RĪGAS TEHNISKĀ UNIVERSITĀTE
Datorzinātnes un informācijas tehnoloģijas fakultāte
Informācijas tehnoloģijas institūts

Inese POĻAKA
Doktora studiju programmas „Informācijas tehnoloģija" studente

KLAŠU BLĪVUMA STRUKTŪRAS IZMANTOŠANA LĒMUMU KOKU KLASIFIKATORU ANSAMBĻU EVOLUCIONĀRĀ INDUKCIJĀ

Promocijas darba kopsavilkums

Zinātniskais vadītājs
Dr.habil.sc.comp., profesors
A. Borisovs

Rīga 2014

UDK 004.85.021(043.2)
Po 185 k

Poļaka I. Klašu blīvuma struktūras izmantošana lēmumu koku klasifikatoru ansambļu evolucionārajā indukcijā. Promocijas darba kopsavilkums.-R.:RTU Izdevniecība, 2014. – 38 lpp.

Iespiests saskaņā ar 2014.gada 7. februāra ITI padomes sēdes lēmumu, protokols Nr. 14-2.

Šis darbs izstrādāts ar Eiropas Sociālā fonda atbalstu projektā „Atbalsts RTU doktora studiju īstenošanai”.

ISBN

PROMOCIJAS DARBS
IZVIRZĪTS INŽENIERZINĀTŅU DOKTORA GRĀDA IEGŪŠANAI RĪGAS TEHNISKAJĀ UNIVERSITĀTĒ

Promocijas darbs inženierzinātņu doktora grāda iegūšanai tiek publiski aizstāvēts 2014.gada 19. maijā plkst. 14:30 Rīgas Tehniskās universitātes Datorzinātnes un informācijas tehnoloģijas fakultātē, Meža ielā 1, 3. korpusā, 202. auditorijā.

OFICIĀLIE RECENZENTI

Profesors, Dr.habil.sc.ing. Zigurds Markovičs
Rīgas Tehniskā universitāte, Latvija

Asociētais profesors, Dr.dat. Jānis Zuters
Latvijas Universitāte, Latvija

Profesors, Dr.sc. Aleksandrs Božeņuks
Dienvidu Federālās Universitātes Taganrogas Tehnoloģiju Institūts, Krievija

APSTIPRINĀJUMS

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

Inese Poļaka …….
paraksts

Datums …….

Promocijas darbs uzrakstīts latviešu valodā, satur ievadu, 5 nodaļas, rezultātu analīzi un secinājumus, 48 tabulas, 37 attēlus, kopā – 141 lappusi. Literatūras sarakstā ir 76 vienības.

SATURS

VISPĀRĒJS DARBA RAKSTUROJUMS

Problēmsfēra

Pasaules medicīna virzās no simptomātiskās diagnostikas uz sistēmbioloģijas pieeju, kurā diagnostikai, ārstēšanai, šo procesu uzraudzībai u.c. tiek izmantoti tādi cilvēka bioloģiskie testi kā ģenētiskais vai imunoloģiskais profils. Šajā pieejā jāsaskaras ar dimensijas ziņā lieliem datu apjomiem, kuros turklāt ir salīdzinoši niecīgs ierakstu skaits, jo testi ir dārgi un vēl nav ieviesti ikdienas veselības aprūpē. Tomēr šie dati ir jāanalizē, lai noteiktu bioloģiskos marķierus, kas norāda uz slimības procesiem. Tāpēc datu analīzes process iekļauj matemātiskās un statistiskās analīzes metodes (jeb biostatistiku) un citas intelektuālas datu analīzes pieejas, to skaitā arī datu ieguves un mašīnapmācības metodes (jeb bioinformātiku).

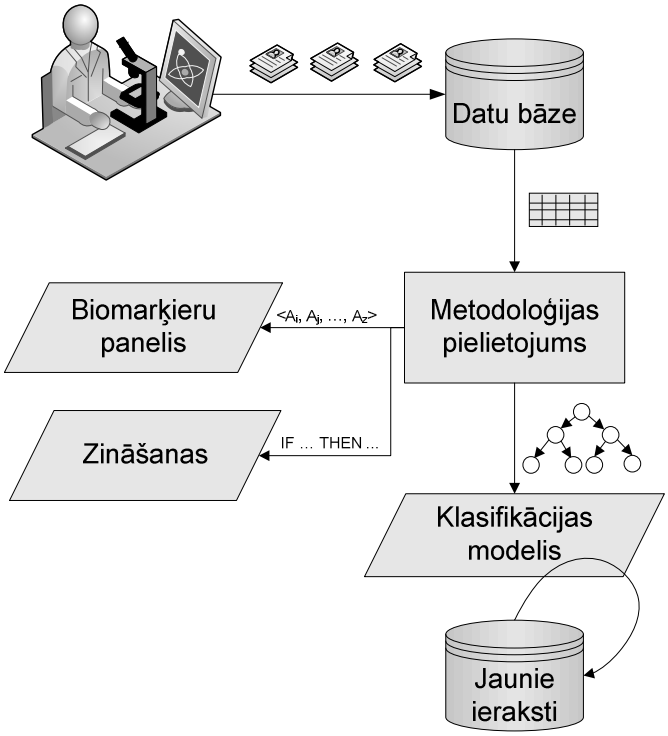
Aktualitāte

Cilvēka genoma projekts tika uzsākts jau 1990. gadā, taču pilnā cilvēka genoma atšifrēšana tika veikta tikai 2003. gadā. Kopš tā laika zinātnieki var izmantot gēnus par slimību raksturojošiem marķieriem, nosakot konkrētos gēnus un to ekspresijas (jeb izteiktības) līmeņus veselos un slimos audos. Šo pārmaiņu rezultātā tika attīstītas gēnu mikromasīvu tehnoloģijas, kas ļāva izmērīt vairākus tūkstošus gēnu viena testa laikā. Šī tehnoloģija arī ļāva attīstīt cilvēka imūnsistēmas reakcijas pētījumus, izmantojot proteīnu mikromasīvus. Visu šo tehnoloģiju attīstība sekmēja jaunas, uz sistēmbioloģiju balstītas perspektīvas rašanos diagnostikā, slimības un ārstēšanas procesu uzraudzībā un prognozēšanā. Taču šajā sfērā vēl ir daudz nezināmā – par gēnu un proteīnu nozīmi, funkcijām un savstarpējām sakarībām, tāpēc pēdējās dekādes laikā bioinformātika ir guvusi milzīgu popularitāti, bet jaunās tehnoloģijas, pieejas un metodes vēl nav pietiekami precīzas un informatīvas, un bioinformātikas pētījumu datos joprojām pastāv daudz neatklātas informācijas un zināšanu.

