Rīgas Tehniskā universitāte
Datorzinātnes un informācijas tehnoloģijas fakultāte
Informācijas tehnoloģijas institūts

Arnis Kiršners
Doktora studiju programmas "Informācijas tehnoloģija" students

ĪSU LAIKA RINDU UN TO RAKSTUROJOŠO PARAMETRU APSTRĀDES SISTĒMA PROGNOZĒŠANAS UZDEVUMIEM

Promocijas darba kopsavilkums

Zinātniskais vadītājs
profesors *Dr. habil. sc. comp.*
A. BORISOVS

RTU Izdevniecība
Rīga 2015

Kiršners A. Īsu laika rindu un to raksturojošo parametru apstrādes sistēma prognozēšanas uzdevumiem. Promocijas darba kopsavilkums. – R.: RTU Izdevniecība, 2015. – 44 lpp.

Iespiests saskaņā ar DITF ITI padomes 2015. gada 19. jūnija lēmumu Nr. 12100-4.1/4.

Šis darbs izstrādāts ar Eiropas Sociālā fonda atbalstu projektā «Atbalsts RTU doktora studiju īstenošanai».

ISBN 978-9934-10-764-1

PROMOCIJAS DARBS
IZVIRZĪTS INŽENIERZINĀTŅU DOKTORA GRĀDA IEGŪŠANAI
RĪGAS TEHNISKAJĀ UNIVERSITĀTĒ

Promocijas darbs inženierzinātņu doktora grāda iegūšanai tiek publiski aizstāvēts 2016. gada 17. februārī plkst. 14.30 Rīgas Tehniskās universitātes Datorzinātnes un informācijas tehnoloģijas fakultātē, Sētas ielā 1, 202. auditorijā.

OFICIĀLIE RECENZENTI

Profesors *Dr. habil. sc. ing.* Jānis Grundspeņķis
Rīgas Tehniskā universitāte, Latvija

Profesors *Dr. sc. ing.* Egils Stalidzāns
Latvijas Lauksaimniecības universitāte, Latvija

Profesors *Dr. sc. comp.* Pavels Ošmera
Brno Tehnoloģiju universitāte, Čehija

APSTIPRINĀJUMS

Apstiprinu, ka esmu izstrādājis šo promocijas darbu, kas iesniegts izskatīšanai Rīgas Tehniskajā universitātē inženierzinātņu doktora grāda iegūšanai. Promocijas darbs zinātniskā grāda iegūšanai nav iesniegts nevienā citā universitātē.

Arnis Kiršners …….
paraksts

Datums …….

Promocijas darbs uzrakstīts latviešu valodā, tajā ir ievads, četras nodaļas, rezultātu analīze un secinājumi, literatūras saraksts, divi pielikumi, 25 tabulas, 53 attēli, kopā – 144 lappuses. Literatūras sarakstā ir 77 nosaukumi.

SATURS

Darba vispārējs raksturojums ... 5
Problēmsfēra ... 5
Problēmas nostādne ... 5
Pētījuma objekts un priekšmets ... 6
Pētījuma metodes ... 6
Pētījuma ierobežojumi ... 6
Darba mērķis un uzdevumi ... 6
Izvirzītās hipotēzes ... 7
Darba aktualitāte un zinātniskā novitāte ... 7
Darba praktiskais nozīmīgums ... 8
Darba aprobācija ... 8
Darba nodaļu satura apraksts ... 11
1. Īsu laika rindu un to raksturojošo parametru apstrādes pamatprincipi ... 11
Klasterizācijas algoritmu izvēle īsu laika rindu analīzei ... 11
Klasifikācijas algoritmu izvēle raksturojošo parametru analīzei ... 12
Uzdevumu definējums ... 12
Uzdevuma formālā nostādne ... 13
Prognozēšanas sistēmas teorētiskais modelis ... 13
2. Prognozēšanas sistēmas izstrādāšanā lietoto metožu un to modifikāciju apskats ... 14
Klasterizācijas kļūdas aprēķināšana apmācības datu kopai ... 15
Paraugmodeļi klasteros ... 16
Testēšanas datu kopas klasterizācija ... 16
Klasterizācijas kļūdas aprēķins testēšanas datu kopai ... 17
3. Apstrādes sistēmas prognozēšanas uzdevumiem ... 18
3.1. Pieprasījuma prognozēšanas sistēma ... 19
PPS uzbūve ... 19
PPS eksperimentu rezultāti ... 21
3.2. Sirds nekrozes riska prognozēšanas sistēma ... 24
SNRPS uzbūve ... 25
SNRPS eksperimentu rezultāti ... 27
3.3. Baktēriju proliferācijas sindroma noteikšanas sistēma ... 29
BPS noteikšanas sistēmas darbības princips ... 30
BPS noteikšanas sistēmas eksperimentu rezultāti ... 32
4. Prognozēšanas sistēmas izstrādāšanas vadlīnijas ... 34
Prognozēšanas sistēmas struktūra ... 34
Prognozēšanas sistēmas scenārija izvēle ... 35
Prognozēšanas sistēmas izstrādāšana un testēšana ... 36
Rezultāti un secinājumi ... 38
Izmantotās literatūras saraksts ... 41

DARBA VISPĀRĒJS RAKSTUROJUMS

Problēmsfēra

Informācijas tehnoloģijām attīstoties un ienākot mūsu dzīvē, palielinās pieprasījums pēc to lietojuma dažādās problēmvidēs. Informācijas tehnoloģijas arvien biežāk tiek izmantotas sarežģītu uzdevumu risināšanai, kurus agrāk varēja atrisināt tikai ar cilvēka – eksperta klātbūtni. Informācijas tehnoloģiju lietojums ļauj radīt jaunu informāciju vai arī veikt vēsturiskas informācijas analīzi, kuras apstrādes rezultātā tiek izgūtas jaunas zināšanas, kas varētu kalpot par pamatu sarežģītu uzdevumu risinājuma iegūšanai. Pieaugot šo tehnoloģiju izmantošanas intensitātei, palielinās pieprasījums pēc ātra un precīza lēmuma – prognozes saņemšanas pēc iespējas īsākā laika posmā. Pastāv vairākas problēmvides, piemēram, tirdzniecība, medicīna vai farmakoloģija, kurās kā datu avots kalpo laika rindas ar vēsturisku informāciju par īsu laika posmu, kas tiek iegūtas, iedarbojoties uz pētāmo objektu, kā arī aprakstoša veida informācija, kas raksturo šo objektu. Lai varētu apstrādāt šādas atšķirīgas datu struktūras, ir nepieciešams izstrādāt sistēmu, kas spētu prognozēt notikumus nākotnē, balstoties uz datu iegūšanas metodēm un algoritmiem.

Problēmas nostādne

Pastāv problēmvides, kurās par datu avotu izmanto īsas laika rindas un to raksturojošos parametrus. Īsas laika rindas raksturo objekta funkcionālās izmaiņas laika periodā, bet raksturojošie parametri – šā objekta īpašības. Piemēram, medicīnā – pacientam tiek iedotas asinsspiedienu pazeminošas zāles, un 20 minūšu garumā katru minūti tiek mērīts asinsspiediens. Šie asinsspiediena mērījumi laika intervālā būs īsas laika rindas, bet pacienta augums, svars, dzimums, vecums u. c. – raksturojošie parametri. Šāda un līdzīga veida problēmvidēs ir jāatrisina prognozēšanas uzdevums (piemēram, vai asinsspiedienu pazeminošās zāles palīdzēs pacientam), izmantojot tikai analizējamā objekta (piemēram, pacienta) raksturojošos parametrus, lai noteiktu iespējamo prognozējamo vērtību. Šāda dažādu veidu datu struktūru atšķirība pirms darba sākšanas izvirza konkrētas prasības prognozēšana sistēmas izstrādei:

- tai jāspēj apstrādāt diskrēti un nepārtraukti dati; datiem jābūt normalizētiem;
- klasterizācijas procesā, analizējot īsas laika rindas, jānosaka piemērotākais klasteru skaits datu kopas apstrādei, vadoties pēc klasterizācijas kļūdas aprēķina, un jānosaka analizējamā objekta piederība kādam no klasteriem;
- iegūtie klasterizācijas rezultāti jāapvieno ar īsu laika rindu raksturojošajiem parametriem, saglabājot datu integritāti, lai tos tālāk varētu izmantot datu klasifikācijā;
- klasifikācijas procesā jānosaka objekta saikne starp klasterizācijā iegūto klasi un raksturojošajiem parametriem, maksimizējot klasifikācijas precizitāti, jutīgumu un specifiskumu;
- jāinterpretē iegūtie klasifikācijas rezultāti atbilstoši problēmsfērā izvirzītajam uzdevumam;

- uz izveidotās sistēmas bāzes jāveic jauna objekta klasifikācija, nosakot šā objekta prognozējamo vērtību.

Pētījuma objekts un priekšmets

Pētījuma objekts ir prognozēšanas sistēma. Pētījuma priekšmets ir datu iegūšanas un mašīnapmācības metodes un algoritmi.

Pētījuma metodes

Promocijas darbā tiek izmantotas datu analīzes un datu iegūšanas metodes. Datu pirmapstrādes procesā tiek lietota z-novērtējuma normalizācija ar standarta novirzi, pieprasījuma normalizācija ar dzīves līkni. Atribūtu informatīvuma noteikšanai tiek lietota *CfsSubsetEval* metode, bet atribūtu pārmeklēšanai *BestFirst* metode. Klasteru analīzē tiek izmantoti k-vidējo sadalošais, darba autora izstrādāts modificēts k-vidējo sadalošais, maksimālās līdzības un aglomeratīvais hierarhiskais algoritmi. Klasifikācijai tiek lietoti *ZeroR, OneR,* k-tuvāko kaimiņu, *CN2, C4.5,* Naivais Baijesa un *JRip* algoritmi. Klasifikācijas precizitātes novērtēšanai izmantota 10-kārtu šķērsvalidācija, izlaist vienu un datu sadalīšana apmācības un testēšanas kopās attiecībā 70:30. No klasifikatoru iegūtajiem rezultātiem tiek veidoti nosacījumu likumi, apvienojot tos ar problēmvides veikto pētījumu rezultātiem. Klasterizācijas rezultātu klašu struktūru transformācijai uz problēmvidē izmantoto klašu struktūru tiek lietota darba autora izstrādāta pieeja. Analizējamā objekta prognozes vērtības aprēķinam tiek lietots matemātiskās cerības aprēķins un darba autora piedāvātā pieeja, kas balstās uz attālumu metriku.

Pētījuma ierobežojumi

Promocijas darba pētījuma ierobežojumi ir saistīti ar īsu laika rindu analīzi, kas izstrādājamajai prognozēšanas sistēmai uzliek papildu nosacījumus datu iegūšanas metožu un algoritmu izvēlē. Prognozēšanas sistēmai ir jārealizē dažādu datu struktūru apvienošana tā, lai šīs datu struktūras būtu iespējams analizēt ar datu iegūšanas metodēm un algoritmiem.

Darba mērķis un uzdevumi

Promocijas darba mērķis ir izstrādāt īsu laika rindu un to raksturojošo parametru apstrādes sistēmu prognozēšanas uzdevumiem, kas būtu lietojama dažādās problēmvidēs un balstītos uz datu iegūšanas metodēm un algoritmiem. Izvirzītā mērķa sasniegšanai nepieciešams realizēt šādus uzdevumus:

1. izanalizēt īsu laika rindu un to raksturojošo parametru apstrādes pamatprincipus;
2. izpildīt īsu laika rindu un to raksturojošo parametru datu pirmapstrādes pieeju analīzi;
3. modificēt atbilstoši problēmvidei klasterizācijas algoritmu, lai tas apstrādātu īsas laika rindas, un salīdzināt to ar citiem klasterizācijas algoritmiem;
4. izstrādāt iegūto klasterizācijas rezultātu un īsu laika rindu raksturojošo parametru datu apvienošanas pieeju;

5. izstrādāt prognozēšanas sistēmu dažādām problēmvidēm, kas apstrādā īsas laika rindas un to raksturojošos parametrus un izdara prognozi, pamatojoties tikai uz sistēmā ievadītajiem analizējamā objekta raksturojošajiem parametriem;
6. novērtēt izstrādātās prognozēšanas sistēmas precizitāti dažādās problēmvidēs;
7. izstrādāt nosacījumu likumu veidošanas un lietošanas pieejas dažādām problēmvidēm;
8. pamatojoties uz izveidoto prognozēšanas sistēmu dažādām problēmvidēm, izstrādāt vadlīnijas līdzīgu sistēmu izstrādāšanai.

Izvirzītās hipotēzes

Promocijas darba izstrādāšanas gaitā tiek izvirzītas vairākas hipotēzes.

1. Īsu laika rindu un to raksturojošo parametru apstrādes sistēmas izstrādāšana nodrošina grūti formalizējama uzdevuma atrisināšanu ar datu iegūšanas metodēm un algoritmiem.
2. Modificēts k-vidējo sadalošais algoritms uzlabo piemērotākā klasteru skaita noteikšanu datu klasterizācijas procesā, analizējot īsas laika rindas.
3. Izstrādātā datu apstrādes sistēma realizē prognozēšanas uzdevumu izpildi dažādās problēmvidēs.

Pirmā hipotēze norāda, ka pastāv datu struktūras, kurās par datu avotu izmanto īsas laika rindas un to raksturojošos parametrus. Hipotēze tiks uzskatīta par apstiprinātu, ja būs radīta sistēma, kas spēj apstrādāt šīs atšķirīgās datu struktūras un rezultātā tiks iegūta risināmā uzdevuma prognoze.

Otro hipotēzi raksturo tas, ka tiks pārbaudīti un salīdzināti vairāki klasterizācijas algoritmi dažādās problēmvidēs, kā tie spēj apstrādāt datus ar īsām laika rindām. Ja modificētais klasterizācijas algoritms uzrādīs labākus rezultātus, hipotēze būs uzskatāma par apstiprinātu.

Trešā hipotēze pamatojas idejā, ka izstrādātā sistēma ir lietojama vai adaptējama dažādām problēmvidēm. Hipotēze tiks uzskatīta par apstiprinātu, ja izstrādātā sistēma būs izmantota vairāku prognozēšanas uzdevumu atrisinājumam dažādās problēmvidēs.

Darba aktualitāte un zinātniskā novitāte

Promocijas darba aktualitāte ir saistīta ar atšķirīga veida (īsas laika rindas un to raksturojošie parametri) datu struktūru analīzi. Nav zināma metode vai algoritms, kas varētu veikt īsu laika rindu un to raksturojošo parametru analīzi. Tāpēc ir svarīgi noteikt pieeju kopumu, kas lieto datu iegūšanas metodes uz algoritmus, lai varētu apstrādāt īsas laika rindas un to raksturojošos parametrus. Īsas laika rindas tiek piedāvāts apstrādāt, izmantojot klasterizāciju, lai noteiktu līdzīgas objektu grupas. Raksturojošie parametri tiek apstrādāti, lietojot klasifikāciju, lai atrastu sakarības starp šiem parametriem un klasterizācijā iegūtajiem rezultātiem. Pētāmā objekta raksturlieluma iespējamā vērtība tiek noteikta, klasificējot šā objekta raksturojošos parametrus uz izveidotā klasifikatora bāzes.

Promocijas darba zinātniskā novitāte ir izstrādātā prognozēšanas sistēma dažādām problēmvidēm, kas realizē īsu laika rindu un to raksturojošo parametru apstrādi. Izstrādātā sistēma spēj analizēt sarežģītas datu struktūras. Sistēmā izstrādāts:

1. modificēts k-vidējo sadalošais algoritms, kas nodrošina īsu laika rindu apstrādi dažādās problēmvidēs, nosakot piemērotāko klasteru skaitu pēc klasterizācijas vidējās absolūtās kļūdas novērtējuma;
2. divu atšķirīgu datu struktūru apvienošanas pieeja dažādās problēmvidēs;
3. sistēmas iegūto klasifikācijas rezultātu atspoguļojums dažādām problēmvidēm un risināmajiem uzdevumiem;
4. nosacījumu likumu veidošanas un lietošanas pieejas dažādām problēmvidēm.

Darba praktiskais nozīmīgums

Promocijas darba praktiskais nozīmīgums ir izstrādātā pieprasījuma prognozēšanas sistēma, kas realizē jaunas preces iespējamo pieprasījumu prognozi nākamajam periodam, sistēmā ievadot tikai preces raksturojošos parametrus. Šo sistēmu iespējams lietot uzņēmumos, kuros ir nepieciešamība prognozēt iespējamo preces pieprasījumu nākamajiem periodiem.
Izstrādāta sirds nekrozes riska prognozēšanas sistēma, kas nosaka sirds nekrozes risku laboratorijas dzīvniekam, sistēmā ievadot tikai šā dzīvnieka raksturojošo informāciju. Izstrādāto sistēmu var lietot pētniecības iestādēs, kurās tiek izmantoti laboratorijas dzīvnieki, lai prognozētu pētāmās vielas ietekmi.
Izstrādāta baktēriju proliferācijas sindroma noteikšanas sistēma, kas nosaka, vai indivīdam ir nepieciešamība veikt laktulozes testu, sistēmā ievadot tikai indivīda pašsajūtas novērtējuma parametrus. Izveidotā sistēma lietojama medicīnā – gastroenteroloģijā, nosakot baktēriju proliferācijas sindromu tievajā zarnā, prognozējot laktozes testa nepieciešamību.
Izpildīts izstrādātās sistēmas precizitātes novērtējums dažādās problēmvidēs.
Izstrādātas prognozēšanas sistēmas vadlīnijas, kas sniedz ieteikumus izstrādātājam līdzīgu sistēmu izveidošanai dažādās problēmvidēs.

Darba aprobācija

Darba izstrādāšanas gaitā tika sagatavotas un publicētas 13 publikācijas.

1. Parshutin S., Kirshners A. Research on Clinical Decision Support Systems Development for Atrophic Gastritis Screening// *Expert Systems with Applications.* – 2013. – Vol. 40, Iss.15, pp. 6041-6046. Citēts: ScienceDirect, SCOPUS, Thomson Reuters ISI Web of Science.
2. Kirshners A., Parshutin S. Application of Data Mining Methods in Detecting of Bacteria Proliferation Syndrome in the Small Intestine // In: *European Conference on Data Analysis 2013: Book of Abstracts: European Conference on Data Analysis 2013.* – 2013. – pp. 139–139.
3. Kirshners A., Parshutin S., Leja M. Research in application of data mining methods to diagnosing gastric cancer// LNAI 7377. Proceedings of the 12th Industrial Conference on Data Mining ICDM'2012. – 2012. – pp. 24–37. Citēts: SpringerLink, SCOPUS.
4. Kirshners A., Liepinsh E., Parshutin S., Kuka J., Borisov A. Risk Prediction System for Pharmacological Problems// Automatic Control and Computer Sciences. – 2012. –Vol. 46, No. 2. – pp. 57–65. Citēts: SpringerLink, SCOPUS.

