %.

4.2.4 Tyrimo rezultatai

Palyginus siūlomą echokardiogramų vaizdo apdorojimo metodą (A) su rankiniu būdu profesionalių kardiologų gautais matavimais (M), gauti šie rezultatai:

1. Diagnostinių parametrų aortos vožtuvo maksimalus sistolinis greitis $AV\ Vmax$ ir aortos vožtuvo laiko–greičio integralas $AV\ VTI$ reikšmės, matuotos dviem metodais, parodė aukštą Pirsono koreliacijos koeficientą: $AV\ Vmax\ r(16) = 0{,}999, p < 0{,}0001$, $AV\ VTI\ r(16) = 0{,}988, p < 0{,}0001$. Tačiau $KSIT\ VTI$ matavimas parodė mažesnę koreliaciją: $r(16) = 0{,}68, p < 0{,}0001$.
2. Bland–Altman kreivės $AV\ Vmax$, $AV\ VTI$ ir $LVOT\ VTI$ parametrai: $AV\ Vmax\ \bar{d} = 0{,}02\ m/s$, $AV\ VTI\ \bar{d} = 0{,}16\ cm$, $KSIT\ VTI\ \bar{d} = 3{,}43\ cm$.

Gautas aortos vožtuvo stenozės laipsnio prognostinis modelis su 100 % jautrumo ir specifiškumo reikšmėmis.

Buvo atlikta (A) ir (M) metodų laiko sąnaudų analizė. Esant vidutiniam 20 pacientų per darbo dieną srautui, vienas kardiologas vidutiniškai sugaišta 24–28 minutes rutininiam sistolės ciklų trasavimui. Taikant sukurtą automatizuotą kraujo tėkmės echokardiogramų apdorojimo metodą, būtų sutaupoma vidutiniškai 22–26 minutės kardiologo darbo laiko.

4.2.5 Klinikinis rezultatų įvertinimas

Visų atrinktų pacientų širdies ritmai buvo reguliarūs, tačiau tėkmės kiekvieno ciklo metu gali šiek tiek skirtis priklausomai nuo diastolės ilgio. Teoriškai visi ciklai turėtų būti

vienodi, tačiau širdis kiekvieno ciklo metu gali išstumti nevienodą kraujo tūrį, priklausomai nuo ciklo trukmės. Esant ilgesnei diastolei, širdis daugiau užsipildo ir atitinkamai išstumia didesnį kraujo tūrį. Kardiologas, vertindamas vaizdus echoskopo monitoriuje, gali tų nedidelių ciklų nereguliarumų nepastebėti. Taip pat ciklai tarpusavyje gali truputį skirtis dėl techninių priežasčių, kai žmogui kvėpuojant širdis šiek tiek juda ir ne visus ciklus pavyksta užregistruoti vienodai. Išvardintos priežastys paaiškina, kodėl gauti parametrai, analizuojant echokardiogramas A metodu, kuris remiasi automatiškai nustatytais vidutiniais sistolės ciklais, skiriasi nuo gydytojų kardiologų atliktų matavimų.

4.2.6 Atitikimas CRISP-MED-DM metodikai

Etapai – Duomenų supratimas (8,9 balo), Duomenų parengimas (10 balų), Modeliavimas (7,8 balo) ir Įvertinimas (10 balų) – rodo gerą metodikos atitikimą. Problemos supratimo etapas vertinamas 6,7 balų. Pateiktas modelis nėra šiuo metu įdiegtas sveikatos priežiūros įstaigose, todėl Diegimo etapas vertinamas 0 balų. Pasiūlytam kraujo tėkmės echokardiogramų apdorojimo ir prognostiniam modeliui įdiegti klinikinėje aplinkoje bus būtina pritaikyti echoskopų programinę įrangą.

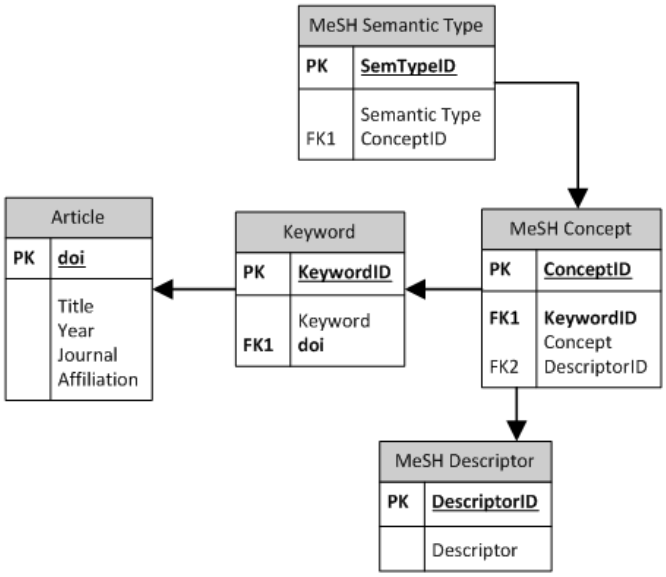
4.3 PubMed duomenų bazės publikacijų metaanalizė