Problēmas nostādne

Cilvēka spēkos nav vienam izanalizēt šādu datu daudzumu, tāpēc pastāv nepārtraukts pieprasījums pēc skaitļošanas tehnoloģijām un metodēm. Lielākā daļa datu analīzes metožu, kas izmantotas līdzīgos pētījumos, ir uzrādījušas labus rezultātus atsevišķās datu kopās, bet nepastāv vienas dominējošas metodes, kas vienlīdz labi darbojas ar visiem datiem. Atbalsta vektoru mašīnas (Support Vector Machines – SVM) un naivā Baijesa metode (NB) dažos pētījumos ir uzrādījušas labu precizitāti, taču arī to sniegums nav stabils. Arī rezultātu interpretācijas iespēja, izmantojot tūkstošus atribūtu, ir tuva nullei. Pastāv neliela gēnu vai proteīnu grupa, kas ir nozīmīga testu mērījumu un rezultējošās klases savstarpējās sakarībās, taču iepriekš minētās metodes nesniedz ne šādas zināšanas, ne samazināto nozīmīgo gēnu/proteīnu (biomarķieru)

paneli. Cita metode, kas ir populāra bioinformātikas pētījumos ir gadījuma meži (Random Forest – RF) un citas metodes, kas balstītas uz induktīvajiem lēmumu koku klasifikatoriem. Lai gan šīs metodes uzrāda nedaudz sliktākas klasifikācijas precizitātes, iegūtie klasifikācijas modeļi ir viegli interpretējami, satur nelielu informatīvo atribūtu kopu (biomarķieru paneli), kā arī attēlo sakarības starp šiem biomarķieriem. Zināšanas, kas tika iegūtas, pētot problēmvidi un tās pašreizējo stāvokli, kā arī dažādus literatūras avotus, noveda pie secinājuma, ka šo datu analīzei jāizmanto uz lēmumu kokiem balstītas klasifikācijas metodes. Taču tās ir jāpielāgo darbam ar specifiskajiem datiem un datu struktūras apraksta izmantošanai klasifikatora veidošanā. Izstrādātās metodoloģijas pielietojums parādīts 1. attēlā.



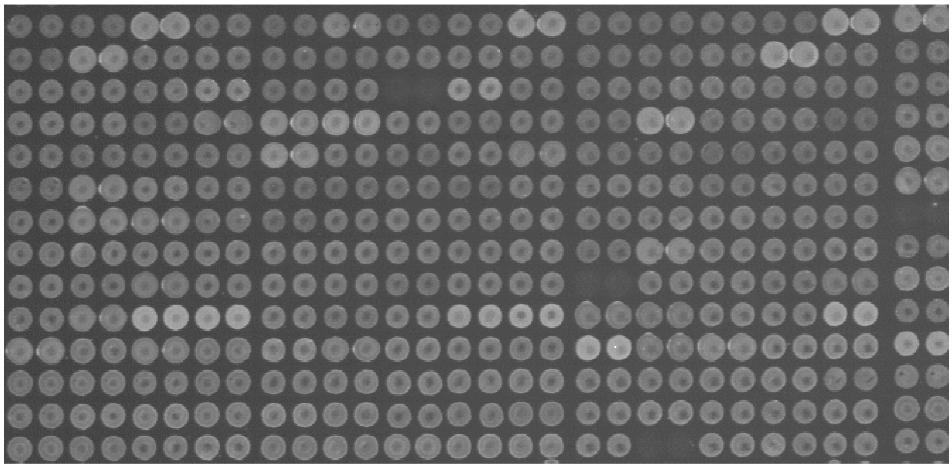
1. attēls. Metodoloģijas pielietošanas process

Motivācija

Iemesls šīs problēmas risināšanai, izmantojot datu ieguves un mašīnapmācības metodes, ir balstīts uz šo metožu specifiku:

- Tās darbojas ar neparametriskiem datiem, neizvirzot specifiskas prasības pret to parametriem;
- Tās labi darbojas ar datiem, kuriem ir augsta dimensionalitāte un dažās metodēs ir iestrādāta atribūtu atlases (dimensionalitātes samazināšanas) pieeja;
- Neskatoties uz to, ka mazs ierakstu skaits samazina precizitāti, tās labi tiek galā ar šādām datu kopām.

Specifiskie dati tiek iegūti, analizējot pacientu bioloģisko materiālu un izmantojot mikromasīvus, kas satur vairākus tūkstošus gēnu vai antivielu (noskenēts mikromasīvs izskatās tā, kā tas parādīts 2. attēlā).



2. attēls. Noskenēta mikromasīva daļa

Skanētā mikromasīva dati tiek transformēti par decimālskaitļiem, kas norāda zaļās un sarkanās krāsas attiecību (procesā gēni un proteīni ir marķēti,ar speciālām krāsām). Datu kopa tiek transformēta par tabulu, kurā rindiņas apzīmē pacientus un to testu rezultātus (x_{ij}), kuros nepārtrauktās atribūtu skaitliskās vērtības apzīmē gēnu vai antivielu izteiktību (gēni vai antivielas norādīti kolonnās (skatīt 1. tabulu)).

1. tabula

Datu kopa ar transformētajiem skenēto mikromasīvu datiem

ID	$Gēns_1$	$Gēns_2$	$Gēns_3$	...	$Gēns_j$
$Pacients_1$	x_{11}	x_{12}	x_{13}		x_{1j}
$Pacients_2$	x_{21}	x_{22}	x_{23}		x_{2j}
$Pacients_3$	x_{31}	x_{32}	x_{33}		x_{3j}
...	...	...	...	...	...
$Pacients_n$	x_{n1}	x_{n2}	x_{n3}		x_{nj}

Lai atrisinātu šādu diagnostisko uzdevumu ar augstu precizitāti un kvalitāti, ir jāizstrādā pielāgota metodoloģija. Tai jāmeklē visprecīzāko klasifikācijas modeli un jāizmanto datu iekšējās struktūras īpašības pat tad, ja tas prasa vairāk laika. Tas ir pieļaujams, jo šajā gadījumā metodoloģija nav plānota pielietošanai reāllaikā arvien papildinātiem datiem.

Pētījuma mērķis un uzdevumi

Pētījuma **mērķis** ir izstrādāt bioinformātikas klasifikācijas metodoloģiju, kas izmanto klašu iekšējo struktūru un ģenētiskos algoritmus klasifikācijas modeļu atrašanai.

Pētījuma uzdevumi, kas izvirzīti uzdevuma sasniegšanai ir sekojoši:

- Izanalizēt citas metodes un pieejas, kas izmantotas līdzīgos pētījumos šajā sfērā un aprakstītas pieejamajā literatūrā;
- Izstrādāt pieeju, kas ļauj attēlot klašu iekšējo struktūru un to izmantot klasifikatoru veidošanā;
- Izstrādāt hibrīdu metodi, kas ļauj atrast kvazi-optimālu lēmumu koku klasifikatoru ansambli, izmantojot ģenētiskos algoritmus;
- Izstrādāt vienotu metodoloģiju, kas veidotu diagnostiskos modeļus, izmantojot iepriekšējos divos punktos izstrādātās metodes;
- Novērtēt izstrādātās metodoloģijas, metožu un pieeju efektivitāti, salīdzinot tās ar pieejamo alternatīvo metožu rezultātiem.

Pētījuma objekts un subjekts

Pētījuma objekts ir biomedicīniskā diagnostika, bet pētījuma **subjekts** – datu ieguves un mašīnapmācības metodes.