5. Kirshners A., Borisov A. A Comparative Analysis of Short Time Series Processing Methods// Scientific Journal of Riga Technical University, Information Technology and Management Science, – 2012. – Vol.15. – pp. 65–69. Citēts: VINITI, EBSCO, CSA/ProQuest.
6. Kirshners A., Borisov A., Parshutin S. Robust Cluster Analysis in Forecasting Task// Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012). – 2012. – pp. 77–81.
7. Parshutin S., Kirshners A. Intelligent Agent Technology in Modern Production and Trade Management// Efficient Decision Support Systems: Practice and Challenges – From Current to Future/ Book Chapter. INTECH. – 2011. – pp. 21–42. Citēts: NetLibrary; Scirus; IntechOpen; WorldCat.
8. Kirshners A. Clustering-based Behavioural Analysis of Biological Objects// Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. – 2011. – Vol. 2. – pp. 24–32. Citēts: SCOPUS.
9. Kirshners A., Borisov A. Multilevel Classifier Use in a Prediction Task// Proceedings of the 17th International Conference on Soft Computing. – 2011. – pp. 403–410. Citēts: Thomson Reuters ISI Web of Science.
10. Kirshners A., Borisov A. Processing short time series with data mining methods// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2011. – Iss. 5, Vol. 49. – pp. 91–96. Citēts: VINITI, EBSCO, CSA/ProQuest.
11. Kirshners A., Parshutin S., Borisov A. Combining clustering and a decision tree classifier in a forecasting task// Automatic Control and Computer Science. – 2010. – Vol. 44, No. 3. – pp. 124–132. Citēts: SpringerLink, SCOPUS.
12. Kirshners A., Borisov A. Analysis of short time series in gene expression tasks// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2010. – Iss. 5, Vol. 44. – pp. 144–149. Citēts: VINITI, EBSCO, CSA/ProQuest.
13. Kirshners A., Kuleshova G., Borisov A. Demand forecasting based on the set of short time series// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2010. – Iss. 5, Vol. 44. – pp. 130–137. Citēts: VINITI, EBSCO, CSA/ProQuest.

Pētījumu un sasniegumu rezultāti ar mutiskiem ziņojumiem prezentēti deviņās konferencēs.

1. *European Conference on Data Analysis 2013,* Luksemburga, Luksemburga, 2013. gada 10.–12. jūlijs.
2. *53rd International Scientific Conference of Riga Technical University,* RTU, Rīga, Latvija, 2012. gada 10.–12. oktobris.
3. *12th Industrial Conference on Data Mining ICDM'2012,* Berlīne, Vācija, 2012. gada 13.–20. jūlijs.
4. *5th International Conference on Applied Information and Communication Technologies AICT2012,* LLU, Jelgava, Latvija, 2012. gada 26.–27. aprīlis.
5. *52nd International Scientific Conference of Riga Technical University,* RTU, Rīga, Latvija, 2011. gada 12.–25. oktobris.

6. *8th International Scientific and Practical Conference,* Rēzekne, Latvija, 2011. gada 20.–22. jūnijs.
7. *17th International Conference on Soft Computing,* Brno, Čehija, 2011. gada 15.–17. jūnijs.
8. *Informācijas tehnoloģija: Zinības un prakse, Latvijas universitāte,* Rīga, Latvija, 2010. gada 7. decembris.
9. *51st International Scientific Conference of Riga Technical University,* RTU, Rīga, Latvija, 2010. gada 11.–15. oktobris.

Darba rezultāti un sasniegumi izmantoti šādos projektos:

06.12.–03.13. – RTU ZP projekts «Jaunajiem zinātniekiem 2012/2013». «Lēmuma atbalsta sistēmas kuņģa vēža diagnosticēšanas uzdevuma pētījums». Vadītājs: S. Paršutins.

01.10.–12.12. – LU projekts 2009/0220/1DP/1.1.1.2.0/09/APIA/VIAA/016 «Agrīnas audzēju diagnostikas un novēršanas starpdisciplināra izpētes grupa». Vadītājs profesors M. Leja.

06.10.–12.11. – Latvijas-Baltkrievijas sadarbības programma zinātnē un tehnikā, līgums Nr. L7631, «Medicīnisko un bioloģisko datu intelektuālo metožu un apstrādes algoritmu kompleksa izstrāde onkoloģisko slimību diagnostikas pilnveidošanai». Vadītājs: profesors A. Borisovs.

Promocijas darbā ir ievads, četras nodaļas, rezultātu analīze un secinājumi, literatūras avotu saraksts un pielikumi.

Pirmajā nodaļā sniegti īsu laika rindu un to raksturojošo parametru apstrādes pamatprincipi. Dots ieskats datu iegūšanā. Apskatīta teorija par īsām laika rindām un to raksturojošajiem parametriem. Pamatota algoritmu izvēle promocijas darbā risināmo prognozēšanas uzdevumu risinājumu realizācijai. Izvirzīta formalizētā uzdevuma nostādne.

Otrajā nodaļā aprakstītas lietotās metodes un algoritmi, kas izmantoti prognozēšanas sistēmu izstrādāšanā. Nodaļā aprakstīta darba autora izstrādātā k-vidējo sadalošā algoritma modifikācija, kas paredzēta īsu laika rindu klasterizācijai ar dažādu objektu skaitu.

Trešajā nodaļā piedāvāta prognozēšanas sistēma dažādām problēmvidēm, uz kuras pamata izstrādātas: pieprasījuma prognozēšanas sistēma, sirds nekrozes riska prognozēšanas sistēma un baktēriju proliferācijas sindroma noteikšanas sistēma. Apskatīti šo sistēmu uzbūves un darbības pamatprincipi. Aprakstīti veiktie eksperimenti un iegūtie rezultāti. Veikts izstrādāto sistēmu precizitātes novērtējums. Izdarīti secinājumi par izstrādātajām prognozēšanas sistēmām.

Ceturtajā nodaļā sniegts izklāsts par prognozēšanas sistēmas izstrādāšanas vadlīnijām, kuras palīdz izstrādātājam izvēlēties piemērotāko sistēmas izveidošanas procesu īsu laika rindu un raksturojošo parametru apstrādāšanai. Sistēmas izveidošanas process balstīts pieredzē, kas iegūta preču pieprasījuma, sirds nekrozes riska un baktēriju proliferācijas sindroma prognozēšanas sistēmu izstrādāšanā dažādās problēmvidēs.

Pēdējā nodaļā veikta sasniegto rezultātu analīze un izdarīti secinājumi par izstrādātajām sistēmām un vadlīnijām.

DARBA NODAĻU SATURA APRAKSTS

1. ĪSU LAIKA RINDU UN TO RAKSTUROJOŠO PARAMETRU APSTRĀDES PAMATPRINCIPI

Nodaļā iepazīstināts ar datu iegūšanu [2, 3, 4, 5, 44, 45, 52, 58, 59, 62, 75]. Aprakstīti datu iegūšanā risināmie uzdevumi un jaunu zināšanu iegūšanas process [76]. Izklāstītas īsu laika rindu un to raksturojošo parametru apstrādāšanas iespējas ar datu iegūšanas metodēm un algoritmiem [7, 9,11, 21, 22, 29, 31, 32, 33, 36, 42, 53, 54, 55, 57, 63, 64, 72, 70, 76].

Klasterizācijas algoritmu izvēle īsu laika rindu analīzei

Lai izvēlētos piemērotākos klasterizācijas algoritmus īsu laika rindu analīzei, tika veikta vairāku algoritmu salīdzinošā analīze, izvēloties dažādus kritērijus: vai algoritms ietilpst 10 populārāko algoritmu sarakstā [73]; algoritma iegūto rezultātu interpretējamība, šeit pozitīvu vērtējumu iegūst tie algoritmi, kuru iegūtie rezultāti ir interpretējami bez eksperta klātbūtnes un to realizācija ir pieejama eksperimentos izmantotajā programmatūra. Salīdzinošie kritēriji tiek vērtēti ar divām vērtībām: pozitīvs vērtējums (+) vai negatīvs vērtējums (–). Klasterizācijas algoritmu salīdzinošās analīzes novērtējuma rezultāti ir parādīti 1. tabulā. Pozitīvo vērtējumu skaita summa ir norādīta kolonā novērtējums.

1. tabula

Klasterizācijas algoritmu salīdzinošā analīze

	Salīdzināšanas kritēriji, iespējamās vērtības (+ vai –)					
Klasterizācijas algoritmi	Vai ir *Top10* algoritms (vieta topā)	Rezultātu interpretējamība	Realizācijai lietotā programmatūra			Novērtējums
			Weka	*Orange Canvas*	*Statistica*	
K-vidējo sadalošais	+(3)	+	+	+	+	**5**
Maksimālās līdzības (*EM*)	+(5)	+	+	-	+	**4**
Pašorganizējošie neironu tīkli (*SOM*)	–	–	–	+	–	1
Aglomeratīvais hierarhiskais	–	+	+	+	–	3
C-vidējo sadalošais	–	+	–	–	–	1

No salīdzinošās analīzes izriet, ka piemērotākie klasterizācijas algoritmi īsu laika rindu analīzei būtu k-vidējo sadalošais un maksimālās līdzības algoritmi. Dažos eksperimentos būtu lietderīgi pārbaudīt arī aglomeratīvā hierarhiskā algoritma lietošanas iespējamību, tā kā šis algoritms ir nākamais sarakstā pēc diviem jau izvēlētajiem algoritmiem.

Klasifikācijas algoritmu izvēle raksturojošo parametru analīzei

Izvēloties klasifikācijas algoritmu, ir jāņem vērā, kā tiks interpretēti iegūtie rezultāti, vai ir nepieciešams caurskatāms un viegli saprotams iegūto rezultātu skaidrojums. Kas izmantos iegūtos rezultātus – eksperts vai ierindas sistēmas lietotājs? Cik ātri ir iegūstams algoritma izpildes rezultāts? Vai algoritms ir piemērots analizējamajai datu struktūrai? Atbildot uz šiem jautājumiem, ir iespējams izvēlēties vairākus klasifikatorus un eksperimentāli noteikt piemērotāko problēmvides uzdevuma risinājumu. Klasifikācijas algoritmu izvēle tika veikta pēc salīdzinošās analīzes, kas parādīta 2. tabulā.

2. tabula

Klasifikācijas algoritmu salīdzinošā analīze

	Salīdzināšanas kritēriji, iespējamās vērtības (+ vai –)				
Klasifikācijas algoritmi	Vai ir *Top10* algoritms (vieta topā)	Rezultātu interpretējamība (lietotā metode)	Realizācijai lietotā programmatūra		Novērtējums
			Weka	*Orange Canvas*	
C4.5	+(1)	+ (induktīvie lēmuma koki)	+	+	**4**
k-tuvāko kaimiņu *(kNN)*	+(8)	+ (attāluma metrika)	+	+	**4**
Naivais Baijesa	+(9)	+ (varbūtības)	+	+	**4**
CN2	–	+ (nosacījumu likumi)	+	+	**3**
OneR	–	+ (nosacījumu likumi)	+	–	2
ZeroR	–	+ (nosacījumu likumi)	+	–	2

Uzdevumu definējums

Ir doti problēmvides dati, kas apraksta informāciju par vēsturiskiem notikumiem un to raksturojošajiem parametriem. Ar piedāvātām datu iegūšanas metodēm un algoritmiem tiek risināts prognozēšanas uzdevums, kam jānosaka analizējamā objekta vērtība nākotnē, balstoties tikai uz šā objekta raksturojošajiem parametriem.

Lietojot datu iegūšanas metodes un algoritmus, ir jānosaka piemērotākie, ar kuriem ir iespējams atrisināt problēmvidē izvirzītos uzdevumus. Datu iegūšanas pieeju kopums, kas lietojams konkrētu uzdevumu izpildei, ir šāds:

1. vēsturisku notikumu (īsas laika rindas) un to raksturojošo parametru (atribūtu) pirmapstrāde;
2. likumsakarību noteikšana vēsturiskos datos, veidojot klasterizācijas modeli, ar kura palīdzību tiek noteiktas līdzīgas objektu grupas – klasteri; atkarībā no problēmvides

izvirzītā uzdevuma, iespējams veidot iegūto līdzīgo objektu grupu attēlojumu ar paraugmodeļu palīdzību;
3. klasterizācijas rezultātā iegūto līdzīgo objektu grupu skaita transformācija uz raksturojošo parametru atkarīgā mainīgā (klašu) skaitu;
4. likumsakarību noteikšana ar klasifikācijas palīdzību starp raksturojošajiem parametriem un klasterizācijas rezultātiem;
5. klasifikācijas rezultātu interpretācija ar nosacījumu likumiem;
6. klasterizācijas un klasifikācijas precizitātes pārbaude;
7. pētāmā objekta raksturlieluma noteikšana, lietojot tikai šā objekta raksturojošos parametrus.

Uzdevuma formālā nostādne

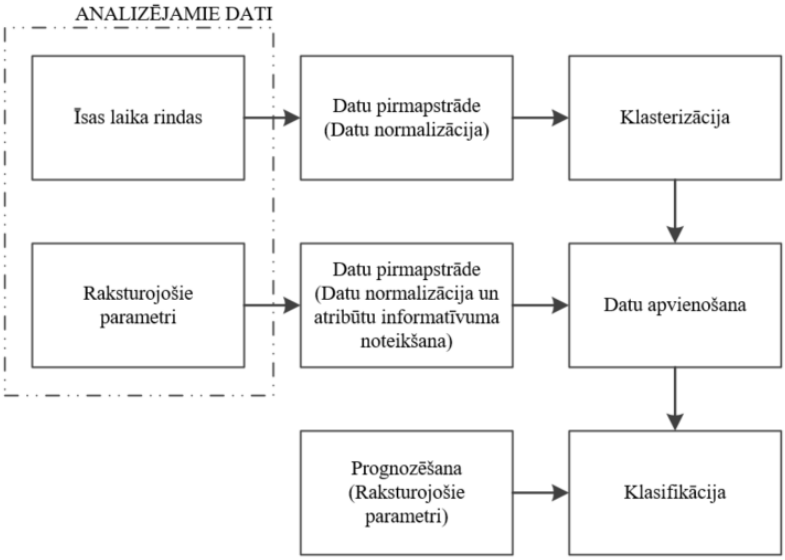
Īsu laika rindu apstrādes process darbā formalizēts kā klasterizācijas uzdevums, kura mērķis ir noteikt līdzīgas objektu grupas, pēc kurām būtu iespējams grupēt objektus analizējamajā datu kopā. Lai pārliecinātos par izvēlēto metožu piemērotību, nepieciešams izpildīt īsu laika rindu klasterizācijas algoritmu salīdzinošo analīzi un izvēlēties piemērotāko. Ir klasterizācijas uzdevumi, kad izvēlētās metodes tikai daļēji nodrošina šo uzdevumu atrisināšanu. Lai pilnībā atrisinātu šos uzdevumus, nepieciešams izpētīt izvēlētos algoritmus un izstrādāt to modifikācijas īsu laika rindu klasterizācijai. Klasterizācijai tiek izmantota apmācība bez skolotāja.

Īsu laika rindu raksturojošo parametru apstrādes process darbā formalizēts kā klasifikācijas uzdevums, kura mērķis ir apmācības procesā ar skolotāju, atrast likumsakarības starp raksturojošajiem parametriem, to vērtībām un mērķa atribūtiem – klasēm, testēšanas procesā noteikt izmantotā klasifikācijas algoritma precizitāti, un, lietojot izveidoto klasifikācijas modeli jeb klasifikatoru, prognozēt jaunā analizējamā objekta mērķa atribūtu uz apmācības procesā izveidotā modeļa bāzes.

Prognozēšanas sistēmas teorētiskais modelis

Pamatojoties uz uzdevuma formālo nostādni, analizējamo datu struktūru un literatūras analīzi, tiek piedāvāts prognozēšanas sistēmas teorētiskais modelis, kas balstīts uz datu iegūšanas metodēm un algoritmiem (skat. 1. attēlu). Analizējamajiem datiem, kas sastāv no diviem datu veidiem, īsām laika rindām un to raksturojošajiem parametriem, tiek piedāvāts veikt datu pirmapstrādi. Datu pirmapstrādes procesā notiek to objektu izslēgšana no analizējamās datu kopas, kam ir trūkstošas vērtības. Datu normalizāciju lieto, lai izvairītos no atribūtu dominances datu kopā, kas rodas no dažādiem atribūtu vērtību intervālu diapazoniem [1]. Raksturojošajiem parametriem lieto atribūtu informatīvuma noteikšanu, kas palīdz izslēgt no datu kopas neinformatīvus atribūtus, kas negatīvi var ietekmēt klasifikācijas precizitāti [27]. Klasterizācijas procesā objekti ar īsām laika rindām tiek apvienoti grupās – klasteros pēc līdzības pazīmēm, lietojot klasterizācijas algoritmu [64]. Iegūtos klasterizācijas rezultātus apvieno ar raksturojošajiem parametriem, tad ar klasifikācijas palīdzību nosaka sakarības starp

raksturojošajiem parametriem (atribūtiem) un klasterizācijas procesā noteikto klasi, lietojot kādu no klasifikācijas algoritmiem [27, 64].



1. att. Prognozēšanas sistēmas teorētiskais modelis

Jauna analizējamā objekta mērķa atribūta jeb klases (prognozēšanas uzdevums) noteikšana tiek veikta, pamatojoties tikai uz šā objekta raksturojošajiem parametriem, lietojot izveidoto klasifikācijas modeli [27, 58, 64, 70].

2. PROGNOZĒŠANAS SISTĒMAS IZSTRĀDĀŠANĀ LIETOTO METOŽU UN TO MODIFIKĀCIJU APSKATS

Nodaļā apskatītas prognozēšanas sistēmu izstrādāšanā izmantotās datu iegūšanas metodes, algoritmi, algoritmu modifikācijas un aplūkoti to darbības principi. Aprakstītas sistēmu izstrādāšanā izmantotās datu pirmapstrādes tehnoloģijas: datu attīrīšana un datu transformācija [1, 27, 57, 58, 63, 64, 76]. Sniegta informācija par datu klasterizāciju, izklāstīti tādi klasterizācijas algoritmi – k-vidējo sadalošais [28, 63], maksimālās līdzības [17, 51], aglomeratīvais hierarhiskais [63] – un darba autora piedāvātais modificētais k-vidējo sadalošais.

Modificētais k-vidējo sadalošais algoritms tika izstrādāts, balstoties uz eksperimentu rezultātiem ar īsām laika rindām [30, 35] dažādās problēmvidēs, kas pierādīja, ka, piemēram, klasiskais k-vidējo sadalošais algoritms nespēj noteikt piemērotāko klasteru skaitu, analizējot īsas laika rindas [34, 37]. K-vidējo sadalošā algoritma iegūtie rezultāti norādīja uz vienotu tendenci – palielinoties klasteru skaitam, samazinās kopējā klasterizācijas kļūda, bet, piemēram, maksimālās līdzības algoritms tieši pretēji mazāko kļūdu noteica pie minimālā klasteru skaita. Tāpēc tika izstrādāta k-vidējo sadalošā algoritma modifikācija, kas spēj novērst minētās nepilnības.