Antrajame disertacijos skyriuje pasiūlytas naujas atstumo metrikos skaičiavimo metodas, kuris tinka duomenims multirealiacinėje formoje. Siūlomas metodas išbandytas ir palygintas su *Tree Edit Distance* (TED) panašumo mato skaičiavimo metodu (Pawlik ir Augsten, 2011), analizuojant *PubMed* duomenų bazėje indeksuotas publikacijas, aprašančias DT taikymus medicinoje. Identifikuojant populiariausias temas tarp *PubMed* esančių publikacijų buvo naudojamas grupavimo klasterizavimo metodas *Partitioning Around Medoids* (PAM), kuris remiasi padalijumu aplink *medoidus* (duomenų rinkinio elementus aproksimuojančius klasterių centrus). Algoritmas, kuris apskaičiuoja pilną atstumų matricą straipsnių rinkiniui, įgyvendintas R kalba. Lyginant su M. Pawlik ir N. Augset panašumo mato skaičiavimo algoritmu, pasiūlyta atstumo metrika leido gauti klasterius su didesnėmis silueto (angl. *silhouette*) reikšmėmis.

4.3.1 Tyrimo duomenys

Eksperimentuose buvo naudojama *PubMed* duomenų bazė kaip didžiausia medicininių publikacijų bazė, turinti aiškią publikacijų hierarchinę semantinio žymėjimo sistemą *MeSH* (National Center for Biotechnology Information, 2009).

Kaip matyti iš 8 paveiksle pateiktos esybių-ryšių diagramos, esybės Konceptas (*MeSH Concept*), Deskriptorius (*MeSH Descriptor*) ir Semantinis Tipas (*MeSH Semantic Type*), atitinkantys MeSH sąvokas, netiesiogiai susieti su centrine Straipsnio (*Article*) esybe. Šios esybės buvo pasirinktos remiantis jų semantine verte ir svarba atliekamai analizei.



8 pav. *PubMed* publikacijų metaduomenų ER diagrama

Norint paskaičiuoti dviejų straipsnių A1 ir A2 panašumą $sim_{A1,A2}$, turime paskaičiuoti visų esybių (*C – concept, D – descriptor, S – semantic type*) panašumus *simC*, *simD* ir *simS*:

$$sim_{A1,A2} = \frac{w_c simC + w_d simD + w_s simS}{\mathrm{w_c + w_d + w_s}},$$

kur,

$$\mathrm{simC} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}(concept_i(A_1), concept_j(A_2))}{\sqrt{m \times n}},$$

$$\mathrm{simD} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}(descriptor_i(A_1), descriptor_j(A_2))}{\sqrt{m \times n}},$$

$$\mathrm{simS} = \frac{\sum_{i=1}^{n}\sum_{j=1}^{m} s_{ij}(semantictype_i(A_1), semantictype_j(A_2))}{\sqrt{m \times n}}.$$

Iš esmės *simC*, *simD* ir *simS* skaičiuoja reikšmių sąrašų (atitinkamai Sąvokų, Deskriptorių ir Semantinių Tipų) panašumus, kurie yra santykinai susieti su centrine Straipsnio esybe. Atitinkamai, dviejų straipsnių atstumas apibrėžiamas:

$$dist_{A1,A2} = 1 - sim_{A1,A2}$$

4.3.2 Tyrimo rezultatai

Buvo sukurtas ir išbandytas sudėtinis panašumo mato skaičiavimo algoritmas multireliacinių duomenų struktūroms. Pasiūlytas sudėtinis panašumo matas sujungia Gower koeficientą su Ochiai–Barkman koeficientu ir yra tinkamas duomenų rinkiniams multireliaciniu pavidalu.

Iš viso buvo suskaičiuotos 2 284 453 straipsnių atstumų reikšmės. Tam, kad būtų įvertinta bendra klasterizavimo kokybė, buvo panaudota klasterio silueto vertė. Silueto vertė nusako kiekvieno klasterio kokybę. Eksperimentiškai ieškant klasterių skaičiaus nuo dviejų iki penkiasdešimties, didžiausia silueto vertė buvo intervale 0,21–0,30.

Gauti rezultatai buvo palyginti su M. Pawlik ir N. Augsten pasiūlytu *TED* medžių struktūros redagavimo panašumo mato skaičiavimo algoritmu *RTED*. Geriausias *RTED* panašumo mato taikymo rezultatas – silueto vertės intervalas $0{,}15-0{,}23$.

Papildomai buvo atliktas palyginimas atlikus pradinio duomenų rinkinio denormalizavimą iki pirmos normalinės formos. Gautų klasterių silueto reikšmės 0–0,16.