Pētījuma hipotēzes

Informācijas tehnoloģiju pētījums ir balstīts uz biomedicīnas faktu: vienai slimībai ar vienu un to pašu simptomātisko izpausmi var būt dažādi biomedicīniskie gēnu vai antivielu profili, kas nedaudz atšķiras un var tikt aprakstīti ar nelielu biomarķieru paneli.

Biomedicīnas diagnostiskā modeļa izveides metodoloģijas, kas balstīta uz datu ieguvi un mašīnapmācību, izstrādes laikā izvirzītas sekojošas hipotēzes:

1. Pastāv zīmīgi mazāka atribūtu kopa, kura satur lielāko daļu zināšanu par slimību (tās diagnostikas iezīmēm), kas atrodas datu kopā.
2. Izmantojot informāciju par datu iekšējo struktūru (nosakot atšķirīgos genotipus), tiek uzlabota klasifikācijas algoritmu precizitāte, veidojot sarežģītus klasifikatorus, kas ne tikai satur informāciju par atšķirīgajām klasēm, bet arī par datu struktūru.
3. Galīgā pārmeklēšanas telpā ģenētiskie algoritmi spēj atrast kvazi-optimālus klasifikatorus, kas ir ne tikai precīzi, bet arī viegli interpretējami tālākai izmantošanai.
4. Klasifikatoru ansambļi ir precīzāki specifiskajos bioinformātikas datos, ja jauni, ansamblī iekļaujami atomāri klasifikatori pievieno jaunu nozīmīgu klasifikācijas informāciju.

Pirmā hipotēze tieši balstās uz medicīniskā fakta. Tā tiks pārbaudīta, realizējot eksperimentu sēriju ar mazākām atribūtu kopām (līdz 200 atribūtu) un novērtējot palikušās zināšanas, izveidojot klasifikācijas modeļus un novērtējot to precizitāti, salīdzinot ar precizitāti, kas iegūta pilnajās datu kopās. Ja klasifikācijas precizitāte

mazākajās datu kopās būs augstāka vai vienāda ar pilnajās datu kopās iegūto, hipotēze tiks uzskatīta par pierādītu un patiesu.

Otrā hipotēze balstās uz fakta, ka datu iekšējās struktūras apraksts ļauj precīzāk aprakstīt klases klasifikatoros. Ar to saistītais trūkums ir tāds, ka datu kopas satur ļoti maz ierakstus, bet no šiem ierakstiem ir jāgūst informācija ne tikai par oriģinālo klašu specifiku, bet arī par klašu iekšējās struktūras aprakstu, kas nozīmē, ka datu ieguves procesā no tā jau nelielās datu kopas jāiegūst papildus zināšanas. Tas palielina klasifikācijas modeļu sarežģītību un rezultātā palielina pārapmācības iespēju, iekļaujot ierakstiem specifisku informāciju, kas neattiecas uz klasēm. Otrā hipotēze tiks pārbaudīta, salīdzinot tās klasifikācijas precizitātes, kuras iegūtas sākotnējās datu kopās, ar tām, kas iegūtas klasifikatoru indukcijā, izmantojot klašu iekšējās struktūras aprakstu. Ja klasifikācijas rezultāti, kuri iegūti izmantojot klašu iekšējās struktūras aprakstu, būs labāki, hipotēze tiks uzskatīta par pierādītu un patiesu.

Trešā hipotēze balstās uz pieņēmuma, ka pastāv optimāli klasifikatori, kurus var iegūt, izmantojot klasiskos algoritmus. Tāpēc pārmeklējot iespējamo klasifikatoru kopu, izmantojot ģenētiskos algoritmu, tajā var tikt atrasti kvazi-optimāli klasifikatori. Šī hipotēze tiks pārbaudīta, veicot datu kopu klasifikāciju, izmantojot klasiskos algoritmus un ģenētiskos algoritmus, kas meklē optimālos lēmumu koku klasifikatorus un to ansambļus. Ja atrastie klasifikatori un/vai to ansambļi būs precīzāki vai tikpat precīzi, uzlabojot klasifikatoru interpretējamību (caurspīdīgumu biomedicīnas ekspertam), šī hipotēze tiks uzskatīta par pierādītu un patiesu.

Ceturtā hipotēze ir balstīta uz pieņēmumu par datu kopas sarežģītību – sarežģītākās datu kopās pastāv nepieciešamība pēc klasifikatoriem, kas spēj izskaidrot vairāk zināšanu un izveidot sarežģītākus modeļus. Viena koka klasifikatori kļūst sarežģītāki, ja tos palielina, bet šādā veidā notiek arī klasifikatoru pārapmācība nelielā ierakstu skaita dēļ. Lēmumu koku ansambļi spēj veidot atsevišķus klasifikatorus, kas ir vienkārši, bet satur sarežģītas zināšanas, kad tiek apskatīti ansamblī. Šī hipotēze tiks pārbaudīta, salīdzinot atsevišķu lēmumu koku klasifikatoru precizitāti ar lēmumu koku klasifikatoru ansambļu precizitāti. Ja ansambļi būs precīzāki, hipotēze tiks uzskatīta par pierādītu un patiesu.

Pētījuma metodes

Pētījums balstās uz matemātisko un statistisko analīzi, datu ieguves, mašīnapmācības, ģenētisko algoritmu un eksperimentālo pētījumu metodēm. Tāpat tiek izmantota literatūras analīze, lai gūtu zināšanas par citiem pētījumiem un pastāvošo situāciju.

Zinātniskā novitāte

Pētījuma zinātniskā novitāte balstās uz izstrādāto metodoloģiju. Tajā ietvertas divas īpaši tai izstrādātas metodes, kas var tikt izmantotas līdzīgos bioinformātikas pētījumos. Metodes ir sekojošas:

- Izstrādātā pieeja izmantot datu struktūras īpašības klasifikācijā ir īstenota klašu dekompozīcijas metodē, kas ļauj attēlot klases iekšējo struktūru tādā veidā, ka klasiskās datu ieguves un mašīnapmācības metodes var to izmantot klasifikatoru veidošanā.
- Ģenētiskais algoritms tika modificēts un adaptēts darbam ar lēmumu koku klasifikatoriem un to ansambļiem. Izstrādātā metode var tikt izmantota gan vienkāršu klasifikatoru, gan to ansambļu meklēšanai.