Sākotnēji tiek ieviests jēdziens – maksimālais klasteru skaits, ar kādu tiks klasterizēta analizējamā datu kopa. Šāda pieeja nodrošina algoritma ātrdarbības palielināšanos atšķirībā no klasiskā k-vidējo sadalošā algoritma, kur manuāli tiek noteikts klasteru skaits, līdz kuram

algoritms veic objektu sadalīšanu [73]. Veicot klasterizāciju ar kādu no algoritmiem, jānosaka meklējamo klasteru diapazons (no minimālā līdz maksimālajam skaitam, kur minimālais skaits parasti ir 2), lai klasterizācijas process būtu efektīvs un neaizņemtu daudz laika resursu. Jāatrod piemērotākais klasteru skaits diapazonā no 2 līdz maksimālajam klasteru skaitam. Šim maksimumam jābūt pietiekami lielam, lai precīzi (iespējams noteikt pēc klasterizācijas vidējās absolūtās kļūdas vai kvadrātiskās kļūdas) veiktu klasterizāciju datu kopā, bet tas nevar būt arī pārlieku liels. Šādā gadījumā pastāv iespēja nepareizai rezultātu interpretācijai, jo, piemēram, izveidotu klasteri algoritms cenšas sadalīt vēl mazākos klasteros. Tāpēc maksimālā klasteru skaita C_{max} aprēķinam izmanto teorētisko pieņēmumu, kas tiek aprēķināts pēc formulas [64, 70]: $C_{\max} = \sqrt{n}$, kur n – objektu skaits analizējamajā datu kopā. Turpinājumā modificētais k-vidējo sadalošais algoritms lieto klasiskā k-vidējo sadalošā algoritma metodiku, līdz algoritms sasniedz summētās kvadrātiskās kļūdas aprēķinu, tad notiek algoritma modifikācija. Katram klasterim tiek iegūta attālumu matrica, kas raksturo objektu (laika rindas {T1, T2, …, T12}) c_n attālumu d_n (iegūts, izmantojot *Eiklīda* attāluma mēru) līdz tuvākajam centroīdam. No iegūtās attālumu matricas rezultātiem tiek aprēķināta klasterizācijas kļūda apmācības datu kopai, izmantojot vidējās absolūtās novirzes aprēķinu katrā klasterī un kopējās klasterizācijas kļūdas aprēķinu. Vadoties pēc minimālās aprēķinātās kopējās klasterizācijas kļūdas vērtības katrā no analizētajiem klasteriem, tiek noteikts piemērotākais klasteru skaits, kas nepieciešams analizējamās datu kopa klasterizācijai.

Ja analizējamās datu kopas objektu skaits nav lielāks par 200, klasterizācijas precizitātes novērtējumam jāizmanto 10-kārtas šķērsvalidācija [43] un klasterizācijas kļūdas aprēķināšana ir jāveic saskaņā ar apakšpunktu «Klasterizācijas kļūdas aprēķināšana apmācības datu kopai», jo, sadalot datu kopu apmācības un testēšanas kopās, objektu skaits tajās būs neliels, kas var iespaidot klasterizācijas rezultātus. Savukārt, ja ierakstu skaits ir lielāks par 200, precizitātes novērtējumam var izmantot kopas sadalīšanu apmācības un testēšanas datu kopās ar proporcijas attiecību 70 % pret 30 % [70]. Šajā gadījumā arī ir jāveic klasterizācijas kļūdas aprēķināšana saskaņā ar apakšpunktu «Klasterizācijas kļūdas aprēķināšana apmācības datu kopai» un jāizveido paraugmodeļi katram noteiktajam klasterim pēc apakšpunkta «Paraugmodeļi klasteros». Tad jāizpilda testēšanas datu kopas klasterizācija atbilstoši apakšpunktam «Testēšanas datu kopas klasterizācija», savukārt klasterizācijas kļūdas aprēķins testēšanas datu kopai tiek veikts saskaņā ar apakšpunktu «Klasterizācijas kļūdas aprēķins testēšanas datu kopai».

Klasterizācijas kļūdas aprēķināšana apmācības datu kopai

Vidējā absolūtā kļūda tiek noteikta pēc formulas (1), iegūstot vidējās absolūtās novirzes AD_i vērtību:

$$AD_i = \frac{d_1 + d_2 + \ldots + d_n}{c_n}, \tag{1}$$

kur $d_1, d_2, \ldots, d_n$ – attiecīgā objekta attālums līdz centroīdam;

c_n – objektu skaits klasterī;

AD_i – vidējā absolūtā novirze i-tajā klasterī.

Pēc tam summē katrā i-tajā klasterī iegūtās vidējās absolūtās novirzes vērtības AD_i un izdala ar noteikto klasteru skaitu C_i analizējamajā datu kopā, iegūstot klasterizācijas vidējo absolūto kļūdu *MeanAE* [52] pēc formulas (2):

$$MeanAE = \frac{AD_1 + AD_2 + \ldots + AD_i}{C_i}, \tag{2}$$

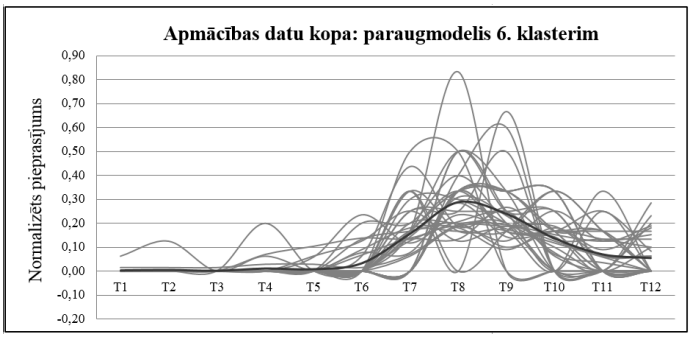
kur C_i – kopējais noteiktais klasteru skaits datu kopā;

MeanAE – klasterizācijas vidējā absolūtā kļūda.

Šāda veida pieeja realizē visu datu kopā ietilpstošo objektu attālumu līdz centroīdiem analīzi katrā no klasteriem, tādējādi nodrošinot klasterizācijas vidējās absolūtās kļūdas aprēķinu, ar kura palīdzību, pēc mazākās iegūtās *MeanAE* vērtības starp klasteriem C_i nosaka to klasteru skaitu, kas būtu piemērotākais analizējamās datu kopas klasterizācijai.

Paraugmodeļi klasteros

Paraugmodelis jeb prototips tiek izveidots katram klasterim, kas noteikts klasterizācijas rezultātā. Paraugmodeļu skaits ir atkarīgs no noteiktā piemērotāko klasteru skaitu. Paraugmodeļa iegūtā līkne, kas izveidota no 6. klastera objektu vidējām vērtībām katrā laika periodā, parādīta 2. attēlā, ar pelēkām līnijām – parādīti šā klastera objekti. Uz *x* ass norādīti periodu numuri, uz *y* ass – īsu laika rindu normalizētās vērtības. Iegūtie paraugmodeļi raksturo klastera objektu uzvedību noteiktā laika periodā.



2. att. Izveidotais paraugmodelis (attēlots ar treknu līniju) 6. klasterim

Testēšanas datu kopas klasterizācija

Testēšanas datu kopas klasterizācija balstās uz apmācības kopas klasterizācijā iegūtajām paraugmodeļa vidējām vērtībām un testēšanas kopas (īsas laika rindas) datiem. Objekta attālumu līdz attiecīgā klastera vidējai vērtībai nosaka pēc formulas (3). Iegūtie rezultāti tiek fiksēti attālumu matricā, kas parādīta 3. tabulā.

Testēšanas datu kopas klasterizācijā katram analizējamajam objektam tiek noteikta piederība klasterim pēc minimālā attāluma [70]. Tiek ņemts katrs testēšanas kopas objekts pa periodiem

un izskaitļots tā attālums līdz katra klastera centroīdam (apmācības kopas paraugmodelim) pēc formulas (3):

$$d_{i,j} = \sqrt{\left(P_1 - C_1\right)^2 + \left(P_2 - C_2\right)^2 + \ldots + \left(Pm_n - Cz_n\right)^2}, \tag{3}$$

kur $d_{i,j}$ – testēšanas datu kopas objekta *i* attālums līdz paraugmodelim *j*;

P_m – objekta *m* vērtība testēšanas datu kopā periodā *n*;

C_z – no apmācības datiem iegūtā *z* paraugmodeļa vidējā vērtība periodā *n*.

Kad veikti testēšanas datu kopas objektu attālumu mēru aprēķini, kuru rezultāti atspoguļoti 3. tabulā, jānosaka katra objekta minimālais attālums katrā rindā $d_{i.}$ un iegūtā vērtība jāreģistrē tabulas laukā minimālais attālums.

3. tabula

Testēšanas datu kopas attālumu matrica klasteru piederības noteikšanai

Objekta numurs	Objekta attālums līdz attiecīgā klastera centram (paraugmodelim)				Minimālais attālums	Piešķirtā klase pēc minimālā attāluma
Nr.	**C1**	**C2**	**…**	**Cj**	**d_i(min)**	**C(j)**
d_1	$d_{1,1}$	$d_{1,2}$	…	$d_{1,j}$		
d_2	$d_{2,1}$	$d_{2,2}$	…	$d_{2,j}$		
…	…	…	…	…		
d_i	$d_{i,1}$	$d_{i,2}$	…	$d_{i,j}$		

Pēc iegūtā minimālā attāluma, kas noteikts katram testēšanas datu kopas objektam, ir jāpiešķir klastera numurs *C(j)*, kuram no *j* klasteriem analizējamais objekts atrodas vistuvāk.

Klasterizācijas kļūdas aprēķins testēšanas datu kopai

Lai varētu spriest par testēšanas kopas klasterizācijas rezultātiem, nepieciešams veikt novērtējumu. To var izdarīt, aprēķinot vidējo absolūto novirzi *MAD* (*mean absolute deviation*) un vidējo absolūto kļūdu *MAE* (*mean absolute error*) katrā klasterī, pēc tam summējot katrā klasterī aprēķinātās vidējās absolūtās kļūdas vērtības. Tādā veidā tiek panākts novērtējums starp neatkarīgām datu kopām. *MAD* [52, 76] aprēķina pēc formulas (4):

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\left|\left(Cvid_{n_i} - P_{m_i}\right)\right|, \tag{4}$$

kur $Cvid_n$– iegūtā apmācības datu kopas paraugmodeļa vidējā vērtība klasterī *n* periodā *i*;

P_m – reāla testēšanas datu kopas objekta *m* vērtība periodā *i*;

N – periodu skaits laika rindā.

Vērtība *MAE* [52, 76] klasterī *i*, kas norāda uz kopējo kļūdu klasterī attiecībā pret apmācības datu kopu, aprēķina pēc formulas (5):

$$MAE_i = \frac{MAD}{k}, \tag{5}$$

kur k – objektu skaits testēšanas datu kopā klasterī i.

Savukārt kopējo absolūto kļūdu *TAE* (*total absolute error*) testēšanas datu kopai aprēķina [52, 76] pēc formulas (6):

$$TAE = \frac{\sum_{i=1}^{n} MAE_i}{n}, \tag{6}$$

kur MAE_i – vidējā absolūtā kļūda *i-tajā* klasterī.

Iegūstot kopējās absolūtās kļūdas novērtējumu, klasterizējot testēšanas datu kopu un validējot iegūtos rezultātus pret paraugmodeļiem, kas iegūti, klasterizējot apmācības datu kopu, var spriest par klasterizācijas rezultātu precizitāti.

Dots ieskats datu klasifikācijā [1, 3, 12, 58, 65, 76]. Aprakstīti *C4.5* [58], k-tuvāko kaimiņu [63], *CN2* [14], naivais Baijesa [63] datu klasifikācijas algoritmi un to precizitātes novērtēšanas kritēriji. Aprakstīts atribūtu informatīvuma noteikšanas process ar metodēm *CfsSubsetEval* un *BestFirst* [70].

3. APSTRĀDES SISTĒMAS PROGNOZĒŠANAS UZDEVUMIEM

Pēc definētās problēmas, ko apraksta teorētiskais modelis, kas parādīts 1. attēlā, sistēma risinās īsu laika rindu un to raksturojošo parametru apstrādi dažādās problēmvidēs ar datu iegūšanas metodēm un algoritmiem. Piedāvātā prognozēšanas sistēma dažādām problēmvidēm parādīta 3. attēlā. Analizējamā datu kopa, kas sastāv no īsām laka rindām un to raksturojošajiem parametriem, tiek pakļauta datu pirmapstrādes procesam. Īsās laika rindās tiek noteiktas objektu grupas – klasteri, lietojot darba autora piedāvātu klasterizācijas algoritma modifikāciju. No iegūtajiem klasteriem tiek veidoti paraugmodeļi, kas raksturo klastera objektu vidējās vērtības katrā laika periodā. Iegūtos klasterizācijas rezultātus (objektu klasteru numurus) apvieno ar pirmapstrādātajiem raksturojošajiem parametriem datu apvienošanas blokos. Ja nepieciešams (pēc uzdevuma nosacījumiem), apstrādāt klasterizācijas rezultātus ar dažādu klašu skaitu, lieto darba autora piedāvātu klašu transformāciju. Ja nepieciešams apvienoto datu kopu sadalīt apakškopās, lieto datu kopas sadalīšanas bloku, kur sadalīšana tiek veikta pēc sadalošā atribūta (nosaka eksperts) vērtību skaita: cik šim atribūtam ir iespējamo vērtību, tik apakškopas *H* tiks izveidotas. Izveidotajās datu kopās ar klasifikācijas palīdzību, lietojot klasifikācijas algoritmu, nosaka sakarības starp raksturojošajiem parametriem un klasi. Nosacījumu likumi kalpo divu klasifikatoru rezultātu salīdzināšanai vai zināšanu bāzes izveidošanai, kurā tiek glabāti likumi par katra objekta klasi, datu kopas sadalošā atribūta vērtību un klīnisko pētījumu rezultātu (izteikts kā skaitliska vērtība vai klase).

Prognozēšanas specifika ir atkarīga no izvirzītā uzdevuma. Prognozēšanā ar paraugmodeli ir jānosaka tā numurs (klase). To nosaka ar apmācīto klasifikatoru, klasificējot analizējamā

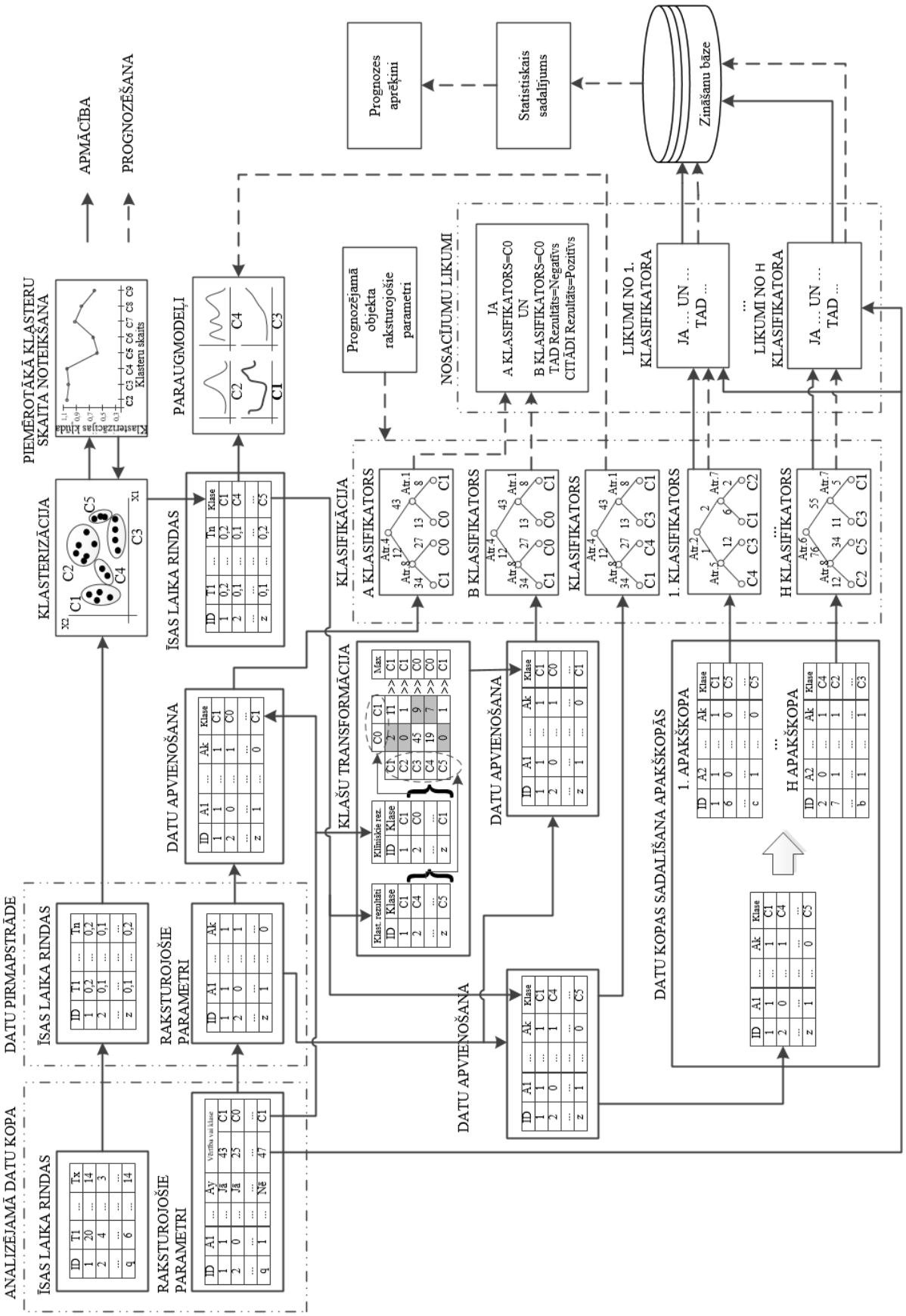
objekta raksturojošos parametrus. Noteiktais numurs norāda uz paraugmodeli, kas raksturo analizējamā objekta pieprasījumu. Cita prognozēšanas pieeja klasificē analizējamā objekta raksturojošos parametrus ar divu klasifikatoru palīdzību, un iegūtos rezultātus salīdzina, lietojot nosacījumu likumus. Nākamā pieeja piedāvā izveidot pagaidu nosacījuma likumu, kas iegūts, klasificējot analizējamā objekta raksturojošos parametrus, nosakot klasi, kā otrs parametrs nosacījumam kalpo sadalošā atribūta vērtība. Pēc tam no zināšanu bāzes atlasa tos likumus, kas atbilst izveidotajam pagaidu nosacītajam likumam. No atlasītajiem likumiem izveidojas statistiskais sadalījums, no kura ar darba autora piedāvātu pieeju aprēķina prognozes vērtību analizējamajam objektam.
Piedāvātā prognozēšanas sistēma dažādām problēmvidēm kalpo kā teorētiskais modelis, uz kura pamata tika izstrādātas preču pieprasījuma, sirds nekrozes riska un baktēriju proliferācijas sindroma prognozēšanas sistēmas. Šo sistēmu uzbūve sīkāk ir aprakstīta nākamajos apakšpunktos.