4.3.3 Atitikimas CRISP-MED-DM metodikai

Etapai – Duomenų supratimas (10 balų), Duomenų parengimas (10 balų), Modeliavimas (7,8 balo) ir Įvertinimas (6,7 balo) – rodo patenkinamą metodikos atitikimą. Priešingai prognostinei DT, tiriamojoje DT nekeliami tikslūs sėkmės kriterijai. Taip pat ne visada yra galimybių validuoti gautą informaciją. Dėl šių priežasčių didžioji dalis Problemos supratimo etapo veiklų nebuvo atliktos ir šio etapo atitikimas vertinamas 3,3 balo. Be to, nėra numatyta diegti įgyvendintą *PubMed* straipsnių metaanalizės metodą kitose informacinėse sistemose, todėl Diegimo etapas vertinamas 0 balų.

Bendros išvados

Šioje disertacijoje atlikti tyrimai leido padaryti tokias išvadas:

1. CRISP-DM pagrindu sukurta duomenų tyrybos taikymo metodika CRISP-MED-DM pasižymi šiomis savybėmis:
   - CRISP-MED-DM metodikoje, palyginti su kitomis duomenų tyrybos taikymo metodikomis, pirmą kartą pasiūlytas detalus procesų modelis, kuris atsižvelgia į medicinos ir sveikatos apsaugos problematiką. CRISP-MED-DM procesų modelis buvo praplėstas 33 užduotimis, kurios penkiuose metodikos etapuose siūlo užduotis ir veiklas medicininių duomenų išankstinio apdorojimo, semantinio sąveikumo, pacientų duomenų privatumo apsaugos problemoms spręsti.
   - Sukurtas atitikties įvertinimo metodas, kuris leidžia atlikti formalų duomenų tyrybos taikymų projektų atitikimo CRISP-MED-DM metodikai įvertinimą. Atitikties įvertinimo modelyje pasiūlytos metrikos ir vertinimo formulės, kurios leidžia vertinti taikymo projektų vykdymo kokybę, lyginti juos tarpusavyje.
   - CRISP-MED-DM metodika buvo sėkmingai aprobuota prognostinio modeliavimo tiriamuosiuose projektuose kardiologijos ir onkologijos srityse.
2. Sukurta kraujo tėkmės echokardiogramos vaizdo apdorojimo metodika ir ją įgyvendinanti programinė įranga leidžia automatizuotai identifikuoti sistolės ciklų kreives iš Doplerio echoskopų gaunamų vaizdų bei išgauti charakterizuojančius parametrus duomenų tyrybos metodų taikymui, sutaupant gydytojo laiką, skiriamą sistolės ciklų trasavimui:
   - Echokardiogramos automatinio vaizdų apdorojimo rezultatai pasižymi stipria koreliacija su kardiologų atliktais pagrindinių diagnostinių parametrų matavimais: aortos vožtuvo maksimalus sistolinio greičio $AV\ V_{max}$ parametro Pirsono koreliacijos koeficientas $r(16) = 0{,}999$ ($p<0{,}0001$); aortos vožtuvo laiko integralo $AV\ VTI$ parametro – $r(16) = 0{,}988$ ($p<0{,}0001$); vidutinio pikinio gradiento ΔP_{max} parametro – $r(16) = 0{,}994$ ($p<0{,}0001$); aortos vožtuvo ploto AVA parametro – $r(16) = 0{,}894$ ($p<0{,}0001$).

- Taikant CRISP-MD-DM metodiką kartu su pasiūlytu vaizdų apdorojimo metodika, buvo sukurtas tikslus prognostinis modelis su 100 % jautrumo ir specifiškumo reikšmėmis.

3. Taikant CRISP-MED-DM metodiką, pavyko pagerinti sukurto *BRCA1* geno mutacijos prognostinio modelio tikslumą:
   - Pagerintos *BRCA1* geno mutacijos prognostinio modelio savybės: bendras tikslumas nuo 0,88 iki 0,94, jautrumas nuo 0,67 iki 0,83, specifiškumas nuo 0,85 iki 0,99, plotas po ROC kreive nuo 0,696 iki 0,81.
   - Pagerintos krūties vėžio remisijos prognostinio modelio savybės: bendras tikslumas nuo 0,73 iki 0,75, jautrumas nuo 0,59 iki 0,96, specifiškumas nepagerėjo, plotas po ROC kreive nuo 0,63 iki 0,65.
4. Sukurta atstumo metrika leidžia atlikti klasterizavimą su multireliaciniais duomenimis nedenormalizuojant jų iki pirmosios normalinės formos:
   - *PubMed* duomenų bazės publikacijų meta-analizė panaudojant pasiūlytą atstumo metriką leido klasterizuoti multireliacinį duomenų rinkinį į grupes pasiekiant silueto reikšmes 0,21–0,31. Tai yra geresnis rezultatas lyginant su *Tree Edit Distance* atstumo metrika (silueto reikšmės 0,15–0,23) ir denormalizuotų duomenų klasterizavimu (silueto reikšmės 0–0,16).
   - Kadangi multireliacinių objektų porų atstumų skaičiavimai yra nepriklausomi vieni nuo kitų, jie gali būti atliekami lygiagrečiai. Sukurta programinė įranga, kuri įgyvendina lygiagrečius pasiūlytos atstumo metrikos skaičiavimus, sutrumpinant vykdymo laiką proporcingai naudojamų procesorių skaičiui.