Praktiskais nozīmīgums

Izstrādātā metodoloģija, kā arī atsevišķās metodes var tikt izmantotas, lai risinātu bioinformātikas uzdevumu ar līdzīgām īpašībām – atrast datos sakarības un zināšanas (diagnostiskas, prognostiskas u.c.), kas spētu noteikt piederību klasei. Izstrādātā metodoloģija un metodes labi darbojas ar augstas dimensionalitātes datiem, kuri satur maz ierakstu, kā, piemēram, gēnu vai proteīnu ekspresijas un citos datos. Pētījumā izstrādātās metodes ne vien uzlabo klasifikācijas precizitāti, bet arī rezultātā iegūtie klasifikācijas modeļi ir caurspīdīgi un viegli interpretējami, kas paplašina to pielietošanas sfēru, ietverot arī jomas, kurās ir nepieciešama ne vien precīza klasifikācija, bet arī paskaidrojums zināšanām, kas ietvertas tās loģikā. Metodes nav ierobežotas tikai vienai sfērai, bet ir jaunas dažādām jomām (pēc izsmeļošas literatūras un līdzīgo pētījumu analīzes autorei nav izdevies atrast līdzīgas metodes ai algoritmus, kas realizē izstrādātajām metodēm līdzvērtīgu funkcionalitāti). Klašu dekompozīcijas metode var lieti noderēt arī uzdevumos, kuros jāsaskaras ar sarežģītiem datiem un pastāv pamatotas aizdomas, ka arī klašu iekšējā struktūra var būt kompleksa – klases neveido radiālas formas, bet gan dažādus augsta blīvuma apgabalus, kurus var aprakstīt ar klašu dekompozīcijas palīdzību un izmantot klasifikatora veidošanā. Tādu pašu pieeju, kāda izmantota klašu dekompozīcijas metodē, var pielāgot darba ar ekspertu, izmantojot to pašu klašu struktūras aprakstu.

Aprobācija

Pētījuma rezultāti ir prezentēti sekojošās starptautiskās konferencēs:

1. *Rīgas Tehniskās universitātes 54. Starptautiskajā zinātniskajā konferencē*, Rīgā, Latvijā, 14.-16. oktobrī, 2013. gadā.
2. *European Conference on Data Analysis 2013*, Luksemburgā, Luksemburgā, 10.-12. jūlijā, 2013. gadā.

3. *Applied Information and Communication Technology 2013*, Jelgavā, Latvijā, 25.-26. aprīlī, 2013. gadā.
4. *Rīgas Tehniskās universitātes 53. Starptautiskajā zinātniskajā konferencē*, Rīgā, Latvijā, 10.-12. oktobrī, 2012. gadā.
5. *Workshop on Data Mining in Life Sciences*, Berlīnē, Vācijā, 20. jūlijā, 2012. gadā.
6. *Applied Information and Communication Technology 2013*, Jelgavā, Latvijā, 26.-27. aprīlī, 2012. gadā.
7. *21st European Meeting on Cybernetics and Systems Research*, Vīnē, Austrijā, 10.-13. aprīlī, 2012. gadā.
8. *Rīgas Tehniskās universitātes 52. Starptautiskajā zinātniskajā konferencē*, Rīgā, Latvijā, 12.-25. oktobrī, 2011. gadā.
9. *8th International and Practical Conference 'Environment. Technology. Resources'*, Rēzeknē, Latvijā, 20.-22. jūnijā, 2011. gadā.
10. *17th International Conference on Soft Computing MENDEL*, Brno, Čehijā, 15.-17. jūnijā, 2011. gadā.
11. *Rīgas Tehniskās universitātes 52. Starptautiskajā zinātniskajā konferencē*, Rīgā, Latvijā, 11.-15. oktobrī, 2010. gadā.

Publikācijas

Rezultāti publicēti sekojošos zinātniskajos rakstos:

1. Poļaka, I., Borisovs, A. The Application of Class Structure to Classification Tasks. *Informācijas tehnoloģija un vadības zinātne*. Nr.16, 2013, 114.-120. lpp. Citēts: VINITI, EBSCO, CSA/ProQuest.
2. Poļaka, I., Borisovs, A. Genethic Algorithm and Tree Based Classification in Bioinformatics. No: European Conference on Data Analysis 2013: Book of Abstracts, Luksemburga, Luksemburga, 10.-12.jūlijs, 2013. Luksemburga: 2013. 107. lpp.
3. Poļaka, I. Clustering Algorithm Specifics in Class Decomposition. No: *Applied Information and Communication Technology 2013 (AICT2013): Proceedings of the 6th International Scientific Conference*, Latvija, Jelgava, 25.-26. aprīlis, 2013. Jelgava: 2013, 29.-36.lpp. Citēts: Thomson Reuters ISI Web of Science.
4. Poļaka, I., Borisovs, A. The Impact of Cluster Stability on Class Decomposition in Antibody Display Data. *Information Technology and Management Science*. Nr.15, 2012, 70.-75.lpp. ISSN 22559086. Citēts: VINITI, EBSCO, CSA/ProQuest.
5. Poļaka, I., Borisovs, A. Class Decomposition in Bioinformatics Analyzing Omics Data. No: *Proceedings of Workshop on Data Mining in Life Sciences*

(DMLS'2012): Workshop on Data Mining in Life Sciences (DMLS'2012), Vācija, Berlin, 20.-20. jūlijs, 2012. Berlin: Springer-Verlag Berlin Heidelberg, 2012, 158.-167.lpp.

6. Poļaka I., Borisovs A. Robust Dimensionality Reduction in Bioinformatics Data // *21st European Meeting on Cybernetics and Systems Research (EMCSR 2012): Book of Abstracts*, Austria, Vīne, 10.-13. April, 2012. - 286-289. lpp.
7. Poļaka I. Genetic Algorithm for Random Tree Generation in Bioinformatics Data // *Proceedings of the 5th International Scientific Conference on Applied Information and Communication Technologies (AICT2012)*, Latvija, Jelgava, 26.-27. aprīlis, 2012. - 335.-340. lpp. Citēts: Thomson Reuters ISI Web of Science.
8. Poļaka I., Borisovs A. Impact of Antibody Panel Size on Classification Accuracy // RTU zinātniskie raksti. 5. sēr., Datorzinātne. - 49. sēj. (2011), 85.-90. lpp. Citēts: VINITI, EBSCO, CSA/ProQuest.
9. Grabusts P., Poļaka I. Estimation of the Efficiency of Knowledge Acquisition Techniques Using Clustering // Proceedings of the Ninth International Scientific School MA SR - 2011, Krievija, Sanktpēterburga, 28.jūnijs-2. jūlijs, 2011. - 131.-137. lpp.
10. Poļaka I., Borisovs A. Impact of Feature Selection on Classifier Testing Validity // Proceedings of the 17th International Conference on Soft Computing MENDEL, Čehija, Brno, 15.-17. jūnijs, 2011. - 411.-418. lpp. Citēts: Thomson Reuters ISI Web of Science.
11. Poļaka I. Feature Selection Approaches in Antibody Display Data Analysis // Proceedings of the 8th International and Practical Conference, June 20-22, 2011, Volume II, Latvija, Rēzekne, 20.-22. jūnijs, 2011. - 16.-23. lpp.
12. Poļaka I., Borisovs A. Using Data Structure Properties in Decision Tree Classifier Design // RTU zinātniskie raksti. 5. sēr., Datorzinātne. - 44. sēj. (2010), 111.-117. lpp. Citēts: VINITI, EBSCO, CSA/ProQuest
13. Poļaka I., Tom I., Borisovs A. Decision Tree Classifiers in Bioinformatics // Scientific Journal of RTU. 5. series., Datorzinātne. - 44. vol. (2010), pp 118-123. Citēts: VINITI, EBSCO, CSA/ProQuest
14. Poļaka, I., Borisovs, A. Clustering-Based Decision Tree Classifier Construction. *Technological and Economic Development of Economy*, 2010, Vol.16, Iss.4, 765.-781.lpp. Pieejams: doi:10.3846/tede.2010.47 Citēts: Taylor&Francis