3.1. Pieprasījuma prognozēšanas sistēma

Pieprasījuma prognozēšanas sistēma (PPS) tika izstrādāta apģērbu uzņēmuma pasūtījumu optimizācijai, pamatojas uz iepriekšējo gadu preču pieprasījuma pieredzi (īsas laika rindas) un jaunā pasūtamā produkta (preces) raksturojošajiem datiem ar šādiem parametriem: preces cena, tips, sezonalitāte, kolekcijas ilgums (mēnešos), krāsa, izmērs u. c.

PPS uzbūve

PPS uzbūve parādīta 4. attēlā. Sākot darbu (apmācības process), sistēmai ir jāsaņem vēsturiskais pieprasījumu periods (gads={2006}), kuru ievada lietotājs – eksperts, balstoties uz savu pieredzi. Pēc perioda saņemšanas sistēma izveido vaicājumu, kuru nodod datubāzei ar valodas *PL/SQL* palīdzību. Datubāze glabā informāciju par preču pieprasījumu dažādos laika periodos, preču veidiem, izmēriem, krāsām, cenām, pavadzīmēm, piegādātājiem, klientiem, svītrkodiem u. c. Vaicājuma izpildes rezultātā tiek atlasīta datu kopa, kas tālāk nodota datu sagatavošanas blokam, kurā informācija sadalās divās datu plūsmās. Pirmo veido īsas laika rindas (preces identifikators un šīs preces pieprasījums gada griezumā pa mēnešiem), otro – raksturojošie parametri (preces identifikators, veids, tips un cena), kas tālāk nonāk prognozēšanas sistēmas modulī.
Prognozēšanas sistēmas modulī īsas laika rindas tiek pakļautas datu pirmapstrādei, kā rezultātā dati tiek attīrīti no trokšņainām vērtībām (objekti ar šāda veida vērtībām tiek izslēgti no datu kopas) un normalizēti (izlīdzināti vērtību diapazoni).
Ar klasterizācijas algoritmu nosaka sakarības starp īsām laika rindām, kas pēc līdzības pazīmēm apvieno grupās jeb klasteros. Piemērotāko klasteru skaitu nosaka klasterizācijas algoritms pēc minimālās klasterizācijas absolūtās kļūdas. Tad katram noteiktajam klasterim tiek izveidots paraugmodelis, kas raksturo klastera objektu vidējās vērtības katrā laika periodā.
Arī otra datu plūsma ir pakļauta datu pirmapstrādei, kuras rezultātā no raksturojošo parametru datu kopas tiek izslēgti objekti ar trūkstošām vērtībām un normalizētas atribūtu vērtības. Pēc tam noteikti informatīvākie atribūti (neinformatīvākie atribūti tiek izslēgti no datu kopas).

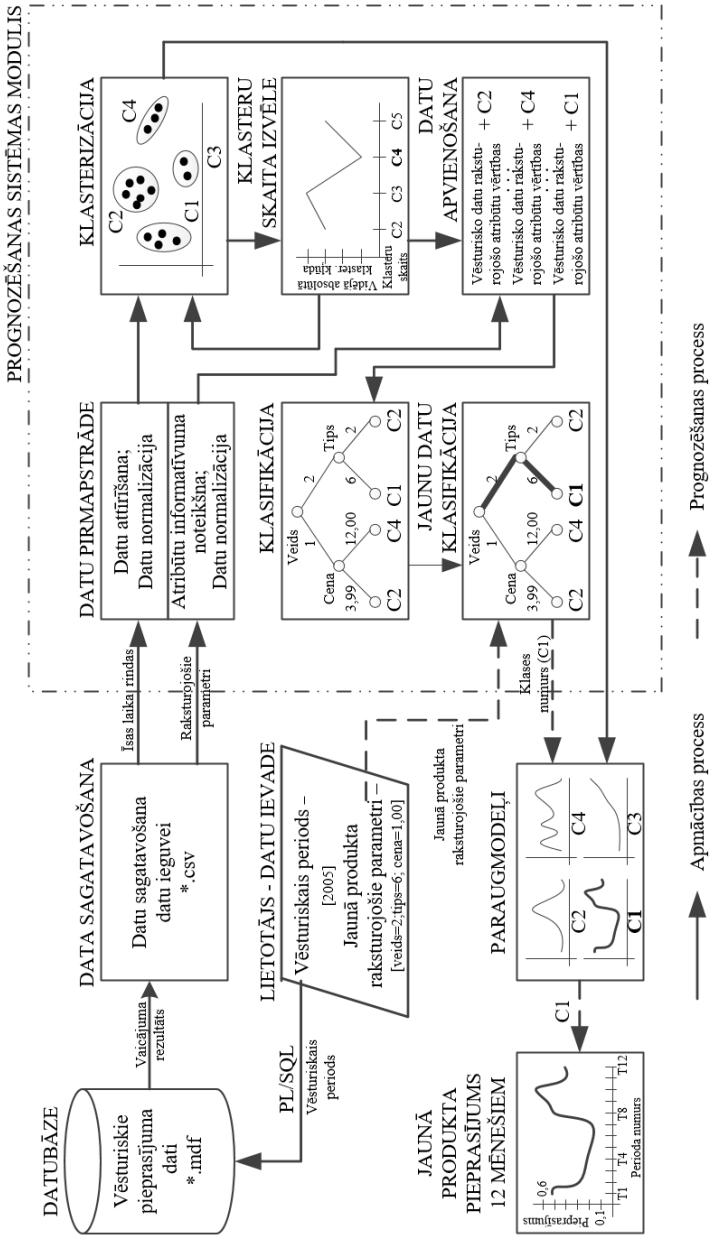


3. att. Prognozēšanas sistēma dažādām problēmvidēm

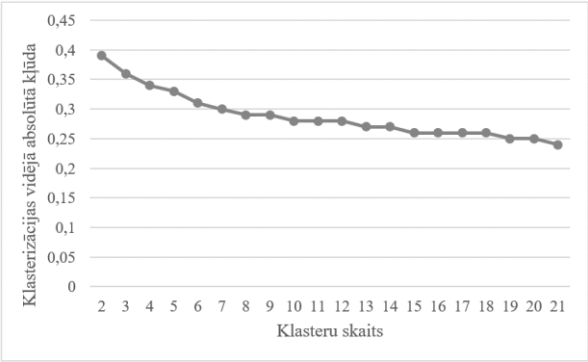
Tālāk vienā datu kopā apvieno pirmapstrādātos raksturojošos parametrus un klastera numuru, kas objektam noteikts klasterizācijas procesā. Apvienošana tiek veikta pēc objektu identifikācijas numuriem. Ar apvienoto datu kopu tiek veikta klasifikācija, lietojot induktīvos lēmuma kokus, rezultātā nosakot sakarības starp paraugmodeli (klastera numuru) un raksturojošajiem parametriem. Pēc apmācības procesa pabeigšanas ir iespējams veikt jaunā produkta pieprasījuma prognozi. Sistēmā tiek ievadīti jaunā produkta raksturojošie parametri (veids={2}, tips={6}, cena={1,00}). Uz klasifikācijas procesā «uzbūvētā» induktīvā lēmuma koka, projicējot tajā (prognozēšanas process) jaunā prognozējamā produkta raksturojošos parametrus, nosaka šā produkta klasi (piemēram, «C1»). Datu projicēšana notiek, virzoties no lēmuma koka saknes uz leju pa koka līmeņiem, līdz tiek sasniegta apakšējā līmeņa koka lapa ar klases numuru. Iegūtās klases numurs (piemēram, «C1») norāda uz attiecīgo paraugmodeli, kas noteikts klasterizācijas procesa rezultātā.
Paraugmodelis («C1 ») norāda kāds būs prognozētā produkta, kura raksturojošie parametri tika ievadīti sistēmā iepriekš, iespējamais pieprasījums 12 mēnešiem (periodiem no T1 līdz T12) nākotnē [40].

PPS eksperimentu rezultāti

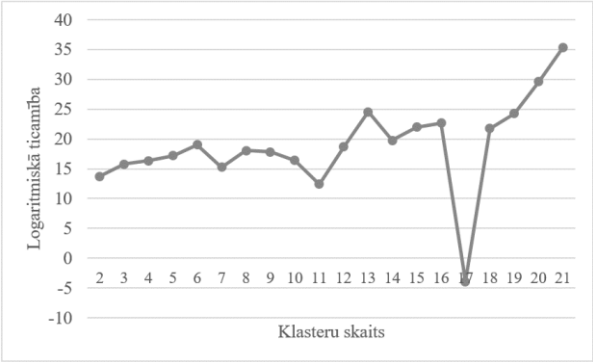
Eksperimentos tika izmantoti reāli apģērbu mazumtirdzniecības uzņēmuma vēsturiski pieprasījuma dati par 2005. gadu, kuros bija 423 objekti pēc datu attīrīšanas procesa, kas tika izmantoti modeļa apmācībai, un 149 objekti par 2006. gadu – modeļa testēšanai. Lai tos varētu izmantot salīdzināšanai ar izveidotajiem paraugmodeļiem, to dzīves ilgums reducēts līdz 12 periodiem (mēnešiem). Apmācības un testēšanas datu kopās bija arī produktu raksturojošie parametri (preces veids, tips un cena) katram objektam. Vēsturisku pieprasījumu apmācības datu kopas normalizācijai tika izmantota z-novērtējuma normalizācija ar standarta novirzi, kas parasti tiek lietota datu iegūšanā dominējošu atribūtu vērtību nolīdzināšanai, ja nav zināmas maksimālās vērtību robežas. Kā otra pieeja tika izmantota pārdošanas apjomu normalizācija ar dzīves likni, kas tika izmantota citu autoru darbos [64] līdzīgu datu struktūru normalizācijai. Eksperimentāli, veicot klasterizāciju ar k-vidējo sadalošo algoritmu pie 10 klasteriem (eksperimentāli noteikts kā piemērotākais klasteru skaits), dažādām normalizācijas pieejām lietojot 10-kārtu šķērsvalidāciju, tika noteikta apmācības kļūda: 3,28 – z-novērtējuma normalizācijai ar standarta novirzi un 0,28 – normalizācijai ar dzīves likni. No rezultātiem izriet, ka labākus rezultātus uzrādīja normalizācija ar dzīves līkni. Klasterizācijas process ar apmācības datu kopu tika veikts, lietojot k-vidējo sadalošo un maksimālās līdzības algoritmus. Iegūtie rezultāti ar k-vidējo sadalošo algoritmu parādīti 5. attēlā, ar maksimālās līdzības algoritmu – 6. attēlā.
Logaritmiskās ticamības samazinājums pie 17 klastera (skat. 6. attēlu), rodas tāpēc, ka pie šī klasteru skaita, salīdzinot ar citiem klasterim, ir neliels objektu skaits, kuriem Baijesa varbūtība bija 1. Iegūtie rezultāti parāda, ka noteiktais piemērotākais klasteru skaits abos gadījumos ir maksimālais – 21, jo vidējā absolūtā klasterizācijas kļūda ir jāņem pēc minimālās vērtības, bet logaritmiskā ticamība – pēc maksimālās vērtības. Abi klasterizācijas algoritmi nespēj veikt klasteru analīzi, apstrādājot vēsturiskas pieprasījuma vērtības – īsas laika rindas.



4. att. Pieprasījuma prognozēšanas sistēma

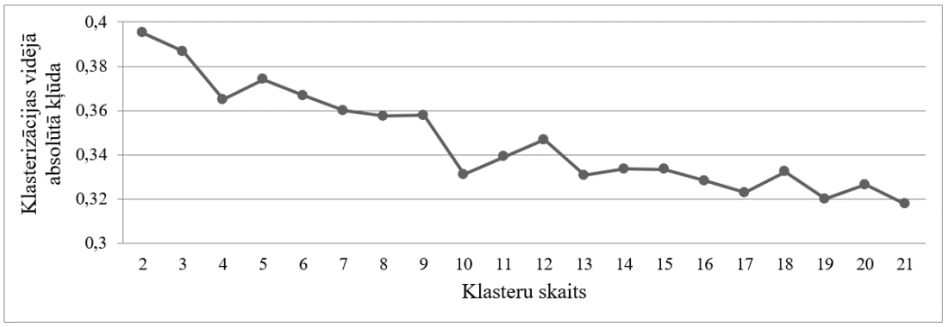


5. att. K-vidējo sadalošā algoritma klasterizācijas kļūdas novērtējums dažādam klasteru skaitam



6. att. Maksimālās līdzības algoritma logaritmiskās ticamības novērtējums dažādam klasteru skaitam

Gadījumā ar k-vidējo sadalošo algoritmu, pieaugot klasteru skaitam, kļūda samazinās, bet ar maksimālās līdzības algoritmu, pieaugot klasteru skaitam, logaritmiskā ticamība palielinās. Tas norāda, ka šie klasterizācijas algoritmi nav robusti, un nozīmē, ka tie nespēj veikt precīzu datu klasterizāciju. Tāpēc tika izmantots modificēts k-vidējo sadalošais algoritms, lai noteiktu piemērotāko klasteru skaitu, kas būtu nepieciešams apmācības datu kopas klasterizācijai.
Piemērotākais klasteru skaits ar modificētu k-vidējo sadalošo algoritmu tiek noteikts, vadoties pēc apakšpunktā «Klasterizācijas kļūdas aprēķināšana apmācības datu kopai» aprakstītās metodikas, nosakot klasterizācijas vidējo absolūto kļūdu, kuras rezultāti parādīti 7. attēlā. Pēc klasterizācijas vidējās absolūtās kļūdas rezultātu analīzes redzams, ka pie 10 klasteriem ir sasniegts pirmais vērā ņemamais minimums. Turpmāk poligona svārstības ir niecīgas, tāpēc par piemērotāko klasteru skaitu apmācības datu kopas klasterizācijai tiek pieņemts 10.



7. att. Klasterizācijas kļūdas aprēķins dažādam klasteru skaitam

Klasifikatoru precizitātes salīdzināšanai tika lietotas vidējās absolūtās kļūdas un vidējās kvadrātiskās kļūdas, jo dažādas kļūdu aprēķinu tehnikas nodrošina precīzāku iegūto rezultātu interpretāciju [64]. Precizitātes novērtēšanai klasifikatoru apmāca, lietojot apmācības datu kopu, bet pārbaudei – testēšanas datu kopu. Programmnodrošinājumā *OrangeCanvas* no vairākiem klasifikācijas algoritmiem tika izveidots modelis, ar kura palīdzību veikti eksperimenti piemērotākā klasifikatora noteikšanai. Iegūtie rezultāti parādīti 4. tabulā, kurā redzams, ka zemāko kļūdas vērtību ar abām kļūdas novērtējuma metodēm uzrādīja *C4.5* algoritms.
Pēc klasifikācijas precizitātes rezultātiem PPS tiek izmantots *C4.5* algoritms.

4. tabula

Dažādu klasifikatoru uzrādītās kļūdas ar testēšanas datu kopu

Kļūdas novērtējuma veids	Klasifikators					
	ZeroR	*OneR*	*JRip*	Naivais Baijesa	k-tuvāko kaimiņu	*C4.5*
Vidējā kvadrātiskā kļūda (*RMSE*)	0,295	0,375	0,284	0,282	0,298	**0,264**
Vidējā absolūtā kļūda (*MAE*)	17,4	14,1	16,1	15,6	16,0	**12,9**

3.2. Sirds nekrozes riska prognozēšanas sistēma

Latvijas Organiskās sintēzes institūts, kas nodarbojas ar medikamentu izstrādi sirds darbības funkciju uzlabošanai un veic laboratoriskus eksperimentus sirds nekrozes riska noteikšanai, izvirzīja uzdevumu izstrādāt sirds nekrozes riska prognozēšanas sistēmu, kas varētu noteikt iespējamo sirds nekrozes risku laboratorijas dzīvniekam, ievadot sistēmā tikai šā dzīvnieka raksturojošos parametrus. Šādas sistēmas izveide palīdzētu farmakoloģijas nozares speciālistiem ietaupīt laiku pētījumu rezultātu iegūšanai un eksperimentos iesaistīt mazāku laboratorijas dzīvnieku skaitu.

Saistībā ar sirds nekrozes riska prognozēšanas sistēmu (SNRPS) bioinformātikas nozares speciālistiem tiek piedāvāts risinājums apstrādāt sirds kontrakcijas spēka datus (īsas laika rindas) un laboratorijas dzīvnieku parametrus (raksturojošie parametri), nosakot «jaunā» indivīda sirds nekrozes risku.

Sistēmas izstrādāšanas procesā tiek analizēti farmakoloģisko pētījumu laboratorisko izmeklējumu dati, kas iegūti, izmantojot «izolētās sirds» išēmijas-reperfūzijas modeli [48, 49]. Eksperimenti veikti, izmantojot *Wistar* līnijas žurkas, kam noteiktu laika periodu (astoņu nedēļu garumā) ievadītas pētāmās sirds darbības stimulējošās vielas, kas pievienotas pārtikai. Katra dzīvnieku grupa tiek barota noteiktu laika periodu ar vienu no pētāmās vielas veidiem. Farmakoloģisko pētījumu mērķis ir noteikt pētāmo vielu efektivitāti sirds šūnu aizsardzībai pret išēmijas-reperfūzijas bojājumu, nosakot sirds nekrozes (atmirušo audu) daudzumu. Pētījumos izmantotais Mildronāts® ir Latvijā izstrādāts antiišēmisks līdzeklis, kas optimizē sirds enerģijas metabolismu [16]. Sirds kontrakcijas spēka un sirds ritma reģistrēšanai oklūzijas laikā tiek izmantota firmas *ADInstruments* aparatūra un programmnodrošinājums, nolasot datus ar intervālu 60 sekundes, iegūstot datu kopu, kas sastāv no 40 laika nolasījumiem oklūzijas periodā (katru minūti tiek reģistrēts sirds kontrakcijas spēks) ar sirds kontrakcijas spēka izmaiņām, un tas veido pirmo datu grupu. Sirds kontrakcijas spēka vērtības katrā laika periodā tiek izteiktas ar dzīvsudraba staba augstumu (mmHg). Iegūtās datu kopas vērtības ir laika rindas, bet tā kā to novērojumu ilgums ir pārāk īss un tas neatkārtojas, šīs vērtības ir uzskatāmas par īsām laika rindām. Analizējot īsas laika rindas, ir praktiski neiespējami noteikt tajās funkcionēšanas likumsakarības, tāpēc šāda veida uzdevumi ir uzskatāmi par grūti formalizējamiem. «Izolētās sirds» eksperiments tika veikts pēc tehnoloģijas, kas aprakstīta zinātniskajā rakstā [47]. Farmakoloģiskā eksperimenta mērķis ir noteikt atmirušo sirds audu (nekrozes) procentuālo attiecību, kas atkarīga no pētāmās vielas veidiem, ar kurām tika baroti dzīvnieki.

Otra datu grupa ir laboratorijas dzīvniekus raksturojošie parametri un farmakoloģisko eksperimentu laikā iegūtais sirds nekrozes riska novērtējums. Raksturojošie parametri apraksta laboratorijas dzīvnieka, piemēram, svaru, pētāmo vielu, asins plazmas analīzes parametrus un sirds nekrozes risku.