Autoriaus mokslinių publikacijų disertacijos tema sąrašas

1. Niakšu, O.; Balčiūnaitė, G.; Kizlaitis, R. J.; Treigys P. 2015. Semi-automation of Doppler Spectrum Image Analysis for Grading Aortic Valve Stenosis Severity. *Methods of Information in Medicine.* (accepted), ISSN: 0026-1270 (IF: 2.248).
2. Niakšu, O. 2015. CRISP Data Mining Methodology Extension for Medical Domain. *Baltic Journal of Modern Computing*. Vol. 3, 2: 92–109, ISSN: 2255-8942.
3. Niakšu, O. 2014. Duomenų tyryba medicinoje: taikymas, problemos ir galimybės. *Visuomenės sveikata*, vol. 4(67), 9–19, ISSN: 1392-2696.
4. Niakšu, O.; Žaptorius, J. 2014. Applying Operational Research and Data Mining to Performance Based Medical Personnel Motivation System. *Studies in Health Technology and Informatics*, vol. 198, 63–70, IOS Press, Inc., ISSN: 0926-9630
5. Niakšu, O.; Skinulytė, J.; Duhaze, H. G. 2014. A Systematic Literature Review of Data Mining Applications in Healthcare. *Workshop Proceedings of Web Information Systems Engineering Conference* – WISE 2013, Springer 2014 Lecture Notes in Computer Science, 313–324, ISBN 978-3-642-54369-2.
6. Niakšu, O.; Gedminaitė, J.; Kurasova, O. 2013. Data mining approach to predict BRCA1 Gene Mutation, *Computational Science and Techniques*, vol. 1, 155–170, ISSN: 2029-9966.
7. Miškinis, P.; Niakšu, O.; Valuntaitė, V. 2013. Mathematical Modelling of Time-related Blood Velocity Changes in Human Aorta. *Laboratorinė medicina*, 15(4), 182–187, ISSN: 1392-6470.

8. Niakšu, O. 2013. Calculating Distance Measure for MRDM Clustering. *Proceedings of the 16th International Multi-conference "Information Society – IS 2013"*, vol. A, 192–194, ISBN: 978-961-264-066-8.
9. Niakšu, O.; Kurasova, O. 2012. Data Mining Applications in Healthcare: Research vs Practice, *Databases and Information Systems BalticDB&IS*, Local Proceedings, 58–70, ISSN: 1613-0073.

Santraukoje cituota literatūra

Abramoff, M. D., Magalhaes, P. J. Ram, S. J. 2004. Image Processing with Imagej. *Biophotonics International*, 11(7), p. 36-43.

Azevedo, A., Santos. M. F. 2008. KDD, SEMMA and CRISP-DM: a Parallel Overview [interaktyvus]. *Prieiga per internetą* <http://disi.unal.edu.co/profesores/eleonguz/cursos/md_2014/documentos/metodologias.pdf> [žiūrėta 01 04 2015].

Baylis, P. 1999. Better Health Care with Data Mining. *SPSS White Paper*, UK, p. 1-8.

Bellazzi, R. Zupan, B. 2008. Predictive Data Mining in Clinical Medicine: Current Issues and Guidelines. *Int J Med Inform*, Feb, 77(2), p. 81-97.

Bodenreider, O. 2008. Biomedical ontologies in action: role in knowledge management, data integration and decision support. *Yearbook of medical informatics*, p. 67.

Cabena, P., Hadjinian, P., Stadler, R. Verhees, J. &. Z. A. 1998. Discovering Data Mining: From Concept to Implementation. *Prentice-Hall, Inc.*

Canlas, R. D. 2009. Data mining in healthcare: Current Applications and Issues. *School of Information Systems & Management, Carnegie Mellon University, Australia*.

Castro, D. 2009. Explaining International Health IT Leadership. *Information Technology and Innovation Foundation, Washington*.

Catley, C., Smith, K., McGregor, C. Tracy, M. 2009. Extending CRISP-DM to Incorporate Temporal Data Mining of Multidimensional Medical Data Streams: A Neonatal Intensive Care Unit Case Study. In *Computer-Based Medical Systems. 22nd IEEE International Symposium*. IEEE, p. 1-5.

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C., Wirth, R. 2000. CRISP-DM 1.0 Step-by-step data mining guide [interaktyvus]. *Prieiga per internetą* <https://the-modeling-agency.com/crisp-dm.pdf> [žiūrėta 01 04 2015].

Chen, H., Fuller, S. S., Friedman, C. Hersh, W. 2006. Medical Informatics: Knowledge Management and Data Mining in Biomedicine. *Springer Science & Business Media,* vol. 8.