Galvenie rezultāti

Pētījumā iegūtie galvenie rezultāti ir sekojoši:

- Veikta citu metožu un pieeju, kas izmantotas līdzīgos pētījumos šajā jomā, analīze.
- Izstrādāta pieeja un realizēta metode, kas ļauj aprakstīt klašu iekšējo struktūru un izmantot to klasifikatoru veidošanai.
- Izstrādāta hibrīda metode, kas meklē optimālo lēmumu koku ansambļa klasifikatoru, izmantojot ģenētiskos algoritmus.
- Izstrādāta vienota metodoloģija, kas veido diagnostiskos modeļus no datiem, izmantojot izstrādātās metodes (klašu dekompozīciju un uz ģenētiskajiem algoritmiem balstīto lēmumu koku klasifikatoru ansambļu metodi).
- Izstrādātā metodoloģija, metodes un pieejas tika novērtētas un salīdzinātas ar pieejamajām alternatīvajām metodēm, kas ļāva izdarīt secinājumus par izstrādātās metodoloģijas un metožu iespējām.

Promocijas darba struktūra un saturs

Pirmajā nodaļā ir sniegtas pētījumā risināmo uzdevumu nostādnes, kā arī problēmsfēras apraksts un darbības specifika, strādājot ar bioinformātikas datiem.

Otrajā nodaļā iekļauts līdzīgu pētījumu apskats, kas balstās uz pieejamo zinātnisko rakstu analīzi, nosakot to uzdevumus, piedāvātos risinājumus, metodes un algoritmus, kā arī rezultātus. Tajā arī sniegta informācija par populārākajām un precīzākajām metodēm, kas izmantotas šajā sfērā.

Trešajā nodaļā sniegts otrajā nodaļā noteikto populārāko metožu apskats un detalizēts apraksts, kā arī aprakstītas metodes un pieejas, kas izmantotas šajā pētījumā ar mērķi izveidot piedāvāto metodoloģiju.

Ceturtajā nodaļā aprakstīta izstrādātā metodoloģija, izskaidrotas izmantotās pieejas, kā arī sniegta detalizēta informācija par izstrādātajām metodēm un to pielietojumu.

Piektajā nodaļā aprakstīta pētījuma empīriskā daļa – eksperimentu kopas, eksperimentu plāna paskaidrojums, norādot uz saistītajām hipotēzēm, kas jāpierāda. Tāpat sniegts izstrādāto metožu un metodoloģijas eksperimentālais novērtējums, salīdzinot tās ar citām bioinformātikā populārām metodēm. Tiek sniegta arī detalizēta analīze par klasifikācijas precizitātes sakarību ar atribūtu kopas lielumu, par uzlabojumiem, kas iegūti, pielietojot klašu dekompozīciju, kā arī salīdzinājums starp piedāvāto klasifikācijas metodi, kas balstās uz ģenētiskajiem algoritmiem un lēmumu koku ansambļiem, un klasiskajām metodēm. Tāpat šajā nodaļā sniegts parametru, metožu un pieeju izvēles pamatojums.

DARBA NODAĻU SATURA APRAKSTS

1. BIOINFORMĀTIKA BIOMEDICĪNISKAJĀ DIAGNOSTIKĀ

Šajā nodaļā aprakstīts biomedicīnisko datu iegūšanas process, kā arī sniegtas pētījuma vai tā daļu uzdevumu nostādnes. Metodes, kas izvēlētas pētījuma uzdevuma risināšanai ir uz lēmumu kokiem balstītās klasifikācijas metodes (to interpretējamības un iebūvētās atribūtu apakškopas atlases dēļ), kas inducē lēmumu koku klasifikatorus, izmantojot ģenētiskos algoritmus un datu iekšējās struktūras aprakstu, lai atrastu optimālos vai kvazi-optimālos klasifikatorus. Datu iekšējā struktūra klašu dekompozīcijas metodes ietvaros tiek analizēta ar mērķi atrast augsta blīvuma apgabalus, kas aprakstītu vienas slimības atšķirīgos apakštipus un kas var tikt izmantoti lēmumu koku klasifikatoru veidošanā. Lai aprakstītu datu iekšējo struktūru, tiek risināts klasteru analīzes uzdevums.

Pētījuma uzdevuma definīcija

Uzdevums, kas šajā pētījumā tiek risināts, izmantojot datorzinātnes metodes, ir sistēmbioloģijas (biomedicīnas) diagnostikas uzdevums. Tiek doti gēnu vai proteīnu ekspresijas dati, kā arī medicīniskā diagnoze katram ierakstam (gēnu/proteīnu ekspresiju vērtību vektors). Uzdevuma risinājums, kas tiek meklēts, ir modelis, kas apraksta gēnu/proteīnu grupas (biomarķieru paneļus), kuras norāda uz specifisku diagnozi katram gēnu/proteīnu vektoram, kas sakrīt ar ‘zelta standarta’ metodi (simptomātiskās diagnostikas metode, kas nav balstīta uz gēniem/proteīniem).

Tāpēc diagnostikas uzdevuma mērķis ir atrast datos zināšanas par biomarķieriem, kas ir atšķirīgo slimību un diagnožu pamatā. Šīs zināšanas šajā pētījumā tiek atklātas, izmantojot datu ieguves un mašīnapmācības pieeju, jo tai ir zemas prasības pret datiem, kā arī tās ir viegli adaptējamas bioinformātikas uzdevuma specifikai.

Katru datu ieguves uzdevumu definē primitīvi [22], kas apraksta uzdevumu. Šī pētījuma uzdevumu definē sekojoši primitīvi:

- Uzdevuma dati: vairāk nekā tūkstotis datu atribūtu ar nepārtraukta tipa skaitļu vērtībām (gēnu vai proteīnu līmeņi) un diagnoze kā mērķa atribūts (slimība vai vesela donora iezīme), kā arī ieraksti, kas atbilst pacientu testiem (viena pacienta viens tests ir visu atribūtu vērtību vektors ar mērķa atribūta iezīmi).
- Meklētās zināšanas: klasifikācijas modelis, kas sasaista atribūtu vērtības un to sakarības ar mērķa klasi.
- Saistītās zināšanas: dati normalizēti, izmantojot fona un trokšņa līmeņus, lai izlīdzinātu signālu stiprumu dažādos testos, bet nepastāv papildus zināšanas par atribūtiem.

- Atklāto zināšanu un profilu novērtēšana: atklātie profili un klasifikācijas (diagnostikas) modeļi tiek novērtēti, izmantojot to klasifikācijas precizitāti; tāpat rezultātiem jābūt viegli interpretējamiem un jāveido biomarķieru panelis.
- Atklāto modeļu vizualizācija: pielietotā klasifikācijas pieeja ir balstīta uz lēmumu koku klasifikatoriem, jo tie ir precīzi un viegli interpretējami. Tāpēc vizualizācija ir koka grafs, kurā virsotnes apzīmē atribūtus (biomarķierus – gēnus vai proteīnus), loki ir šķelšanas vērtības, bet lapas – mērķa klases.