SNRPS uzbūve

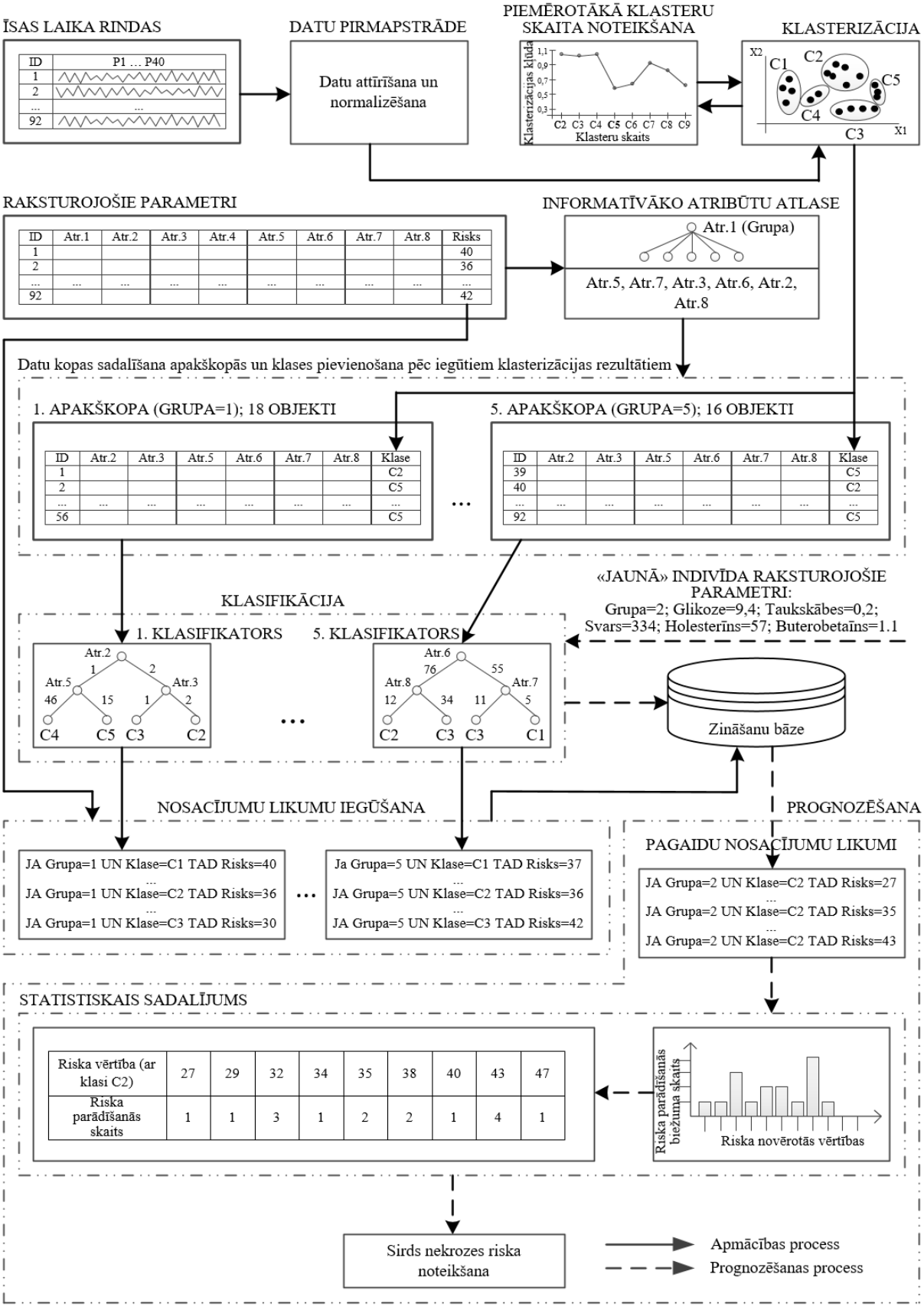
Sākotnēji tiek veikta īsu laika rindu pirmapstrāde, no kuras tiek izslēgts pirmais un pēdējais mērījums, lai izvairītos no iespējamā trokšņa (kļūdaini rādījumi). Tad tiek izveidota laika rinda ar 38 periodiem katram datu kopas objektam, atlasot datus par sirds kontrakcijas spēku. Datu normalizācija tiek veikta, izmantojot normalizāciju ar dzīves līkni [40, 64] un *z*-novērtējuma normalizāciju ar standarta novirzi [1] pieejas, kas tika izmantotas arī PPS izstrādāšanā. Sirds nekrozes riska prognozēšanas sistēmas darbības princips parādīts 8. attēlā. Tās uzbūve balstās uz vairāku datu iegūšanas metožu lietošanu. Matemātiskās statistiskas metožu pieejas ir ierobežotas, risinot uzdevumus, kur jāanalizē sakarības starp īsām laika rindām. Tāpēc šādu uzdevumu atrisināšanai tiek lietotas datu iegūšanas pieejas, kas tiek uzskatītas par piemērotākām [76].

Īsas laika rindas tiek klasterizētas ar modificētu k-vidējo algoritmu, nosakot piemērotāko klasteru skaitu, kas nepieciešams līdzīgu objektu apvienošanai klasteros. Indivīda raksturojošajos parametros tiek meklēta atribūtu savstarpējā korelācija, nosakot to mijiedarbības pakāpi un atlasot informatīvākos atribūtus tālākajai datu analīzei. Raksturojošo atribūtu datu kopa tiek sadalīta piecās vienādās apakškopās, tādēļ, ka barībai pievieno piecu veidu pētāmās vielas. Sadalītajām apakškopām pievieno klasterizācijas rezultātā iegūto klasi, vadoties pēc objekta identifikatora. Katrai no izveidotajām apakškopām ar klasifikācijas palīdzību tiek noteikta saikne starp klasterizācijā iegūto objekta klasi un šā objekta raksturojošajiem parametriem, izmantojot induktīvo lēmumu koku algoritmu *C4.5* [58]. Objekta klasifikācijas procesā, no uzbūvētā lēmuma koka, tiek iegūts nosacījuma likums «JA …UN … TAD …» veidā, kas glabā informāciju par klasterizācijā iegūto klasi, pētāmās vielas grupu un farmakoloģiskajos eksperimentos iegūto sirds nekrozes riska novērtējumu. Iegūtie nosacījumu likumi tiek saglabāti zināšanu bāzē. Sirds nekrozes riska prognozēšana tiek veikta, ievadot sistēmā «jaunā» indivīda raksturojošos parametrus. Vadoties pēc pētāmās vielas grupas, raksturojošie parametri tiek projicēti uz attiecīgo klasifikatoru, iegūstot klases vērtību un izveidojot nosacījumu «JA … UN …» (piemēram, JA Grupa=2 UN Klase=C2), kas tiek nodots zināšanu bāzei. Tālāk no zināšanu bāzes tiek atlasīti visi tie gadījumi, kas atbilst izveidotajam nosacījumam (JA Grupa=2 UN Klase=C2). Katrs nosacījums tiek papildināts ar atrasto informāciju zināšanu bāzē, veidojot pagaidu nosacījumu likumu kopu «JA … UN … TAD …», piemēram:

- JA Grupa=2 UN Klase=C2 TAD Risks=27;
- JA Grupa=2 UN Klase=C2 TAD Risks=32;
- …;
- JA Grupa=2 UN Klase=C2 TAD Risks= 40.

No atlasītās pagaidu nosacījumu likumu kopas izveidojas sirds nekrozes riska parādīšanās biežuma statistika (statistiskais sadalījums), kas veido riska sadalījuma funkciju, no kuras iegūstam, cik reizes attiecīgā riska vērtība parādās pagaidu likumu kopā.

Sirds nekrozes riska noteikšanai tiek lietots matemātiskās cerības aprēķins un darba autora piedāvāta pieeja, kas balstīta uz attāluma aprēķina.



8. att. Sirds nekrozes riska prognozēšanas sistēmas uzbūve

Matemātiskās cerības aprēķinam tiek izmantots sirds nekrozes riska statistiskais sadalījums, kur katrai biežuma statistikas vērtībai piešķir varbūtību pēc riska biežuma skaita, jo biežāk šī vērtība ir sastopama pagaidu nosacījumu likumu kopā, jo varbūtība būs lielāka. Tālāk, saskaitot katras riska aprēķinātās matemātiskās cerības vērtības, tiek iegūta sirds nekrozes riska prognoze. Darba autora izstrādātā pieeja lieto attālumu aprēķinu starp riska parādīšanās biežuma skaitu un riska vērtībām.

SNRPS eksperimentu rezultāti

Eksperimentiem izmantota 92 objektu datu kopa, kas raksturo sirds kontrakcijas spēka vērtības išēmijas stadijā (īsas laika rindas), un šo objektu raksturojošie parametri (atribūtiem), tādi kā: grupa(Atr.1.) – pētāmās vielas tips, ar kuru tika baroti indivīdi; svars(Atr.2.); asins plazmas parametri: karnitīns(Atr.3.), triglicerīdi(Atr.4.), taukskābes(Atr.5.), glikoze(Atr.6.), holesterīns(Atr.7.), butirobetaīns(Atr.8.) un risks(Klase) – sirds nekrozes riska novērtējums, kas iegūts farmakoloģisko pētījumu rezultātā. Īsu laika rindu garums pēc datu attīrīšanas bija 38 periodi. Iegūtās īsās laika rindas tika normalizētas, lietojot z-novērtējuma normalizāciju ar standarta novirzi un normalizāciju ar dzīves līkni pieejas, iegūstot katrai no tām divas dažādas datu kopas, ar kurām eksperimentējot, tika noteikts, ka piemērotākā bija normalizācija ar dzīves līkni. Izveidotās datu kopas tika klasterizētas ar k-vidējo sadalošo, maksimālās līdzības, aglomeratīvo hierarhisko un modificētu k-vidējo sadalošo algoritmiem. Ar k-vidējo sadalošo algoritmu, kura rezultāti parādīti 5. tabulā, pēc summētās kvadrātiskās kļūdas abām normalizācijas pieejām mazākā kļūda sasniegta pie 9 klasteriem, kas norāda uz klasterizācijas algoritma nepiemērotību īsu laika rindu klasterizācijai, jo, palielinoties klasteru skaitam, samazinās kļūda.

5. tabula

K-vidējo sadalošā algoritma summētā kvadrātiskā kļūda

	Klasteru skaits							
Normalizācijas pieeja	2	3	4	5	6	7	8	9
Z-novērtējuma normalizācija ar standarta novirzi	141,67	138,42	128,36	125,89	122,64	120,56	116,34	**113,92**
Normalizācija ar dzīves līkni	115,73	112,99	101,81	96,36	92,54	89,44	87,39	**85,79**

Ar maksimālās līdzības algoritmu, kura rezultāti parādīti 6. tabulā, pēc logaritmiskās ticamības pirmajai datu kopai, lietojot z-novērtējuma normalizāciju ar standarta novirzi, tika noteikts piemērotākais klasteru skaitu 9, savukārt otrajai datu kopai, lietojot normalizāciju ar dzīves līkni – 8. Kā parāda rezultāti, arī šis klasterizācijas algoritms nespēj noteikt piemērotāko klasteru skaitu, jo, pieaugot to skaitam, palielinās logaritmiskā ticamība, kas rada šaubas par algoritma robustumu.

Ar modificētu k-vidējo sadalošo algoritmu, kura rezultāti parādīti 7. tabulā, pēc vidējās absolūtās kļūdas, lietojot normalizāciju ar dzīves līkni, kā piemērotākais klasteru skaits tika noteikts 5. Raksturojošajos parametros tika meklēta informatīvāko atribūtu kopa, kas būtu lietojama klasifikatora uzbūvē. Datu klasifikācija ir atkarīga no atribūtu informatīvuma. Ja starp atribūtiem pastāv korelācija, uzskata, ka atribūti savstarpēji ir saistīti [70]. Savukārt, ja atribūti nekorelē, tas nozīmē, ka starp šiem atribūtiem ir vāja saite un tiem nav nozīmes lietot klasifikāciju. Pētījumos tika izmantota *CfsSubsetEval* atribūtu novērtējuma un *BestFirst* pārmeklēšanas metodes, kas tiek bieži lietotas datu iegūšanā atribūtu informatīvuma noteikšanai [70].

6. tabula

Maksimālās līdzības algoritma logaritmiskā ticamība

	Klasteru skaits							
Normalizācijas pieeja	2	3	4	5	6	7	8	9
Z-novērtējuma normalizācija ar standarta novirzi	–49,64	–48,64	–47,93	–47,39	–46,36	–45,47	–45,25	**–44,13**
Normalizācija ar dzīves līkni	144,49	145,40	148,42	149,06	150,06	150,32	**151,13**	150,87

7. tabula

Modificēta k-vidējo sadalošā algoritma vidējā absolūtā kļūda

	Klasteru skaits							
Normalizācijas pieeja	2	3	4	5	6	7	8	9
Z-novērtējuma normalizācija ar standarta novirzi	1,233	1,192	1,158	1,150	1,128	1,095	1,093	0,946
Normalizācija ar dzīves līkni	1,026	1,017	1,026	**0,59**	0,644	0,91	0,824	0,615

Tika iegūti šādi rezultāti: taukskābes, holesterīns, karnitīns uzrādīja 100 %, glikoze 90 %, svars 80 % un butirobetaīns 20 % korelāciju, savukārt triglicerīdi un grupa 0 %. Pēdējos divus parametrus pēc loģikas vajadzēja izslēgt, bet parametram grupa ir ļoti svarīga informācija par pētāmās vielas veidiem, kas ir šā pētījuma pamatā, tāpēc šis parametrs tika saglabāts. Tad tika pieņemts lēmums sadalīt pilno datu kopu apakškopās, balstoties pēc parametra grupa vērtībām, iegūstot piecas apakškopas. Šāda datu kopas sadalīšana ļauj atbrīvoties no neinformatīva atribūta izmantošanas klasifikatora apmācības procesā un tajā pat laika patur svarīgu informāciju par pētāmo vielu, kas tālāk tiks izmantota nosacījumu likumu veidošanai.

Piemērotākais klasifikators tika izvēlēts pēc klasifikācijas precizitātes rezultātiem, kas parādīti 8. tabulā, lietojot modificētā k-vidējo sadalošā algoritma noteikto piemērotāko klašu skaitu, un 9. tabulā, lietojot k-vidējo sadalošā algoritma noteikto piemērotāko klašu skaitu.

Eksperimentāli starp Naivo Baijesa, k-tuvāko kaimiņu, *C4.5* un *CN2* algoritmiem noteikts, ka piemērotākais ir *C4.5* algoritms. No iegūtajiem klasifikācijas rezultātiem tika veidoti nosacījumu likumi, kas glabājas zināšanu bāzē. Sirds nekrozes riska prognoze tiek veikta, pamatojoties uz klasifikācijas procesā iegūtā lēmuma koka un «jaunā» indivīda raksturojošajiem parametriem.

8. tabula

Ar klasifikācijas algoritmiem, lietojot modificētā k-vidējo sadalošā algoritma atrasto piemērotāko klašu skaitu, noteiktā klasifikatoru precizitāte

	Klasifikators			
	Naivais Baijesa	*C4.5*	k-tuvāko kaimiņu	*CN2*
1. apakškopa	0,45	0,7	0,65	0,6
2. apakškopa	0,25	0,3	0,2	0,15
3. apakškopa	0,25	0,6	0,4	0,5
4. apakškopa	0,45	0,55	0,3	0,75
5. apakškopa	0,2	0,25	0,5	0,3
Vidējā vērtība	0,32	**0,48**	0,41	0,46
Nesadalīta datu kopa		0,38		

9. tabula

Ar klasifikācijas algoritmiem, lietojot k-vidējo sadalošo algoritmu ar iepriekš atrasto piemērotāko klašu skaitu, noteiktā klasifikatoru precizitāte

	Klasifikators			
	Naivais Baijesa	*C4.5*	k-tuvāko kaimiņu	*CN2*
1. apakškopa	0,35	0,45	0,4	0,1
2. apakškopa	0	0,1	0,05	0,05
3. apakškopa	0,45	0,35	0,4	0,2
4. apakškopa	0,05	0,05	0,2	0,05
5. apakškopa	0,27	0,4	0,25	0,25
Vidējā vērtība	0,224	0,27	0,26	0,13
Nesadalīta datu kopa		0,3		

Šos raksturojošos parametrus klasificē uz tā klasifikatora, kura numurs atbilsts atribūta grupa vērtībai, iegūstot klasi. No iegūtas klases un atribūta grupa vērtības tiek izveidots pagaidu nosacījuma likums, ar kuru no zināšanu bāzes atlasa riska vērtības, izveidojot nosacījumu likumus. Izveidotie nosacījumu likumi veido statistisko sadalījumu, no kura ar attālumu mēru palīdzību, kas parādīti 10. tabulā, tiek aprēķināts sirds nekrozes risks.

3.3. Baktēriju proliferācijas sindroma noteikšanas sistēma

Baktēriju proliferācija tievajā zarnā ir šīs zarnas kolonizācija ar resnās zarnas mikroorganismiem, kas var radīt plašu klīnisko spektru, sākot no viegliem un nekonkrētiem simptomiem līdz smagiem gremošanas traucējumiem. Par baktēriju proliferāciju tievajā zarnā tās sākumstadijā, kad baktērijas pārvietojas no resnās zarnas uz tievās zarnas sākumdaļu,

sūdzības parasti ir tikai indivīdiem ar hroniskām gremošanas trakta slimībām. Simptomi, ko rada baktēriju proliferācija tievajā zarnā, ir ļoti netipiski. Šie slimnieki sūdzas par zarnu diskomfortu, pastiprinātu gāzu veidošanos (meteorisms), nekārtīgu vēdera izeju [50, 56].