Cios, K. J. Moore, W. G. 2002. Uniqueness of Medical Data Dining. *Artificial Intelligence in Medicine*, 26(1), p. 1-24.

Cleveland, W. S. Loader, C. 1996. Smoothing by Local Regression: Principles and Methods. *Statistical Theory and Computational Aspects of Smoothing*. *Springer*, p. 10-49.

Esfandiari, N., Babavalian, M. R., Moghadam, A.-M. E. Tabar, V. K. 2014. Knowledge Discovery in Medicine: Current Issue and Future trend. *Expert Systems with Applications*, 41(9), p. 4434-4463.

Fayyad, U., Piatetsky-Shapiro, G. Smyth, P. 1996. From Data Mining to Knowledge Discovery in Databases. *AI magazine*, 17(3), p. 37.

Gower, J. C. 1971. A General Coefficient of Similarity and Some of Its Properties. *Biometrics*, p. 857-871.

Kaufman, L. Rousseeuw, P. 1987. Clustering by Means of Medoids. *North-Holland.*

Koh, H. Tan, G. 2005. Data Mining Applications in Healthcare. *J Healthc Inf Manag*, 19(2), p. 64-73.

Martin Bland, J. Altman, D. 1986. Statistical Methods ror Assessing Agreement Between Two Methods of Clinical Measurement. *The Lancet*, 327(8476), p. 307-310.

National Center for Biotechnology Information. 2009. PubMed - database of references and abstracts on life sciences and biomedical topics [interaktyvus]. Prieiga per internetą <http://www.ncbi.nlm.nih.gov/pubmed> [žiūrėta 01 04 2015].

Niakšu, O. Kurasova, O. 2012. Data Mining Applications in Healthcare Theory vs Practice. *DB&IS Local Proceedings*, p. 58-70.

Otto, C. M. 2012. The Practice of Clinical Echocardiography. *Elsevier Health Sciences.*

Panchal, S. M., Ennis, M., Canon, S. Bordeleau, L. J. 2008. Selecting a BRCA Risk Assessment Model for Use in a Familial Cancer Clinic. *BMC medical genetics*, 9(1), p. 116.

Pawlik, M. Augsten, N. 2011. RTED: a Robust Algorithm for the Tree Edit Distance. *Proceedings of the VLDB Endowment*, Volume 5.4, p. 334-345.
Piatetsky-Shapiro, G. 2014. CRISP-DM, still the top methodology for analytics, data mining, or data science projects [interaktyvus]. Prieiga per internetą <http://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html> [žiūrėta 01 04 2015].
R Core Team. 2014. R: A Language and Environment for Statistical [interaktyvus]. Prieiga per internetą <http://www.R-project.org> [žiūrėta 01 04 2015].
Rudnick, A. 2004. An Introductory Course in Philosophy of Medicine. *Medical Humanities*, 30(1), p. 54-56.
Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. 2012. NIH Image to ImageJ: 25 Years of Image Analysis. *Nature Methods*, 9(7), 671-675.
Špečkauskienė, V. Lukoševičius, A. 2009. Methodology of Adaptation of Data Mining Methods for Medical Decision Support: Case study. *Electronics and Electrical Engineering*, 2(90), p. 25-28.
Stroetmann, K. A., Artmann, J., Stroetmann, V. N. Whitehouse, D. 2011. European Countries on Their Journey Towards National eHealth Infrastructures. *Final European Progress Report*, p. 1-47.
Wilson, A., Thabane, L. Holbrook, A. 2004. Application of Data Mining Techniques in Pharmacovigilance. *British Journal of Clinical Pharmacology*, 57(2), p. 127-134.
Yang, Q. Wu, X. 2006. 10 Challenging Problems in Data Mining Research. *International Journal of Information Technology & Decision Making*, 5(04), p. 597-604.

Trumpos žinios apie autorių

Olegas Niakšu gimė 1979 m. rugsėjo 14 d. Vilniuje. 1997 m. baigė Vilniaus Karolinos vidurinę mokyklą. Vilniaus Gedimino technikos universitete 2001 m. įgijo informatikos bakalauro laipsnį, 2003 m. – informatikos inžinerijos magistro laipsnį. Nuo 2009 iki 2014 metų Vilniaus universiteto Matematikos ir informatikos instituto doktorantas.

Aktyviai dirba medicininių informacinių sistemų diegimo, nacionalinių e. sveikatos projektų srityse. Konsultavo ir vadovavo įgyvendinant e. sveikatos ir ligoninių informacinių sistemų projektus Lietuvoje, Pietų Afrikos Respublikoje, Kazachstane, Albanijoje, Irane.

Nuo 2006 m. yra IPMA sertifikuotas projektų vadovas, nuo 2009 m. profesinių asociacijų HIMMS ir ACM narys.