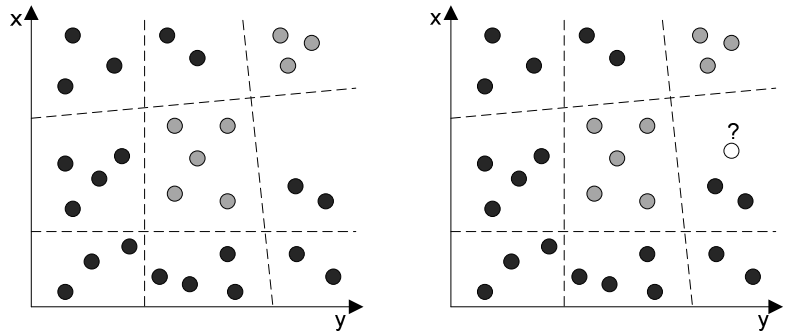
Klasifikācijas uzdevuma formālā nostādne

Tā kā datos ir iekļauti gēnu vai proteīnu ekspresijas līmeņi un mērķa klases iezīmes, diagnostiskā modeļa indukcija tiek veikta, izmantojot apmācību ar skolotāju. Tas nozīmē, ka risināmais uzdevums ir klasifikācijas uzdevums.

Klasifikācijas uzdevumā ir dota datu kopa ar ierakstiem $x_i = x_i^{A1}, x_i^{A2} \ldots, x_i^{Ai} \ldots, x_i^{An}, x_i^C$, kas ir atribūtu *A1, A2, …, An* (gēnu vai proteīnu ekspresijas līmeņu) vērtību vektori un mērķa atribūta *C* vērtība. Risinājums ir klasifikācijas modelis, kas saista mērķa atribūta vērtību ar citu atribūtu vektoru: $x_i^C = f(x_i^{A1}, x_i^{A2} \ldots, x_i^{Ai} \ldots, x_i^{An})$. Uzdevuma mērķis ir piešķirt mērķa atribūta vērtību iepriekš „neredzētam" atribūtu vērtību vektoram.

Apmācības jeb klasifikatora konstruēšanas fāzē algoritms izmanto datu kopu X, kurā *A1 … An* vērtības un mērķa atribūta *C* vērtības ir dotas. Tad algoritms meklē sakarības, kas vektoru $x_i = x_i^{A1}, x_i^{A2} \ldots, x_i^{Ai} \ldots, x_i^{An}$ atēlo mērķa atribūta vērtību kopā S_C. Šis attēlojums jeb funkcija ir rezultējošais klasifikācijas modelis, ko arī sauc par klasifikatoru. Kad visi apmācības datu kopas atribūtu vērtību vektori (ieraksti) ir izmantoti, lai atrastu funkcijas vai likumus, kas attēlo atribūtu vērtību vektorus klašu kopā, inducētais klasifikators tiek novērtēts. Tam izmanto iepriekš "neredzētu" testa datu kopu, kas satur vektorus $x_j = x_j^{A1}, x_j^{A2} \ldots, x_j^{Ai} \ldots, x_j^{An}$, kuri jāattēlo mērķa atribūta C vērtību kopā, izmantojot iepriekš inducēto klasifikatoru.

Klasifikācijas uzdevuma ģeometriskā interpretācija divdimensionālā telpā ir parādīta 3. attēlā. Dažādi datu punkti (vektori jeb ieraksti) ir attēloti divdimensionālā telpā atbilstoši diviem atribūtiem (*x* un *y*). Raustītās līnijas norāda uz klasifikatoru hiperplaknēm, kas nošķir atšķirīgas klases (parādītas kā zili un sarkani punkti). Labajā pusē ir redzams jauns vai testa ieraksts (baltais punkts), kurš jāklasificē atbilstoši esošajam klasifikatoram (hiperplaknēm). Tā kā neklasificētais punkts pieder laukam, kurā dominē tumšā klase, jaunajam ierakstam arī tiek piešķirta tumšā klase.



3. attēls. Klasifikatora vizualizācija un klases piešķiršana jaunam ierakstam

2. BIOMEDICĪNISKĀS DIAGNOSTIKAS UZDEVUMA RISINĀJUMI AR BIOINFORMĀTIKAS METODĒM

Šajā nodaļā aprakstīti citi pētījumi, kas saistīti ar bioinformātikas metožu izmantošanu biomedicīniskās diagnostikas uzdevumā. Sākotnēji gandrīz visi biomedicīniskie dati tika analizēti tikai ar statistikas metodēm, bet kopš Goluba pētījuma 1999. gadā [16] datu ieguves un mašīnapmācības metodes kļūst arvien populārākas datu analīzē ar mērķi atklāt jaunas zināšanas, kā arī diagnostiskos un prognostiskos modeļus [62].

Klasifikācija

Golubs ar kolēģiem [16] izmantoja klasifikācijas un klasteru analīzes pieejas gēnu ekspresiju datos, lai atklātu sakarības un zināšanas par leikēmiju. Šis bija biomedicīnisko datu analīzes lūzuma punkts, raksts ticis citēts vairāk nekā 800 reizes dažādos biomedicīniskajos un bioinformātikas rakstos IEEE, ACM un citos žurnālos.

Kopš tā laika daudzas dažādas mašīnapmācības metodes ir izmantotas gēnu un proteīnu ekspresijas datos. Daudzi pētnieki dod priekšroku metodēm, kas balstītas uz lēmumu kokiem, ieskaitot to ansambļu metodes, to precizitātes un interpretējamības dēļ [13, 29, 30, 31, 42, 43].

Taču vislabāko precizitāti visbiežāk sasniedz atbalsta vektoru mašīnas (SVM) un klasifikatori, kas balstīti uz naivo Baijesa klasifikatoru, taču tās nesniedz pilnīgu informāciju par biomarķieriem un to sakarībām datos [42, 43, 74].

Daudzi no šiem pētījumiem iekļauj arī atribūtu atlases uzdevumu, kas uzlabo klasifikācijas precizitāti, taču arī liek klasifikatoriem zaudēt to saturēto informāciju un klasifikatoru caurspīdīgumu [13, 41, 42, 74].

Klasteru analīze

Golubs ar kolēģiem [16] arī norādīja, ka pastāv morfoloģiski līdzīgas slimības ar atšķirīgām patoģenēzēm (vienu slimību izraisa atšķirīgi mehānismi), kas arī iedvesmoja šajā pētījumā izstrādāto klašu dekompozīcijas pieeju. Atšķirīgie slimību apakštipi simptomātiski izpaužas kā viena slimība, taču tiem ir atšķirīga norise un

atbildes reakcija uz ārstēšanu. Tas vēlreiz pierāda klašu struktūras iekšējās analīzes nepieciešamību.

Klasteru analīze visbiežāk bioinformātikā izmantota klašu atklāšanai, kas ir klasifikācijai līdzīgs uzdevums, pievēršoties problēmai no citas puses, neizmantojot informāciju par zināmajām klasēm. Visbiežāk šo pieeju izmanto datos, kas satur informāciju par atšķirīgiem terapijas iznākumiem – tiek meklētas pacientu apakšgrupas, kas izskaidrotu dažādos terapijas iznākumus. Populārākā klasteru analīzes metode bioinformātikā ir hierarhiskā klasterizācija [2, 47, 66, 73], kura bieži tiek izmantota arī atribūtu klasteru analīzei, lai noteiktu sakarības starp atribūtiem.