10. tabula

Iespējamā sirds nekrozes riska aprēķins pēc riska parādīšanās biežuma attāluma novērtējuma

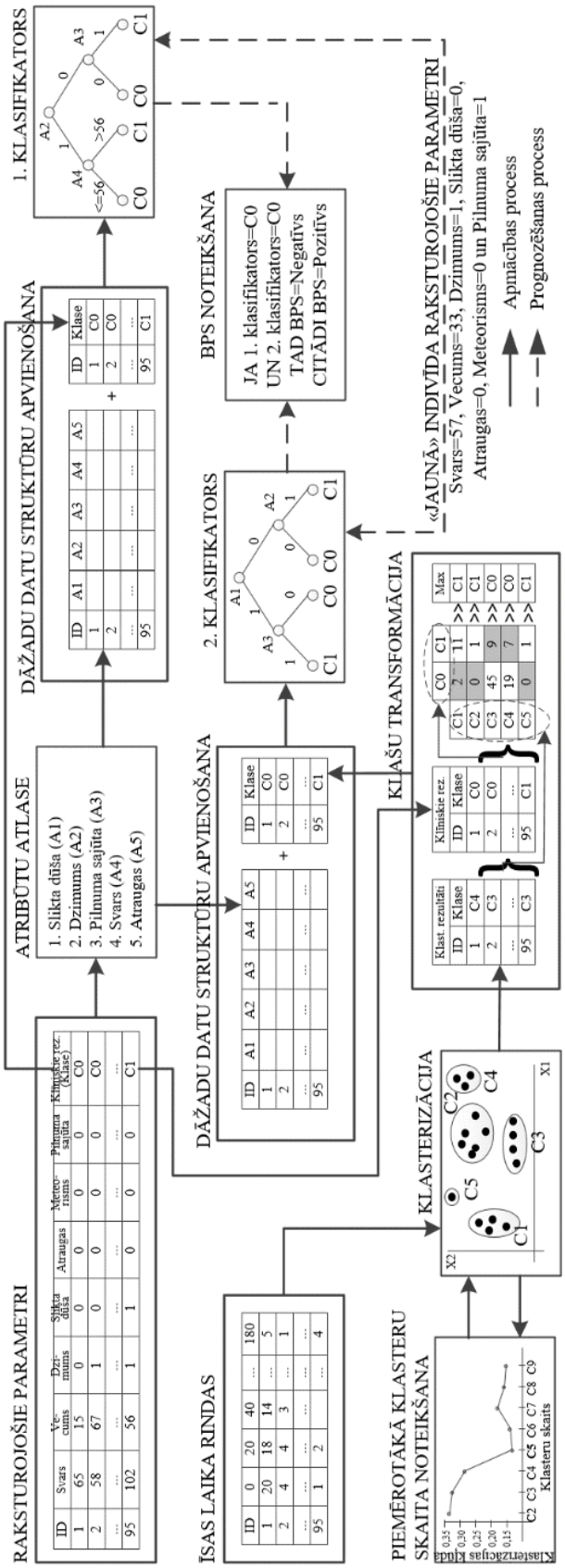
	Aprēķini								
Riska parādīšanās biežuma skaits (RPBS)	1	1	3	1	2	2	1	4	1
Riska vērtības	27	29	32	34	35	38	40	43	47
Attālums starp riska vērtībām	–	2	3	2	1	3	2	3	4
Starpība starp RPBS min un RPBS vērtībām	0	0	–2	0	–1	–1	0	–3	0
Summa: Attālums + starpība		2	1	2	**0**	2	2	**0**	4
Novērtējums					35			43	
Novērtējuma vidējā vērtība	**39**								

Baktēriju proliferācijas sindroma (BPS) noteikšanai diagnostikas testi tiek iedalīti invazīvajos un neinvazīvajos testos. BPS noteikšana vēl nav standartizēta [20], tāpēc šobrīd klīniskajā praksē tiek izmantoti glikozes (GET) un laktulozes (LET) elpas testi. Klīniski BPS diagnostikas algoritms nosaka, ka indivīdam sākotnēji tiek veikts GET tests. Ja tas izrādījies pozitīvs, tiek veikts LET tests. Tikai abu pozitīvu rezultātu gadījumā indivīdam diagnosticē BPS. Elpas testos izmanto noteiktu ūdeņraža koncentrāciju izelpas gaisā. Indivīdiem, kas cieš no BPS, glikozes elpas tests (GET) parasti uzrāda vienu «agro» maksimālo ūdeņraža izdalīšanos. Vismaz sešas stundas pirms testa veikšanas indivīdam ir jāatturas no ēdienreizēm. Vismaz 30 minūtes pirms substrāta uzņemšanas indivīdam ir jāatturas no smēķēšanas un fiziskas slodzes. Alveolārais gaiss, t. i., pēdējā daļa no vienas izelpas (aptuveni 150 ml), tiek lietota mērījumam. GET procedūra: 75 g glikozes, kas ir izšķīdināta 400 ml ūdens, ievada orāli [46]. Ja H_2 koncentrācija izelpā palielinās vairāk kā $\Delta = 20$ Pm, pieņem, ka tests ir pozitīvs un ir apstiprināts BPS tievajā zarnā.

GET un LET testi ir laikietilpīgi, un tie prasa ilgu pirmssagatavošanās procesu, kas parasti indivīdiem izraisa negatīvu attieksmi pret šo procedūru. Tāpēc mērķis ir piedāvāt alternatīvas pieejas BPS noteikšanai tievajā zarnā, izmantojot datu iegūšanas metodes un algoritmus. No izmeklējumu rezultātiem, veicot GET testus un atbildot uz anketēšanas jautājumiem par indivīda pašsajūtas novērtējumu, tiek piedāvāta sistēma, kas nosaka BPS tievajā zarnā. GET testu rezultāti ir uzskatāmi par īsām laika rindām [21], jo nolasījumu skaits vienai laika rindai ir 10 periodi (laika intervāli), kas iegūti no izelpas paraugiem (ievākti nekavējoties pirms procedūras, kā arī katras 20 minūtes pēc substrāta ieņemšanas trīs stundu laikā). Indivīda pašsajūtas novērtējums tiek aprakstīts ar virkni jautājumu, uz kuriem atbildes sniedz indivīds GET testa izpildes laikā. Jautājumus uzdod kvalificēts medicīnas darbinieks, atbildes reģistrējot speciālā protokolā.

BPS noteikšanas sistēmas darbības princips

BPS noteikšanas sistēmai, kas parādīta 9. attēlā, ir jāapstrādā divu veidu datu plūsmas. Viena



9. att. Baktēriju proliferācija sindroma noteikšanas sistēma

ir GET testu vēsturiskie rezultāti, kas attēloti kā īsas laika rindas ar 10 laika intervāliem, otra – indivīda pašsajūtas novērtējums, ko raksturo atribūti – dzimums; slikta dūša; atraugas; meteorisms; pilnuma sajūta – un klīnisko izmeklējumu rezultāti, ko raksturo atribūts klase. Sistēmai jāveic informatīvāko atribūtu atlasi, datu klasterizāciju, piemērotākā klasteru skaita noteikšanu, klašu transformāciju, jaunas datu kopas izveidi, apvienojot dažādas datu struktūras, divu veidu datu plūsmu klasifikāciju un BPS noteikšanu. Atribūtu informatīvuma noteikšanai tiek lietotas *CfsSubsetEval* atribūtu novērtējuma un *BestFirst* pārmeklēšanas metodes, kas bieži tiek izmantotas datu iegūšanā [70]. Pēc indivīda pašsajūtas novērtējuma un raksturojošo informatīvāko atribūtu atlases datu kopai, kas darbā tiks apzīmēta ar *DS1*, pievieno klīnisko izmeklējumu rezultātus (atribūts klase – klīniski apstiprināts baktēriju proliferācijas sindroms). Izveidotā datu kopa *DS1* tiek nodota 1. klasifikatoram, lietojot k-tuvāko kaimiņu, *C4.5* un *CN2* klasifikācijas algoritmus, nosaka sakarības starp indivīda raksturojošajiem parametriem un klasēm. Īsas laika rindas tiek apstrādātas klasterizācijas modelī, lietojot modificētu k-vidējo sadalošo algoritmu, tiek noteikts piemērotākais klasteru skaits datu kopas klasterizācijai. Klasterizācijas rezultātā iegūtās klases transformācijas modelī tiek pārveidotas uz klīnisko izmeklējumu klašu struktūru. Iegūtā klašu struktūra, vadoties pēc indivīda identifikācijas numura, tiek pievienota atlasītajai informatīvāko atribūtu kopai, kas darbā tiks apzīmēta ar *DS2*, kas tālāk tiek apstrādāta 2. klasifikatorā, lietojot k-tuvāko kaimiņu, *C4.5* un *CN2* klasifikācijas algoritmus, nosakot sakarības starp indivīda

raksturojošajiem parametriem un klasterizācijas rezultātiem. No 1. klasifikatora, kas apstrādā *DS1* datu kopu, tiek iegūti rezultāti, kas raksturo saikni starp indivīda pašsajūtas novērtējuma raksturojošiem parametriem un klīniskajiem rezultātiem. No 2. klasifikatora, kas apstrādā *DS2* datu kopu, tiek iegūti rezultāti, kas raksturo saikni starp elpas testu mērījumiem un klasterizācijā iegūto klasi, kas tiek reducēta uz klīnisko rezultātu klašu struktūru.
Divu dažādu datu plūsmu apstrādes modelis nodrošina to, ka iegūtie rezultāti tiek analizēti no dažādām pusēm. Tiek iegūti divi neatkarīgi novērtējumi, kuru rezultātus salīdzina BPS noteikšanas modelī, kas izvada paziņojumu, vai indivīdam ir nepieciešama tālāka izmeklēšana. Šis piedāvātais divu plūsmu apstrādes modelis garantē, ja indivīds negodprātīgi atbildējis uz pašsajūtas novērtējuma jautājumiem, pastāv iespēja, ka otrs klasifikators atklās nepatiesi sniegtās ziņas.

BPS noteikšanas sistēmas eksperimentu rezultāti

Eksperimentiem izmantoti dati no retrospektīva pētījuma, kurā iekļauti abu dzimumu indivīdi bez vecuma ierobežojuma. Pētījumā piedalījās 95 indivīdi, kuriem tika veikti glikozes elpas testi. Tika izstrādāts protokols, kurā ietverts: glikozes elpas testa ūdeņraža nolasījums bez CO_2 korekcijas laika periodā no 0 līdz 180 minūtēm ar intervālu 20 minūtes, kā arī indivīda svars un vecums iekļaušanas brīdī pētījumā; dzimums; glikozes testa slēdziens jeb klīniskā pētījuma rezultāts – klase (negatīvs-0; pozitīvs-1); indivīda pašsajūtas novērtējuma parametri: slikta dūša; atraugas; meteorisms un pilnuma sajūta. Negatīvs glikozes testa rezultāts norāda, ka tālāka pārbaude nav nepieciešama, pozitīvs rezultāts – indivīdam ir nepieciešama papildu izmeklēšana.
Tika noteikti informatīvākie atribūti, lietojot *CfsSubsetEval* atribūtu novērtējuma un *BestFirst* pārmeklēšanas metodes. No sākotnējās datu kopas, kurā bija septiņi atribūti, tika atlasīti pieci atribūti: slikta dūša (100 %), dzimums (70 %), pilnuma sajūta (50 %), svars (40 %) un atraugas (40 %). Iegūtie rezultāti iekavās norāda, cik procentos gadījumu atribūts ir ticis iekļauts atribūtu kombinācijās ar lielāko novērtējumu.
Eksperimentu gaitā tika noteikts piemērotākais klasteru skaits, kas jāizvēlas datu kopas klasterizācijai, vadoties pēc vidējās klasterizācijas kļūdas pie dažādu klasteru skaita. Kā redzams 11. tabulā, piemērotākais klasteru skaits ir pieci, jo pie šāda klasteru skaita tiek sasniegta mazākā vidējā absolūtā kļūda (*MAE*).

11. tabula

Klasterizācijas rezultāti ar modificētu k-vidējo sadalošo algoritmu

	Klasteru skaits							
	2	3	4	5	6	7	8	9
Vidējā absolūtā kļūda	0,335	0,329	0,283	**0,125**	0,146	0,183	0,167	0,154

Rezultāti norāda arī uz klasterizācijas robustumu, lietojot šo algoritmu īsu laika rindu apstrādei, kas nav iegūstams, lietojot, piemēram, klasisko k-vidējo sadalošo algoritmu, kura iegūtie rezultāti parādīti 12. tabulā.

12. tabula

Klasterizācijas rezultāti ar k-vidējo sadalošo algoritmu

	Klasteru skaits							
	2	3	4	5	6	7	8	9
Summēta kvadrātiskā kļūda	20,56	19,15	17,42	15,93	15,23	14,74	13,76	**12,59**

Lai noteiktu piemērotāko klasifikācijas algoritmu, ko varētu izmantot par pamatu realizējamās sistēmas izveidē, tika veikta virkne eksperimentu ar *C4.5*, *CN2* un k-tuvāko kaimiņu klasifikācijas algoritmiem. Iegūtie klasifikācijas eksperimentu rezultāti parādīti 13. tabulā, lietojot 1. klasifikatoru ar *DS1* datu kopu un dažādas precizitātes novērtēšanas pieejas.

13. tabula

Klasifikācijas rezultāti ar *DS1* datu kopu

Klasifikācijas algoritmi	10-kārtas šķērsvalidācija			Izlaist vienu		
	Klasifikācijas precizitāte	Jutīgums	Specifiskums	Klasifikācijas precizitāte	Jutīgums	Specifiskums
C4.5	0,65	0,87	0,08	0,64	0,87	0,04
kNN	0,63	0,86	0,04	0,68	0,87	0,12
CN2	0,67	0,91	0,04	0,63	0,86	0,04

Novērojams, ka izmantotie klasifikatori labāk ir atpazinuši negatīvo klasi («C0») – indivīdus, kuriem nav nepieciešam tālāka izmeklēšana. Ja raugās no izvirzītā darba uzdevuma, sasniegts pozitīvs rezultāts, bet, ja skatās – cik labi klasifikators ir atpazinis pozitīvo klasi («C1»), rezultāts nav iepriecinošs, jo specifiskums, piemēram, ar *C4.5* algoritmu pie 10-kārtas šķērsvalidācijas ir tikai 0,08.

Nākamajā eksperimentā, lietojot 2. klasifikatoru ar *DS2* datu kopu, tika iegūti rezultāti, kas parādīti 14. tabulā. Rezultāti parāda, ka *kNN* algoritms vienlīdz labi ir atpazinis gan pozitīvo, gan negatīvo klasi. *CN2* algoritms uzrāda gandrīz 100 % negatīvās klases atpazīstamību, taču pozitīvās klases atpazīstamību varētu vēlēties augstāku. Savukārt *C4.5* algoritms ar abām precizitātes novērtējuma pieejām uzrādīja līdzīgus rezultātus, atpazīstot negatīvo klasi, jutīgums bija attiecīgi 0,65 un 0,69, savukārt pozitīvo klasi – specifiskums bija 0,39 abos gadījumos.

14. tabula

Klasifikācijas rezultāti ar *DS2* datu kopu

Klasifikācijas algoritmi	10-kārtas šķērsvalidācija			Izlaist vienu		
	Klasifikācijas precizitāte	Jutīgums	Specifiskums	Klasifikācijas precizitāte	Jutīgums	Specifiskums
C4.5	0,53	0,65	0,39	0,55	0,69	0,39
kNN	0,46	0,51	0,39	0,42	0,43	0,41
CN2	0,6	0,98	0,16	0,63	0,98	0,23

C4.5 algoritms uzrādīja salīdzinoši vienādus rezultātus ar abām precizitātes novērtējuma pieejām, tāpēc, klasificējot *DS1* un *DS2* datu kopas attiecīgi ar 1. klasifikatoru un

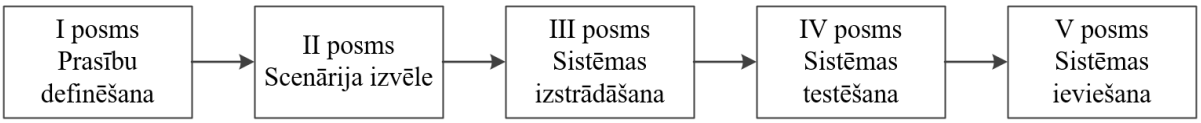
2. klasifikatoru, par pamatu datu kopu klasifikācijai BPS noteikšanas sistēmā tika izvēlēts *C4.5* algoritms.

4. PROGNOZĒŠANAS SISTĒMAS IZSTRĀDĀŠANAS VADLĪNIJAS

Prognozēšanas sistēmas izstrādāšanas vadlīnijas (PSIV) sniedz skaidrojumus, kā realizēt datu apstrādes sistēmas, kur par datu avotu kalpo īsas laika rindas un to raksturojošie parametri, balstoties uz iegūto pieredzi līdzīgu sistēmu izveidošanā. Vadlīnijas veidotas, balstoties uz iegūto pieredzi preču pieprasījumu, sirds nekrozes riska un baktēriju proliferācijas sindroma prognozēšanas sistēmu izstrādāšanā. Vadlīnijas norāda izpildītājam, kādu modeļa struktūru ieteicams veidot, balstoties uz pasūtītāja iesniegto datu aprakstu. Pasūtītājs ir persona vai organizācija, kas definē prasības un iesniedz analizējamo datu aprakstu. Izpildītājs ir persona vai organizācija, kas novērtē iesniegtās prasības, izanalizē tās un sniedz pasūtītājam atbildi. Ja puses vienojas par nākamo soli, izpildītājs uz iesniegto prasību pamata izstrādā piemērotāko sistēmas risinājumu, balstoties uz vadlīnijām. Pēc prognozēšanas sistēmas risinājuma izvēles izstrādātājs realizē sistēmu un veic tās darbības pārbaudi, vadoties pēc klasifikācijas precizitātes novērtējuma. Ja novērtējums ir pietiekams, izpildītājs veic sistēmas ieviešanu un integrāciju pasūtītāja informācijas sistēmās.

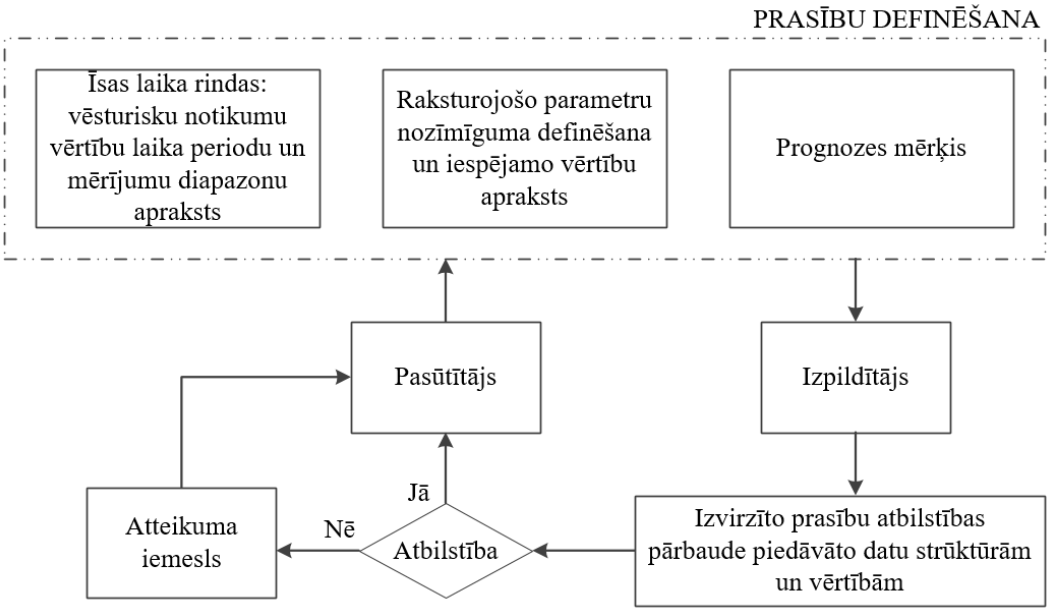
Prognozēšanas sistēmas struktūra

Prognozēšanas sistēmas struktūra ir parādīta 10. attēlā, tai ir pieci posmi. Pirmajā posmā tiek definētas prasības, kas ietver pasūtītāja un izpildītāja dialogu, analizējot pasūtītāja sniegto datu atbilstību izstrādājamajai sistēmai un izvirzītajam uzdevumam – prognozes mērķim. Otrajā posmā notiek piemērotākā sistēmas izstrādāšanas scenārija izvēle. Trešajā posmā tiek izstrādāts sistēmas koncepts. Ceturtajā posmā notiek izstrādātā sistēmas koncepta testēšana, ja izveidotā koncepta precizitāte ir pietiekama, tas tiek pārveidots sistēmā. Piektajā posmā tiek veikta izveidotās sistēmas integrācija pasūtītāja informācijas sistēmā.