E. paštas *niaksu@acm.org*

THE DEVELOPMENT AND APPLICATION OF DATA MINING METHODS IN MEDICAL DIAGNOSTICS AND HEALTHCARE MANAGEMENT

Research Context and Motivation

The healthcare domain is known for its ontological complexity and variety of medical data standards and variable data quality (Cios and Moore, 2002; Chen et al., 2006; Bodenreider, 2008; Esfandiari et al., 2014). With the addition of patient data privacy issues, making an effective and practically usable medical knowledge discovery is of ongoing importance over recent decades. Modern clinical practices also undertake a transformation not only in diagnosis and treatment methods, but also in the understanding of health and illness concepts, moving from disease-oriented problem solving to a patient-centric approach, where computer-aided knowledge discovery methods play an important role (Rudnick, 2004).

Although data mining methods and tools have already been applied in various domains for more than 40 years, their applications in healthcare are relatively young. R.D. Wilson et al. (Wilson et al., 2004) have started to classify and collect medical publications where knowledge discovery and data mining techniques were applied or researched from 1966 until 2002.

Starting from the twentieth century, many countries have chosen e-Health as a prioritized national program, which in essence proposes to benefit from the standardized aggregation of patients' clinical information and healthcare services rendered by providing instant access to this information for healthcare professionals as well as to patients themselves (Castro, 2009; Stroetmann et al., 2011). According to the strategic plans of EU member states, the USA and many other nations from all continents, a considerable amount of investments are allocated to enable the global computerization of healthcare data. Taking a linear progression would propose that in 10 years all new medical encounters will be thoroughly digitalized, at least in the developed countries. For the first time in history, the research community is going to get a full set of a person's medical history from the birth date until the decease date. This anticipated scenario forecasts a tremendous potential for machine learning and in particular for data mining applications in healthcare.

Problem Statement

The application of data mining in healthcare raises additional challenges which require specific methods, tools, and methodology. Moreover, cross-domain knowledge is of key importance to achieve practical results. The rapid progress in the computerization of the healthcare industry gave a vast amount of heterogeneous, both structured and unstructured, data available for research and secondary use. There are hundreds of algorithms implemented to classify, cluster, and find hidden patterns in data. However, domain specific issues of healthcare are still to be resolved. As it was discussed by Cios et al., Bellazi et al., Špečkauskienė et al. (Cios and Moore, 2002; Bellazzi and Zupan, 2008; Špečkauskienė and Lukoševičius, 2009), specific problems shall be resolved to successfully apply data mining methods. According to their studies, without resolving

depersonalization, multi-relational and media data pre-processing, clinical data heterogeneity, and quality issues, data mining application is sub-optimal or impossible.

The surveys conducted by the data mining community KDNuggets (Piatetsky-Shapiro, 2014) in 2009 and 2014 have revealed the most widely used data mining application methodology is the Cross-Industry Standard Process for the Data Mining (CRISP-DM). However, due to its generic purpose CRISP-DM is not well suited for applications in the medical domain. Furthermore, a survey of university hospitals (Niakšu and Kurasova, 2012) has revealed that frequently data mining research projects remain theoretical, have no clinical follow-up, and rarely go beyond the institutions directly involved in the research. In order to apply data mining methods for clinical data, the researchers shall additionally resolve the problems related to patient privacy, semantic interoperability, heterogeneous data sources, and unstructured data presented in text or media formats.
Thus, there is a need for a methodology with a data mining process model to tackle the problems of the medical domain.

Tasks and Objectives of the Research

The main goal of this thesis is to develop and evaluate a medical domain specific methodology for predictive and explorative data mining in medicine and healthcare. The methodology shall address the issues typical for data mining in medicine, by defining the activities and the deliverables to tackle them. In addition, an evaluation model is needed to provide the compliance assessment to the methodology.
In order to achieve this goal, the following objectives and corresponding tasks have been formulated:

1. To analyze the existing data mining application methodologies by investigating data mining as part of a knowledge discovery process model.
2. To propose a novel, specific to the medical domain, data mining application methodology, which resolves the issues of the existing methodologies.
3. To evaluate the proposed data mining methodology in several medical specialty domains by creating the required medical data, such as diagnostic images, multi-relational data, analysis and processing methods.
4. To propose a multi-relational clustering method implementation for mining data in a multi-relational format.

Practical Significance of the Results

The practical significance of the thesis is as follows:

- The proposed CRISP-MED-DM methodology facilitates a data mining process in the medical domain by proposing the improved reference model and the compliance evaluation method.
- The proposed *BRCA1* gene mutation prediction model can be used as a decision support tool, to indicate the gene mutation risk before an expensive genetic test is carried out.
- The proposed echocardiography image analysis and feature extraction method and its software implementation allows automating the labor-intensive manual systole tracing performed by cardiologists when assessing aortic stenosis. The created aortic valve stenosis predictive model can be used as a decision support tool.