Ģenētisko algoritmu un uz lēmumu kokiem balstīto klasifikatoru hibrīdmetodes

Pastāv pārsteidzoši maz pētījumu par lēmumu koku klasifikatoru un ģenētisko algoritmu hibrīdu metodēm. Populārākais algoritms un rīks GATree, kurš tika izstrādāts 2010. gadā [34, 48] izmanto stipri adaptētu ģenētisko algoritmu ar savdabīgu koku reprezentāciju un līdz ar to stipri pielāgotiem operatoriem. Tāpat algoritmā pārmeklēšanai izmantota visa klasifikatoru telpa, kas ir pārāk darbietilpīgi bioinformātikas datiem ar to milzīgo dimensionalitāti. Arī citi izstrādātie algoritmi [1, 3, 15, 19] izmanto sarežģītu koku kodēšanu un pilnas klasifikatoru telpas pārmeklēšanu.

3. BIOINFORMĀTIKĀ IZMANTOTĀS MAŠĪNAPMĀCĪBAS METODES

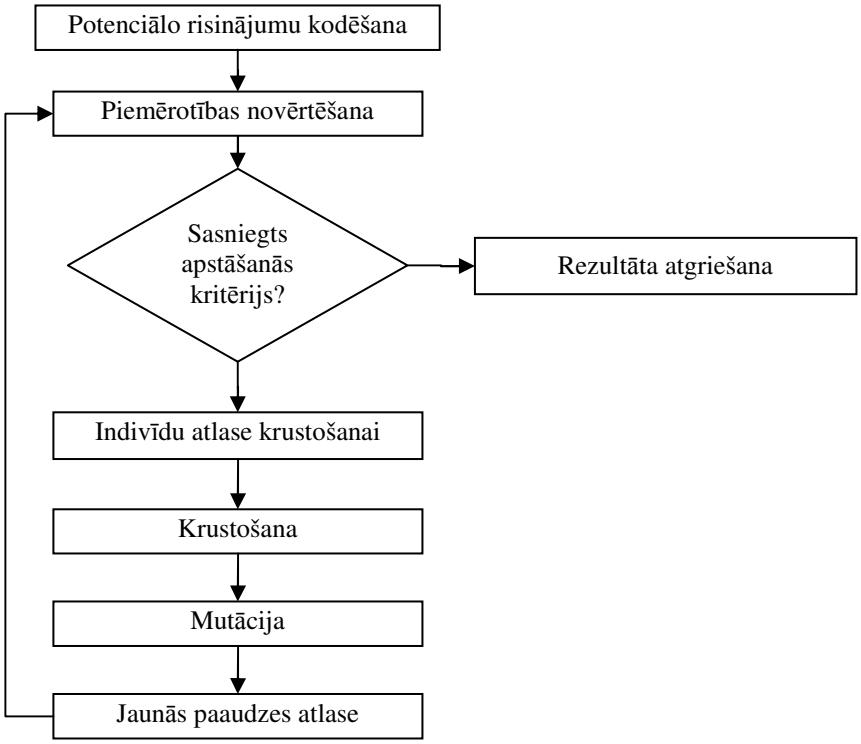
Trešajā nodaļā aprakstītas klasiskās datu ieguves un mašīnapmācības metodes, kas izmantotas pētījumā – gan metožu izstrādē, gan salīdzinošajai analīzei. Pirmajā apakšnodaļā aprakstīts pirmapstrādes posms, otrajā dots detalizēts apraksts „dimensionalitātes lāsta" pārvarēšanā [55, 57, 59, 60]. Trešajā apakšnodaļā aprakstītas bioinformātikā populārākās klasifikācijas metodes, bet ceturtajā – klasterizācijas metodes. Piektajā apakšnodaļā sniegta informācija par ģenētiskajiem algoritmiem.

Izstrādātā klašu dekompozīcijas metode pielieto klasteru analīzi [58, 61, 62, 63], tāpēc šajā nodaļā aprakstītas bioinformātikā populārākās klasterizācijas metodes (skatīt literatūras analīzi otrajā nodaļā):

- K vidējo klasterizācija [70], kas veido klasterus, balstoties uz to centriem un attālumiem starp tiem un objektiem, pievienojot objektu tuvākajam klasterim;
- Hierarhiskā klasterizācija [11], kas īsteno objektu apvienošanu (vai klasteru dalīšanu) soli pa solim, balstoties uz tuvākajiem (vai tālākajiem) objektiem.

Attālums starp objektiem, kas izmantots salīdzinošajā analīzē un klasteru analīzē, kas ir daļa no klašu dekompozīcijas, ir Eiklīda attālums [17]. Metrika, kas izmantota, lai aprēķinātu attālumus starp klasteriem ir Varda attālums [72].

Izstrādātā klasifikācijas metode izmanto lēmumu koku indukcijas algoritmu, kas balstīts uz C4.5 algoritmu [64] un *Information Gain* metriku, lai noteiktu šķelšanas atribūtu un šķelšanas vērtību. Bet koka struktūru ierobežo nosacījums par bināro dalīšanu, kā arī ierobežotais koku dziļums.
Izstrādātā klasifikācijas metode izmanto ģenētisko algoritmu [26], kura darbība parādīta 4. attēlā.



4. attēls. Ģenētisko algoritmu darbības shēma

Izstrādāto metožu un metodoloģijas salīdzinošā analīze ir veikta, izmantojot četras populāras un precīzas klasiskās klasifikācijas metodes (balstoties uz esošās situācijas analīzi, kas veikta otrajā nodaļā):