10. att. Prognozēšanas sistēmas izstrādāšanas posmi

Prasību definēšanas posms izvērstā veidā parādīts 11. attēlā, kur pasūtītājs iesniedz izpildītājam analizējamo datu kopumu un definē prognozes mērķi. Lai izveidotu prognozēšanas sistēmu, analizējamiem datiem ir jāsatur vēsturisku notikumu vērtības un to raksturojošie parametri. Prognozes mērķim ir jābūt skaidri definētam, tā sasniegšanai jālieto tikai «jaunā» analizējamā objekta raksturojošie parametri. Gadījumā, ja sistēmas apmācības procesā ir jālieto noteikts vēsturisku notikumu periods (īsas laika rindas par noteiktu laika periodu, piemēram, preču pārdošanas apjoms par 2012. gadu), pasūtītājam jāpieprasa šīs iespējas integrēšana sistēmā. Pēc pasūtītāja iesniegto prasību definēšanas izpildītājs novērtē datu struktūru atbilstību izvirzītajam mērķim, ja atbilstība ir, var pāriet pie nākamā posma izpildes, ja nē, izpildītājs izveido atskaiti par atteikuma iemeslu un veic neatbilstības novēršanu sadarbībā ar pasūtītāju.



11. att. Prognozēšanas sistēmas prasību definēšana (I posms)

Prognozēšanas sistēmas scenārija izvēle

Prognozēšanas sistēmas scenārija izvēle sākas no lietotāja datubāzes (skat. 12. attēlu), no kuras ar vaicājuma palīdzību tiek atlasīta analīzei nepieciešamā informācija (vēsturiskas īsas laika rindas un to raksturojošie parametri), kas nonāk datu sadalīšanas blokā. Datu sadalīšanas blokā tiek izdalītas divas datu plūsmas: vēsturiskas laika rindas un to raksturojošie parametri. Vēsturiskām laika rindām ir iespēja lietot datu sagatavošanu un/vai datu pirmapstrādi. Datu sagatavošanas procesā tiek piedāvāts veikt periodu izlīdzināšanu, iztrūkstošu vērtību apstrādāšanu. Datu normalizācijas procesā tiek piedāvāts veikt dominējošu vērtību normalizāciju ar dzīves līkni. Līdzīgu objektu grupu noteikšana tiek veikta ar klasteru analīzi, lietojot modificētu k-vidējo sadalošo algoritmu, nosakot piemērotāko klasteru skaitu datu kopas klasterizācijai. Vadoties pēc uzdevuma specifikas, ir iespējams veikt klasterizācijas procesā iegūto klašu skaita transformāciju uz raksturojošo parametru klašu skaitu. Iespējams arī klasterizācijas rezultātus pārveidot paraugmodeļos, kas attēlo klasteru vidējās vērtības katrā laika periodā. Raksturojošo parametru datu sagatavošanas procesā tiek piedāvāts veikt iztrūkstošu vērtību apstrādāšanu, datu normalizāciju, lietojot z-novērtējuma normalizāciju ar standarta novirzi, un datu diskretizāciju. Raksturojošo parametru klasifikācijas procesam tiek piedāvātas vairākas iespējas:

a) klasifikācijai izmantot datu kopu, kas sastāv no raksturojošajiem parametriem un šo parametru raksturojošās klases (noteikta klīnisko, farmakoloģisko vai citu pētījumu rezultātā un tiek uzskatīta par «zelta standartu»);
b) ja nepieciešams, veikt raksturojošo parametru datu kopas sadalīšanu apakškopās, vadoties pēc atribūtu nozīmīguma sadalījuma, kas noteikts eksperimentāli, vai arī to ir definējis pasūtītājs;
c) klasifikācijai izmantot datu kopu, kas sastāv no raksturojošajiem parametriem un klasterizācijā iegūtās klases.

Dažādu datu tipu apvienošana tiek realizēta, ja izvēlēta pieeja, kas aprakstīta apakšpunktos «b» un «c». Sakarību noteikšanai starp dažādu struktūru datiem jeb klasifikācijai tiek piedāvāts izmantot *C4.5* algoritmu, atkarībā no uzdevuma specifikas klasifikatoru skaits var mainīties.
Nosacījumu likumu veidošana balstās uz vairāku procesu darbības rezultātu apkopošanu «JA … UN … TAD …»likumu veidā, kur:

a) pirmais nosacījums ir izveidotās apakškopas sadalošā parametra vērtība (piemēram, analizējamās vielas veids – atr. grupa, kas pievienots barībai; vērtību diapazons 1–5);
b) otrais nosacījums ir analizējamā objekta klase (iegūta klasterizācijas rezultātā);
c) slēdziens – atribūts risks ir objekta «zelta standarta» vērtība.

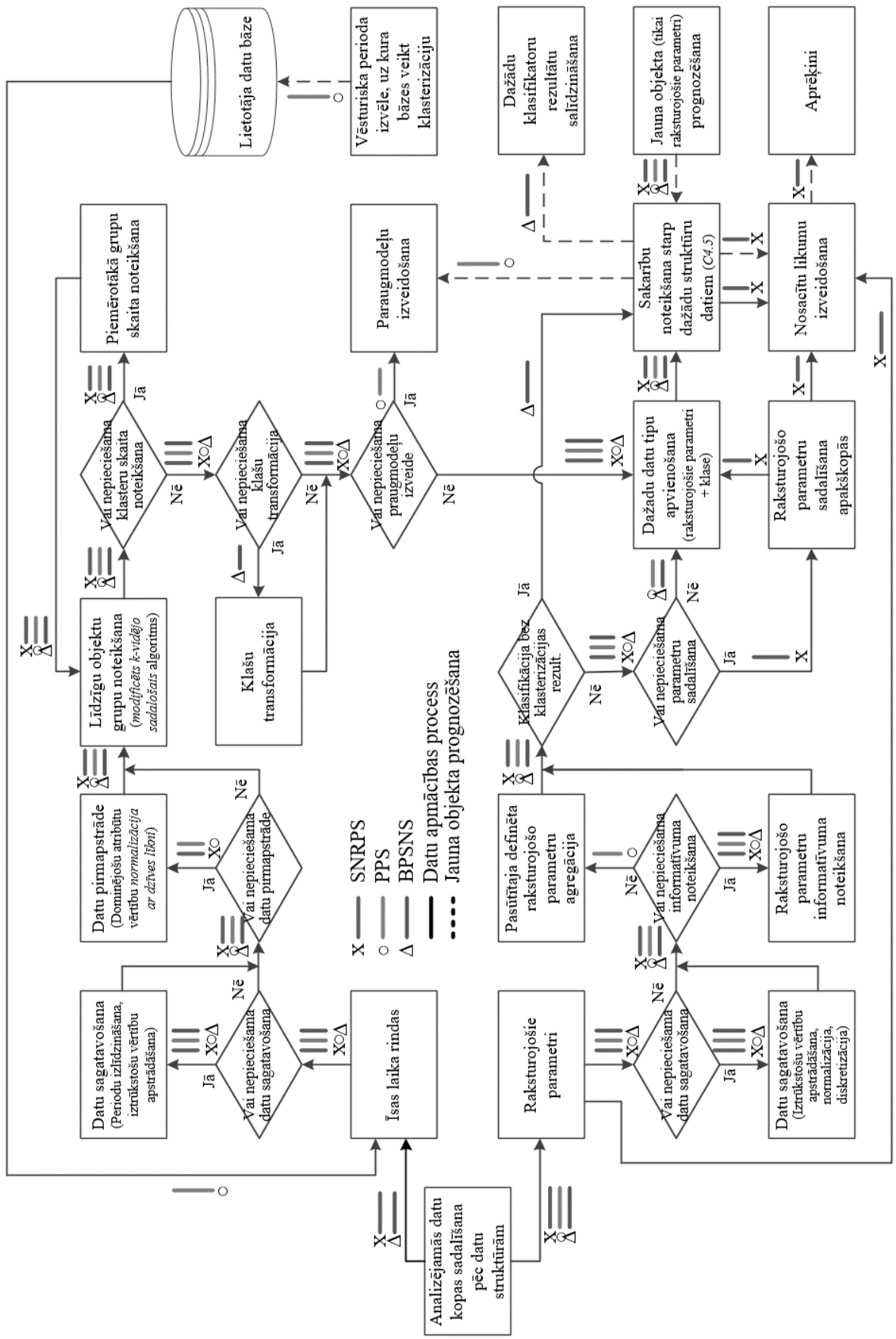
Katra objekta klasifikācijas rezultātā tiek izveidots viens nosacījumu likums, kas tiek saglabāts likumu datubāzē, piemēram, JA Grupa=1 UN Klase=C2 TAD Risks=35.
Promocijas darbā dažādām problēmvidēm izstrādāto sistēmu scenāriju attēlojumi ir parādīti 12. attēlā ar «○ ▬▬”» – pieprasījumu prognozēšanas sistēma (PPS), ar «× ▬▬» – sirds nekrozes riska prognozēšanas sistēma (SNRPS) un ar «Δ ▬▬» – baktēriju proliferācijas sindroma noteikšanas sistēma (BPSNS).
Jauna objekta prognozēšana tiek veikta, ievadot sistēmā objekta raksturojošos parametrus, kur ar klasifikatora palīdzību nosaka objekta piederību kādai no klasēm un tālāk atkarībā no uzdevuma specifikas tiek realizēta iegūto klasifikācijas rezultātu interpretācija:

a) ja prognozēšana tiek veikta, izmantojot paraugmodeļus, klasifikatora iegūtā klase norāda uz attiecīgā paraugmodeļa numuru, kas raksturo objekta pieprasījumu nākotnē noteiktā laika posmā. Lietotājam pastāv iespēja, ka pirms prognozes veikšanas sistēma tiek apmācīta uz konkrēta izvēlēta vēsturiska pieprasījuma perioda;
b) ja prognozēšana tiek veikta, izmantojot nosacītus likumus, klasifikatora iegūtā klase un sadalošā parametra vērtība norāda kādus nosacītus likums, ir jāatlasa no nosacītu likumu datubāzes. Balstoties uz atlasītajiem likumiem, kas raksturo nosacīto likumu parādīšanās biežumu, tiek veikti aprēķini izskaitļojot prognozēto vērtību;
c) ja prognozēšana tiek veikta, izmantojot datu kopas ar un bez klasterizācijas rezultātu izmantošanas, prognozes noteikšana balstās uz dažādu klasifikatoru rezultātu salīdzināšanu. Ja abi klasifikatori nosaka klasi «C0», prognoze nozīmē, ka nav jāveic turpmākas darbības. Ja rezultāts ir klase «C1» vai rezultāts netika noteikts, prognoze nozīmē, ka jāveic turpmākas darbības.

Scenārija izvēles patiesumu var noskaidrot tikai eksperimentāli, izveidojot sistēmas konceptu un veicot testēšanu ar pasūtītāja reālo datu kopu.

Prognozēšanas sistēmas izstrādāšana un testēšana

Pēc scenārija izvēles tiek izstrādāts prognozēšanas sistēmas koncepts, uz kura bāzes tiek veikti eksperimenti ar pasūtītāja reālu datu kopu, rezultātā nosakot klasifikatora precizitāti, jutīgumu un specifiskumu. Atsevišķos gadījumos, it īpaši medicīnā, var tikt lietotas arī papildu precizitātes novērtējuma vērtības, ko var aprēķināt no paplašinātās neskaidrības matricas. Ja analizējamā datu kopa sastāv no 200 un vairāk objektiem, izveidojamās sistēmas klasifikators tiek apmācīts ar apmācības kopas datiem un pēc tam pārbaudīts ar testēšanas kopas



12. att. Prognozēšanas sistēmas izveidošanas scenārija izvēle (II posms)

datiem. Ja analizējamās datu kopas objektu skaits ir mazāks par 200, tiek lietota 10-kārtu šķērsvalidācija, kas garantē novērtējuma drošumu pie neliela objektu skaita. Skaitlis 200 ir aptuvenā robeža. Ja iegūtais izstrādātās sistēmas klasifikācijas rezultātu novērtējums ir pietiekams, ko nosaka eksperts – sistēmas izstrādātājs, no šā koncepta var izveidot lietojumu, kas tālāk tiktu integrēts pasūtītāja informācijas datu nesējos.

Gadījumā, ja izstrādātās sistēmas iegūto klasifikācijas rezultātu novērtējums ir izrādījies nepietiekams, ko nosaka eksperts – sistēmas izstrādātājs, tad, balstoties uz vadlīnijām, iespējams izvēlēties citus risinājumus un veikt atkārtotu izveidotās sistēmas precizitātes novērtēšanu. Svarīgi ir pareizi izvēlēties piemērotākās datu sagatavošanas un datu pirmapstrādes pieejas, jo lielākās neprecizitātes datu ieguves procesā veido tieši nepiemērotu datu pirmapstrādes pieeju lietošana vai nekvalitatīvi pirmapstrādāta analizējamā datu kopa.

Piedāvātās vadlīnijas īsu laika rindu un raksturojošo parametru apstrādāšanai varētu būt arī nepilnīgas jaunu sistēmu izveidošanai dažādās problēmvidēs, jo tās veidos, balstoties uz iegūto pieredzi preču pieprasījuma prognozēšanā tirdzniecībā, sirds nekrozes riska prognozēšanā farmakoloģijā un baktēriju proliferācijas sindroma noteikšanā medicīnā. Tāpēc, veidojot jaunas sistēmas citās problēmvidēs, vadlīnijas noteikti varētu papildināt ar jauniem blokiem, iespējamiem notikumiem un citām datu plūsmām.

REZULTĀTI UN SECINĀJUMI

Promocijas darbā piedāvāta uz datu iegūšanu balstīta īsu laika rindu un to raksturojošo parametru apstrādes sistēma prognozēšanas uzdevumiem, kas izstrādāta dažādām problēmvidēm. Darba izstrādāšanas gaitā tika atrisināti šādi uzdevumi:

1. izanalizēti īsu laika rindu un to raksturojošo parametru apstrādes pamatprincipi, nosakot metodes, kas izmantojamas sistēmas realizācijai;
2. analīzes rezultātā noteiktas piemērotākas īsu laika rindu un to raksturojošo parametru datu pirmapstrādes metodes;
3. izstrādāts atbilstoši problēmvidei modificēts k-vidējo sadalošais algoritms, kas realizē īsu laika rindu klasterizāciju, nosakot piemērotāko klasteru skaitu. Izstrādātā algoritma modifikācija salīdzināta ar citu klasterizācijas algoritmu darbību;
4. izstrādāta klasterizācijas rezultātu un īsu laika rindu raksturojošo parametru datu apvienošanas tehnika, piedāvājot klašu transformācijas pieejas, lai izlīdzinātu klašu skaitu analizējamajās datu struktūrās;
5. izstrādātas tirdzniecības, farmakoloģijas un medicīnas nozaru prognozēšanas sistēmas, kas veic jauna objekta prognozi, balsoties tikai uz sistēmā ievadītajiem objekta raksturojošajiem parametriem;
6. veikts sistēmā lietoto klasterizācijas un klasifikācijas modeļu precizitātes novērtējums;
7. izstrādātas nosacījumu likumu veidošanas un lietošanas pieejas dažādām problēmvidēm;
8. izstrādātas vadlīnijas, uz kuru bāzes iespējams izveidot jaunas prognozēšanas sistēmas, kur kā datu avots kalpo īsas laika rindas un to raksturojošie parametri.

Preču pieprasījumu prognozēšanas sistēmā realizēta klasterizācijas rezultātu attēlošana ar paraugmodeļiem, kas pēc tam tika izmantoti arī prognozēšanā. Realizēta vairāku klasifikatoru izmantošana sirds nekrozes riska prognozēšanas sistēmā un baktēriju proliferācijas sindroma noteikšanas sistēmā. Izstrādāta klasifikācijas apmācības procesā iegūto rezultātu un farmakoloģisko pētījumu rezultātu apvienošanas pieeja, kas realizēta nosacījumu likumu veidā.

Izstrādātā klasterizācijas algoritma modifikācija, un visas trīs izveidotās prognozēšanas sistēmas tika eksperimentāli testētas, lai pārbaudītu izvirzītās hipotēzes, un to rezultāti ir šādi:

1. izveidotā sistēma dažādām problēmvidēm apstiprina pirmo izvirzīto hipotēzi, ka īsu laika rindu un to raksturojošo parametru apstrādes sistēmas izstrādāšana nodrošina grūti formalizējama uzdevuma atrisināšanu ar datu iegūšanas metodēm un algoritmiem;
2. otro hipotēzi apstiprina fakts, ka ir izstrādāts modificēts k-vidējo sadalošais algoritms, kas uzlabo piemērotākā klasteru skaita noteikšanu datu klasterizācijas procesā, analizējot īsas laika rindas. Izstrādātā algoritma modifikācija eksperimentāli pārbaudīta dažādās problēmvidēs un salīdzināta ar citu klasterizācijas algoritmu darbību;
3. trešo hipotēzi apstiprina fakts, ka ir izstrādātas preču pieprasījuma prognozēšanas, sirds nekrozes riska prognozēšanas un baktēriju proliferācijas sindroma noteikšanas sistēmas, kas tiek lietotas dažādās problēmvidēs: tirdzniecībā, farmakoloģijā un medicīnā, apstrādājot datu kopas, kas sastāv no īsam laika rindām un to raksturojošajiem parametriem.

Darba izstrādāšanas gaitā eksperimentāli tika noskaidrots, ka:

a) piemērotākā īsu laika rindu normalizācijas metode ir normalizācija ar dzīves līkni, kas uzrādīja labāko novērtējumu visās izstrādātajās sistēmās;
b) piemērotākais īsu laika rindu klasterizācijai ir modificēts k-vidējo sadalošais algoritms, pie tam tas saglabāja robustumu visā analizējamajā klasteru diapazonā un uzrādīja labākos novērtējumus visās izstrādātajās sistēmās;
c) piemērotākais klasifikators ir *C4.5* algoritms, jo tas uzrādīja labākos rādītājus klasifikācijas precizitātes novērtējumos;
d) datu kopām ar ierakstu skaitu līdz 200 klasifikācijas precizitātes novērtējumam ieteicams izmantot 10-kārtu šķērsvalidāciju, bet datu kopām virs 200 ierakstiem –datu sadalīšanu apmācības un testēšanas kopās attiecībā 70:30;
e) no statistiskā sadalījuma aprēķinot iespējamo sirds nekrozes risku, ieteicams lietot darba autora piedāvātu pieeju, kas balstās uz attālumu mēriem.

Promocijas darba sasniegtie zinātniskie rezultāti

1. Izstrādāta klasterizācijas algoritma modifikācija īsu laika rindu klasteru analīzei.
2. Izstrādāta pieeja dažāda skaita klašu struktūru transformācijai vienotā klašu struktūrā.
3. Izstrādātas vadlīnijas, kas norāda, kā veidojamas līdzīgas prognozēšanas sistēmas.