- The proposed and implemented distance metric can be applied to any exploratory analysis problem in a multi-relational environment, which cannot be reduced to a "single-table" form without a significant loss of information. The developed software calculates the distance matrix for multi-relational objects.

Research Methods

The exploratory research and systematic literature review were used to analyze and apply the results of other research. Various methods of statistical analysis, operation research, data mining, and image processing techniques were applied. Experimental research was used to evaluate the proposed methods and compare them with alternative approaches.

Statements to be Defended

1. The data mining methodology CRISP-DM can be specialized and extended to improve data mining performance in the medical domain.
2. Applying CRISP-MED-DM to create the breast cancer susceptibility gene BRCA1 prediction model improves the model's accuracy.
3. The application of CRISP-MED-DM with the novel echocardiography image transformation techniques results in a highly accurate aortic valve stenosis prediction model sufficient for aortic valve stenosis grading.
4. The partitioning clustering with the proposed multi-relational similarity measure is more precise in multi-relational settings where data generalization to one-table format leads to information loss.

Scientific Novelty and Results

The scientific novelty and results of the thesis are as follows:

1. A novel data mining application methodology CRISP-MED-DM is created. It defines tasks and activities to resolve the issues typical to the medical domain. Application of the CRISP-MED-DM allowed the *BRCA1* gene mutation risk prediction model's accuracy to improve from 0.88 to 0.94, sensitivity from 0.67 to 0.83.
2. A novel method for cardiology echocardiography image analysis, transformation and feature extraction is created, allowing the prediction of the aortic valve stenosis grade. The proposed method implements semi-automated systole cycle tracing and provides the cardiologists a time saving of up to two minutes per patient. The derived aortic valve stenosis predictive model has 100 % sensitivity and specificity for the research dataset.
3. A novel similarity (distance) measure for multi-relational data is created. The proposed metric when compared with a propositionalized dataset clustering and multi-relational clustering with RTED metric showed higher clustering accuracy with silhouette values of 0.21–0.31 against 0–0.16 (propositionalized) and 0.15–0.23 (RTED).

Approval of the Research

Articles in the reviewed scientific periodical publications:

1. Niakšu, O.; Balčiūnaitė, G.; Kizlaitis, R. J.; Treigys P. Semi-automation of Doppler Spectrum Image Analysis for Grading Aortic Valve Stenosis Severity. *Methods of Information in Medicine.* 2015 (accepted), ISSN: 0026-1270 (IF: 2.248).
2. Niakšu, O. CRISP Data Mining Methodology Extension for Medical Domain. *Baltic Journal of Modern Computing*. 2015. Vol. 3, 2: 92-109, ISSN: 2255-8942.
3. Niakšu, O.; Gedminaitė, J. & Kurasova, O. Data mining approach to predict BRCA1 gene mutation, *Computational Science and Techniques*, 2013, vol. 1, 155–170, ISSN: 2029-9966.
4. Miškinis, P.; Niakšu, O.; & Valuntaitė, V. Mathematical Modelling of Time-related Blood Velocity Changes in Human Aorta. *Laboratorinė medicina*. 2013, 15(4), 182–187, ISSN: 1392-6470.
5. Niakšu, O. Duomenų tyryba medicinoje: taikymas, problemos ir galimybės. *Visuomenės sveikata.* 2014, vol. 4(67), 9–19, ISSN: 1392-2696.
6. Niakšu, O., & Žaptorius, J. Applying operational research and data mining to performance based medical personnel motivation system. *Studies in health technology and informatics*, 2014, vol. 198, 63-70, IOS Press, Inc., ISSN: 0926-9630.
7. Niakšu, O.; Skinulytė, J. & Duhaze, H. G. A Systematic Literature Review of Data Mining Applications in Healthcare. *Workshop proceedings of Web Information Systems Engineering Conference* – WISE 2013, Springer 2014 Lecture Notes in Computer Science, 2014, 313-324, ISSN 0302-9743.

Articles in other peer-reviewed editions:

1. Niakšu, O. & Kurasova, O. Data Mining Applications in Healthcare: Research vs Practice, *Databases and Information Systems BalticDB&IS*, Local Proceedings, 2012, 58–70, ISSN: 1613-0073.
2. Niakšu, O. Calculating distance measure for MRDM clustering. *Proceedings of the 16th International Multi-conference "Information Society – IS 2013"*, 2013, vol. A, 192–194, ISBN: 978-961-264-066-8.

Outline of the Dissertation

The text of the thesis consists of 3 chapters, conclusions, references, the list of publications, and appendixes. Each chapter is provided with an introduction. The total scope of this thesis is 154 pages (without annexes), 44 figures, 22 tables.

Chapter 1 outlines in detail the issues of data mining in medicine and healthcare. In addition, the results of a literature analysis and university hospitals' survey are provided. A novel process model for data mining and knowledge discovery in the medical domain is proposed in Chapter 2. Further, the theoretical part of the proposed process model's evaluation in the fields of Oncology and Cardiology is described. Moreover, a novel multi-relational clustering method, supporting the multi-relational nature of medical data, is proposed. Chapter 3 provides experimental results of the proposed methods. The first two sections present use-cases of the applied methodology for predictive data mining in the Oncology and Cardiology domains. The third section illustrates the usage of multi-relational clustering. The Conclusions section presents the main conclusions of the thesis.

Conclusions

The topics investigated and experimentally proved in the thesis allow us to conclude that:

1. The created data mining application methodology CRISP-MED-DM extends the Cross Industry Standard Process for Data Mining (CRISP-DM) methodology with the following distinguished features:
   - The CRISP-MED-DM methodology, in comparison with other data mining applications methodologies, for the first time outlines the detailed process model specific to the issues and constraints of the medical domain. To achieve that, the initial CRISP-DM process model's five phases were extended with thirty three activities, addressing the issues of medical data pre-processing, semantic interoperability, and patient data privacy protection.
   - The created compliance evaluation model allows for performing the formal assessment of the data mining application projects' compliance to CRISP-MED-DM. The model provides metrics and evaluation formulas to assess the overall quality of application projects and allows for their comparison.
   - The CRISP-MED-DM has been successfully tested in predictive modelling research projects in the oncology and cardiology domains.
2. The accuracy of the created breast cancer susceptibility gene *BRCA1* mutation predictive model has been increased by applying the CRISP-MED-DM methodology:
   - The improvement of the *BRCA1* gene mutation predictive model is as follows: overall accuracy from 0.88 to 0.94, sensitivity from 0.67 to 0.83, specificity from 0.85 to 0.97, ROC AUC from 0.70 to 0.81.
   - The improvement of breast cancer reoccurrence predictive model is as follows: overall accuracy from 0.73 to 0.75, sensitivity from 0.59 to 0.96, specificity has not changed, ROC AUC from 0.63 to 0.65.
3. The developed blood flow echocardiography image analysis technique saves the cardiologist time spent for systolic cycle tracing by extracting a systole cycle curve from standard Doppler ultrasound images and extracting features for further application of predictive data mining methods:
   - The developed software implementation of the proposed echocardiography images analysis technique, compared to the manual measurement of the professional cardiologists resulted in high accuracy for the main aortic valve stenosis diagnostic parameters: maximum aortic valve systolic velocity $AV\ V_{max}$ Pearson coefficient $r(16) = 0.999$ (p-value<0.0001); aortic valve time integral $AV\ VTI$ – $r(16) = 0.988$ (p-value<0.0001); mean peak gradient ΔP_{max} – $r(16) = 0.994$ (p-value<0.0001); aortic valve area AVA – $r(16) = 0.894$ (p-value<0.0001).
   - Applying CRISP-MD-DM with the proposed echocardiography image pre-processing techniques showed that the resulting accuracy is sufficient for practical decision support usage for aortic stenosis grading and diagnosis. The resulting predictive model had 100 % sensitivity and specificity on the research dataset.

4. Partitioning the clustering method with the proposed novel similarity measure allows clustering multi-relational data without its de-normalization and generalization to one-table format:
    - Application of the created similarity measure for PubMed database articles meta-analysis allowed for grouping multi-relational data into clusters with silhouette values 0.21–0.31, which showed better results in comparison with Tree Edit Distance measure results 0.15–0.23, and propositional approach results 0–0.16.
    - The calculation of the distance of each multi-relational object pair is independent and therefore can be successfully parallelized. The developed software implementation of multi-relational clustering supporting parallel calculation allows decreasing similarity measure calculation time in proportion to available processor nodes.

About the Author

Olegas Niakšu was born in Vilnius on the 14th of September in 1979. After finishing Vilnius Karolinos secondary school in 1997, he graduated from Vilnius Gediminas Technical University in 2001 acquiring a bachelor's degree in informatics and in 2003 he acquired a master's degree in informatics engineering. PhD student at Vilnius University Institute of Mathematics and Informatics from 2009 to 2015.

He has a proven professional track record implementing medical information systems and e-health projects in Lithuania, South African Republic, Kazakhstan, Albania and Iran.

Olegas is IPMA certified project manager since 2006, and member of professional organizations ACM and HIMMS since 2009.

E-mail: niaksu@acm.org

Olegas NIAKŠU

DUOMENŲ TYRYBOS METODŲ, SKIRTŲ MEDICININEI DIAGNOSTIKAI IR SVEIKATOS APSAUGOS VADYBAI, VYSTYMAS IR TAIKYMAS

Daktaro disertacijos santrauka

Technologijos mokslai,
Informatikos inžinerija (07 T)

Redaktorė Lina Navickaitė

Olegas NIAKŠU

DEVELOPMENT AND APPLICATION OF DATA MINING METHODS IN MEDICAL DIAGNOSTICS AND HEALTHCARE MANAGEMENT

Summary of Doctoral Dissertation

Technological Sciences,
Informatics Engineering (07 T)

Editor Aurelija Juškaitė