Promocijas darba sasniegtie praktiskie rezultāti

1. Izstrādāta preču pieprasījuma prognozēšanas sistēma, kas nosaka analizējamā objekta iespējamo pieprasījumu, kas tika pārbaudīta ar reāliem preču pieprasījuma datiem.

2. Izstrādāta sirds nekrozes riska prognozēšanas sistēma, kas nosaka analizējamā objekta sirds nekrozes riska iespējamo vērtību. Sistēma tika pārbaudīta ar reāliem farmakoloģisko pētījumu datiem.
3. Izstrādāta baktēriju proliferācijas sindroma noteikšanas sistēma tievajā zarnā, kas nosaka, vai analizējamajam indivīdam ir nepieciešams veikt laktozes testu. Sistēma tika pārbaudīta ar reāliem medicīnas datiem.

Promocijas darba izstrādāšanas gaitā par izveidoto īsu laika rindu un to raksturojošo parametru apstrādes sistēmu prognozēšanas uzdevumiem ir iegūti šādi secinājumi:

1. izstrādājot PPS un SNRPS, eksperimentāli tika pierādīts, ka datu normalizēšanai abās sistēmās labākus rezultātus uzrādīja pieeja normalizācija ar dzīves līkni;
2. raksturojošo parametru informatīvuma noteikšana un atlasīšana uzlabo klasifikācijas rezultātus, samazina analizējamās datu kopas apjomu un palielina algoritma izpildes ātrumu;
3. izstrādātā klasterizācijas algoritma modifikācija ir lietojama arī citu klasteru analīzes uzdevumu risināšanā, kur par datu avotu kalpo īsas laika rindas;
4. izstrādātā klasterizācijas algoritma modifikācija pieļauj dažādu novērtēšanas pieeju izmantošanu: 10-kārtu šķērsvalidāciju vai datu kopas sadalīšanu apmācības un testēšanas kopās;
5. izstrādātā klašu transformācijas pieeja nodrošina dažāda klašu skaita datu struktūru salīdzināšanu;
6. izstrādājot līdzīgas prognozēšanas sistēmas medicīnā, klasifikācijas precizitātes novērtējumam paralēli jāsalīdzina arī jutīgums un specifiskums, jo šie parametri būtiski ietekmē klasifikatora izvēli;
7. visās trijās izstrādātajās prognozēšanas sistēmās par piemērotāko klasifikācijas algoritmu eksperimentāli tika noteikts *C4.5* algoritms;
8. vairāku klasifikatoru lietošana sistēmu izstrādāšanā uzlabo klasifikatora kopējo precizitāti, kas eksperimentāli tika pierādīts ar sadalītu un nesadalītu datu kopām;
9. Izstrādātās īsu laika rindu un to raksturojošo parametru prognozēšanas sistēmas pamato un nodrošina grūti formalizējama uzdevuma atrisinājumu, lietojot klasterizācijas, to modifikācijas un klasifikācijas algoritmu apvienošanu.
10. izstrādātās īsu laika rindu un to raksturojošo parametru prognozēšanas sistēmas, realizē «jauna» objekta prognozēšanu, pamatojoties tikai uz šā objekta raksturojošajiem parametriem;
11. izstrādātās prognozēšanas sistēmas vadlīnijas palīdz nodrošināt jaunu prognozēšanas sistēmu izveidošanu, kā arī pieļauj iespēju papildināt esošas vadlīnijas.

Turpmākie pētījumi ir saistīti ar medicīnu – kuņģa vēža riska mazināšanas skrīninga sistēmas izstrādāšanu [41], kur promocijas darba gaitā izstrādātās prognozēšanas sistēmas vadlīnijas varētu izmantot kuņģa vēža skrīninga sistēmas izveidošanai, sasaistot neinvazīvu izmeklējumu rezultātus – īsas laika rindas ar respondenta raksturojošajiem parametriem. Šā veida skrīninga sistēmas varētu integrēt veselības aprūpes centros, kas palīdzētu nozares speciālistiem lēmumu pieņemšanā, nosakot diagnozes vai nosūtot pacientu pie speciālista izmeklējumu veikšanai.

IZMANTOTĀS LITERATŪRAS SARAKSTS

1. Datu ieguve: Pamati/ A. Sukovs, L. Aleksejeva, K. Makejeva u. c. – Rīga: Rīgas Tehniskā universitāte, SIA «Drukātava», 2006. – 130 lpp.
2. Dravnieks J. Matemātiskās statistikas metodes sporta zinātnē. – Rīga, 2004. – 76 lpp.
3. Klasifikācija un klasterizācija izplūdušajā vidē/ L. Aleksejeva, O. Užga-Rebrovs, A. Borisovs – Rīga: Rīgas Tehniskā universitāte, 2012. – 248 lpp.
4. Latvijas Zinātņu Akadēmijas TK ITTEA terminu datubāze/ Internets. – http://termini.lza.lv - Resurss apskatīts 2015. gada 22. janvārī.
5. Smotrovs J. Varbūtības teorija un matemātiskā statistika. – Rīga: Apgāds Zvaigzne ABC, 2004. – 264 lpp.
6. Varbūtību teorijas un matemātiskās statistikas elementi medicīnas studentiem/ U. Teibe, U. Berķis – Rīga: AML/RSU, 2001. – 88 lpp.
7. Alba E., Mendoza M. Bayesian forecasting methods for short time series// Foresight: The International Journal of Applied Forecasting, International Institute of Forecasters. – 2007. – Issue 8. – pp. 41–44.
8. Armstrong J. S., Collopy F. Error measures for generalizing about forecasting methods: Empirical comparisons// International Journal of Forecasting 8. – 1992. – pp. 69–80.
9. Armstrong J. S., Collopy F., Yokum J. T. Decomposition by causal forces: A procedure for forecasting complex time series// International Journal of Forecasting 21. – 2005. – pp. 25–36.
10. Armstrong J. S., Fildes R. Correspondence on the selection of error measures for comparisons among forecasting methods// International Journal of Forecasting 14. – 1995. – pp. 67–71.
11. Berndt D. J., Clifford J. Using dynamic time warping to find patterns in time series// Association for the Advancement of Artificial Intelligence, Workshop on Knowledge Discovery in Databases (AAAI), – 1994. – pp. 229–248.
12. Berry M. W., Browne M. Lecture notes in data mining. – World Scientific Publishing Co. Pte. Ltd., 2006. – 222 p.
13. Boyer K. K., Verma R. Operations and Supply Chain Management for the 21st Century. – USA: South-Western Cengage Learning, 2010. – 560 p
14. Clark P., Niblett T. The CN2 induction algorithm. Machine Learning, 3(4), – 1989. – pp. 261–283.
15. Cost S., Salzberg S. A Weighted Nearest Neighbor Algorithm for Learning with Symbolic Features // Machine Learning. – 1993. – Vol.10. – pp. 57–78.
16. Dambrova M., Liepinsh E., Kalvinsh I., Mildronate. Cardioprotective Action through Carnitine-Lowering Effect// Trends Cardiovasc. Med. – 2002. – Vol.12. – pp. 275–279.
17. Dellaert F. The Expectation Maximization Algorithm. College of Computing, Georgia Institute of Technology, Technical Report number GIT-GVU-02-20, – 2002.
18. Dempster A. P., Laird N. M., Rubin D. B. Maximum likelihood from incomplete data via the EM algorithm// Journal of the Royal Statistical Society. Series B (Methodological). – 1977. – Vol.39 (1). – pp. 1–38.
19. Devisscher M., De Baets B., Nopens I. Pattern discovery in intensive care data through sequence alignment of qualitative trends: proof of concept on a diuresis dataset// Appearing in the Proceedings of the ICML/UAI/COLT 2008 Workshop on Machine Learning for Health-Care Applications, Helsinki, Finland. – 2008.
20. Donald I. P., Kitchingmam G., Donald F., Kupfer, R. M. The diagnosis of small bowel bacterial overgrowth in elderly patients// J. Am. Geriatr. Soc. – 1992. – Vol. 40 (7). – pp. 692–696.
21. Ernst J., Nau G. J., Bar-Joseph Z. Clustering short time series gene expression data// Bioinformatics. – 2005. – Vol.21. – pp. 159–168.

22. Flores J. J., Loaeza R. Financial time series forecasting using a hybrid neural-evolutive approach// Proceedings of the XV SIGEF International Conference, Lugo, Spain. – 2009. – pp. 547–555.
23. Gardner E. S. Jr., Exponential Smoothing: The State of Art// Journal of Forecasting. – 1985. – Vol.4. – pp. 1–28.
24. Golub T. R., Slonim D. K., Tamayo P., Huard C., Gaasenbeek M., Mesirov J. P., Coller H., Loh M. L., Downing J. R., Caligiuri M. A., Bloomfield C. D., Lander E. S. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring// Science. – 1999. – Vol. 286 (5439). – pp. 531–537.
25. Grabusts P. The choice of metrics for clustering algorithms// Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. – 2011. – Vol. 2. – pp. 70–76.
26. Graves S. C., Kletter D. B., Hetzel W. B. A dynamic model for requirements planning with application to supply chain optimization// Operation Research. – 1998. – Vol. 46 (3). – pp. 35–49.
27. Hall M. A. Correlation-based feature selection for machine learning/ Doctoral Thesis, Hamilton: University of Waikato. – 1999. – 178 p.
28. Han J., Kamber M. Data Mining: Concepts and Techniques. Second Edition. – Morgan Kaufmann, Elsevier Inc., 2006. – 800 p.
29. Hsieh C. H., Anderson C., Sugihara G. Extending nonlinear analysis to short ecological time series// Am Nat. – 2008. – Vol. 171 (1). – pp. 71–80.
30. Kirshners A. Clustering-based behavioural analysis of biological objects// Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. – 2011. – Vol. 2. – pp. 24–32.
31. Kirshners A., Borisov A. A Comparative analysis of short time series processing methods// Scientific Journal of Riga Technical University, Information Technology and Management Science. – 2012. – Vol. 15. – pp. 65–69.
32. Kirshners A., Borisov A. Analysis of short time series in gene expression tasks// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2010. – Iss. 5, Vol. 44. – pp. 144–149.
33. Kirshners A., Borisov A. Multilevel classifier use in a prediction task// Proceedings of the 17th International Conference on Soft Computing. – 2011. – pp. 403–410.
34. Kirshners A., Borisov A. Processing short time series with data mining methods// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2011. – Iss. 5, Vol. 49. – pp. 91–96.
35. Kirshners A., Borisov A., Parshutin S. Robust cluster analysis in forecasting task// Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012). – 2012. – pp. 77–81.
36. Kirshners A., Kornienko Y. Time-series data mining for e-service application analysis// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2009. – Iss. 5, Vol. 40. – pp. 94–100.
37. Kirshners A., Kuleshova G., Borisov A. Demand forecasting based on the set of short time series// Scientific Proceedings of Riga Technical University, Information Technology and Management Science. – 2010. – Iss. 5, Vol. 44. – pp. 130–137.
38. Kirshners A., Liepinsh E., Parshutin S., Kuka J., Borisov A. Risk prediction system for pharmacological problems// Automatic Control and Computer Sciences. – 2012. – Vol.46, No.2. – pp. 57–65.

39. Kirshners A., Parshutin S. Application of data mining methods in detecting of bacteria proliferation syndrome in the small intestine// In: European Conference on Data Analysis 2013: Book of Abstracts: European Conference on Data Analysis 2013. – 2013. – pp. 139–139.
40. Kirshners A., Parshutin S., Borisov A. Combining clustering and a decision tree classifier in a forecasting task// Automatic Control and Computer Science. – 2010. – Vol. 44, No. 3. – pp. 124–132.
41. Kirshners A., Parshutin S., Leja M. Research in application of data mining methods to diagnosing gastric cancer// LNAI 7377. Proceedings of the 12th Industrial Conference on Data Mining ICDM'2012. – 2012. – pp. 24–37.
42. Kirshners A., Sukov A. Rule induction for forecasting transition points in product life cycle data// Scientific Proceedings of Riga Technical University, Information Technology and Management Sciences. – 2008. – Iss. 5, Vol. 36 – pp. 170–177.
43. Kohavi R. A study of cross-validation and bootstrap for accuracy estimation and model selection// Proceedings of the 14th International Conference on Artificial Intelligence (IJCAI-95). – San Mateo, CA: Morgan Kaufman. – 1995. – pp. 1137–1143.
44. Kohavi R., Quinlan J. R. Decision-tree discovery// Handbook of Data Mining and Knowledge Discovery. – Klosgen W., Zytkow J. M., Eds. – Oxford: Oxford University Press. – 2002. – pp. 267–276.
45. Koza J. R. Genetic Programming: On the Programming of Computers by Means of Natural Selection. – Cambridge, MA: The MIT Press, 1992. – 840 p.
46. Lembcke B. Atemtests bei Darmkrankheiten und in der gastroenterologischen Funktionsdiagnostik. Schweiz// Rundschau Medizin (Praxis). – 1997. – Vol. 86. – pp. 1060–1067.
47. Liepinsh E., Vilskersts R., Loca D., Kirjanova O., Pugovichs O., Kalvinsh I., Dambrova M. Mildronate, an inhibitor of carnitine biosynthesis, induces an increase in gamma-butyrobetaine contents and cardioprotection in isolated rat heart infarction// J Cardiovasc Pharmacol. – 2006. – Vol. 48 (6). – pp. 314–319.
48. Liepinsh E., Vilskersts R., Skapare E., Svalbe B., Kuka J., Cirule H. et al. Mildronate decreases carnitine availability and up-regulates glucose uptake and related gene expression in the mouse heart// Life Sci. – 2008. – Vol. 83. – pp. 613–619.
49. Liepinsh E., Vilskersts R., Zvejniece L., Svalbe B., Skapare E., Kuka J. et al. Protective effects of mildronate in an experimental model of type 2 diabetes in Goto-Kakizaki rats// British Journal of Pharmacology. – 2009. – Vol. 157. – pp. 1549–1556.
50. Lupascu A., Gabrielli M., Lauritano E. C., Scarpellini E., Scantoliquido A., Cammarota G., Flore R., Tondi P., Pola P., Gasbarrini G., Gasbarrini A. Hydrogen glucose breath test to detect small intestinal bacterial overgrowth: a prevalence case-control study in irritable bowel syndrome// Aliment Pharmacol Ther. – 2005. – 22 (11–12). – pp. 1157–1160.
51. McLachlan G., Krishnan T. The EM algorithm and extensions, 2nd edition. Wiley series in probability and statistics. – John Wiley & Sons, 2008. – 400 p.
52. Montgomery D. C., Jennings C. L., Kulachi M. Introduction to Time Series Analysis and Forecasting. – Wiley-interscience, 2008. – 445 p.
53. Parshutin S., Aleksejeva L., Borisov A. Forecasting product life cycle phase transition points with modular neural networks based system// Proceedings of 9th Industrial Conference on Data Mining ICDM`2009, Springer-Verlag. – 2009. – LNAI 5633. – pp. 88–102.
54. Parshutin S., Kirshners A. Intelligent agent technology in modern production and trade management// Efficient Decision Support Systems: Practice and Challenges – From Current to Future/ Book Chapter. INTECH. – 2011. – pp. 21–42.

55. Parshutin S., Kirshners A. Research on clinical decision support systems development for atrophic gastritis screening// Expert Systems with Applications. – 2013. – Vol. 40, Iss. 15. – pp. 6041–6046.
56. Posserud I., Stotzer P. O., Bjornsson E. S., Abrahamsson H., Simren M. Small intestinal bacterial overgrowth in patients with irritable bowel syndrome// Gut. – 2007. – Vol.56(6). - pp. 802-808.
57. Pyle D. Data Preparation for Data Mining. – San Francisco etc.: Morgan Kaufmann, 1999. – 540 p.
58. Quinlan J. R. C4.5: Programs for Machine Learning. – San Mateo: Morgan Kaufmann Pub., 1993. – 302 p.
59. Russell S. J., Norvig P. Artificial Intelligence: A Modern Approach – Prentice-Hall, Inc., 1995. – 932 p.
60. Salam A. Najim, Zakaria A. M. Al-Omari, Samir M. Said. On the application of artificial neural network in analyzing and studying daily loads of Jordan power system plant// Computer Science and Information Systems. – 2008, – Vol. 5, Iss. 1. – pp. 127–136.
61. Sjakste N., Gutcaits A., Kalvinsh I. Mildronate: An antiischemic drug for neurological indications// CNS Drug Reviews. – 2005. – Vol. 11 (2). – pp. 151–168.
62. Starzyk, J. A., Haibo H., Yue L. A Hierarchical Self-organizing Associative Memory for Machine Learning// Advances in Neural Networks. – 2007. – pp. 413–423.
63. Tan P. N., Steinbach M., Kumar V. Introduction to Data Mining. – Boston: Pearson Addison-Wesley, 2006. – 769 p.
64. Thomassey S., Fiordaliso A. A hybrid sales forecasting system based on clustering and decision trees// Decision Support Systems. – 2006. – Vol. 42, Iss. 1. – pp. 408–421.
65. Thomassey S., Happiette M. A neural clustering and classification system for sales forecasting of new apparel items// Applied Soft Computing. – 2007. – Vol. 7. – pp. 1177–1187.
66. Thomassey S., Happiette M., Castelain J. A global forecasting support system adapted to textile distribution// International Journal of Production Economics. – 2005. – Vol. 96. – pp. 81–95.
67. Thomassey S., Happiette M., Castelain J. A short and mean - term automatic forecasting system – application to textile logistics// European Journal of Operations Research. – 2005. – Vol. 161. – pp. 275–284.
68. Toshniwal D., Joshi R. C. Similarity search in time series data using time weighted slopes// Informatica, An International Journal of Computing and Informatics. – 2005. – Vol. 29, No. 1. – pp. 79–88.
69. Ward J. H., Jr. Hierarchical Grouping to Optimize an Objective Function// Journal of the American Statistical Association. – 1963. – Vol. 58. – pp. 236–244.
70. Witten I. H., Frank E. Data mining: Practical machine learning tools and techniques (Second edition). – San Francisco, CA: Morgan Kaufmann, 2005. – 560 p.
71. Wiener N. Cybernetics: Or Control and Communication in the Animal and the Machine. – Boston, MA: Technology Press, 1948. – 219 p.
72. Wang X., Wu M., Li Z., Chan C. Short time-series microarray analysis: Methods and challenges// BMC Systems Biology – 2008. – 2:58. – pp. 1–6.
73. Wu X., Kumar V., Quinlan J. R., et al. Top 10 algorithms in data mining// Knowl. Inf. Syst. – 2007. – 14. – pp. 1–37.
74. Zhu W., Zeng. N., Wang N. Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementation// NESUG Proceedings: Health Care and Life Sciences, Baltimore, Maryland. – 2010. – pp. 1–9.
75. Zurada J. M. Introduction to Artificial Neural Systems. – West: St. Paul, MN, 1992. – 679 p.
76. Барсегян А., Куприянов М., Степаненко В. Технологии анализа данных: Data Mining, Visual Mining, Text Mining, OLAP. – Санкт-Петербург: БХВ - Петербург, 2007. – 384 с.
77. Гринглаз Л., Копытов Е. Математическая статистика с примерами решения на компьютере: Учеб. Пособие. – 2-е изд. – Рига: ВШЭК, 2002. – 326 с.