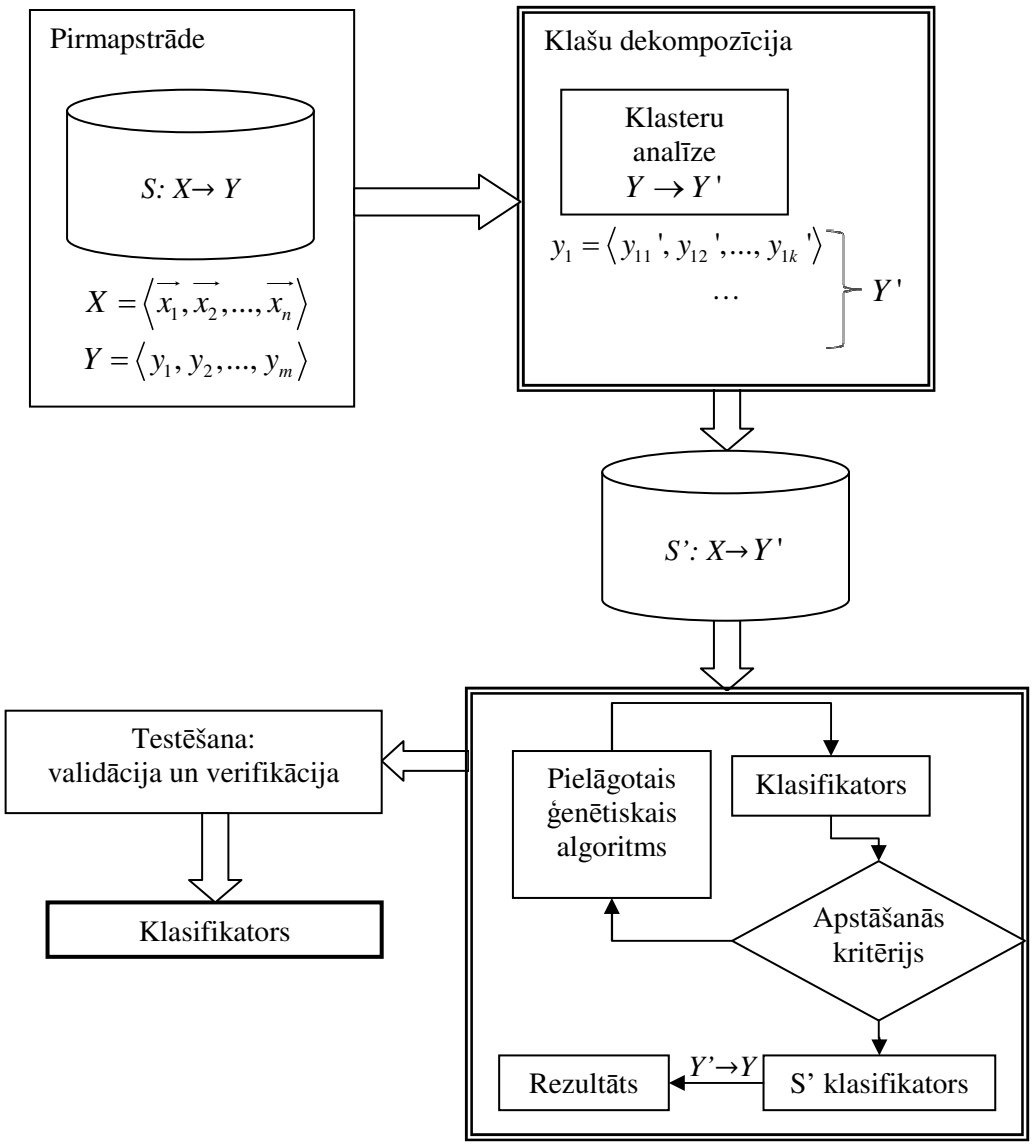
- Naivā Baijesa (NB) metode [32], kas izmanto uz varbūtībām balstītus klasifikācijas modeļus un tajos pielieto visus atribūtus bez to atlases;
- Atbalsta vektoru mašīnas (SVM) [6], kas ir populārākā un bieži arī precīzākā metode, lai gan tās inducētie modeļi ir sarežģīti un tos ir gandrīz neiespējami interpretēt medicīniskajam personālam, jo sevišķi tik augstā dimensionalitātē;
- C4.5 [64] ir populārākā lēmumu koku klasifikācijas metode; tajā ir iestrādāta atribūtu atlases metode un tās inducētie klasifikatori ir viegli interpretējami un saprotami arī speciālistiem no citām darbības jomām (finanses, medicīna u.c.);
- *Random Forest* [7] ir lēmumu koku ansambļu metode, kas izmanto gadījuma apakštelpas metodi (līdzīgi šajā pētījumā izstrādātajai metodei) un bioinformātikas pētījumos uzrāda labu precizitāti.

4. UZ MAŠĪNAPMĀCĪBU BALSTĪTAS METODOLOĢIJAS IZSTRĀDE BIOMEDICĪNISKU DIAGNOSTISKO MODEĻU INDUKCIJAI

Metodoloģija inducē diagnostiskos modeļus, izmantojot mašīnapmācības metodes. Tāpēc metodoloģijas svarīgākie soļi ir sekojoši:

- Datu sagatavošana un pirmapstrāde;
- Klašu dekompozīcija [54];
- Klasifikatora indukcija (precīzāko uz lēmumu kokiem balstīto klasifikatoru atrašana, izmantojot ģenētisko algoritmu) [51];
- Klasifikatoru testēšana un precizitātes novērtēšana [50];
- Rezultātu interpretācija.

Metodoloģijas procesi un to soļi ir parādīti 5. attēlā. Darbā izstrādātās metodes ir apvilktas ar dubultu līniju.



5. attēls. Izstrādātā metodoloģija

Datu sagatavošana un pirmapstrāde

Dati, kas iegūti, izmantojot gēnu ekspresijas mikromasīvus vai fāgu displejus, tiek sākotnēji apstrādāti un ieskenēti kā dažādas intensitātes punkti. Tiek veikta normalizācija ar skenera programmatūru, lai izlīdzinātu mikromasīvu punktu intensitātes pret fonu viena masīva ietvaros, kā arī tie tiek izlīdzināti starp atšķirīgiem skenējumiem. Ieskenētie attēli tiek pārvērsti par datu tabulām ar nepārtrauktu datu tipu vērtībām katram no punktiem, kur katrs punkts pārstāv atšķirīgu gēnu vai antivielu.

Tad iegūtie dati tiek transponēti, atbilstoši gēnu vai proteīnu kartējumam tabulās, kas satur datus par pacientiem. Šīs datu kopas, atbilstoši kopīgam gēnu vai antivielu kartējumam, tiek apvienotas ar salīdzinošajām pacientu grupām (piemēram, veselajiem donoriem). Latvijas Biomedicīnas pētījumu un studiju centra sniegtajos datos iekļauto antivielu nosaukumi tiek aizvietoti ar identifikatoriem, jo šī informācija ir patentjutīga.

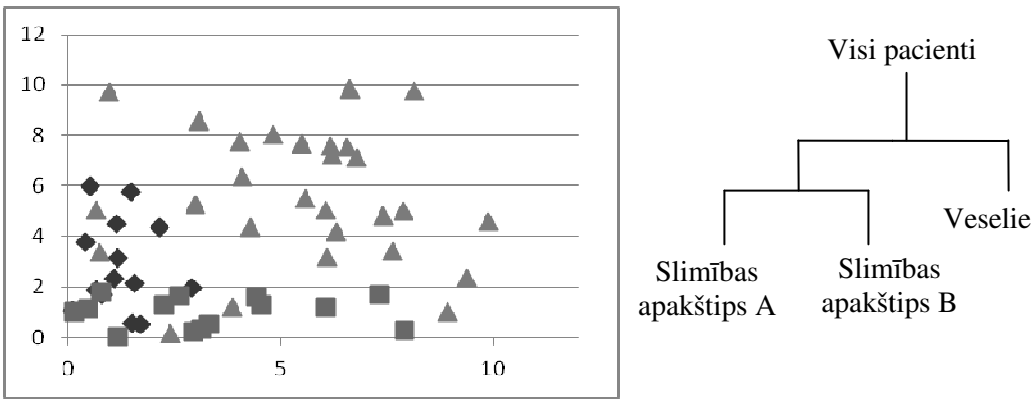
Datu kopu trūkstošās vērtības tiek aizvietotas, aprēķinot divu tuvāko „kaimiņu" vērtības, atbilstoši pirmās pakāpes Minkovski attālumu. Atribūta A_n trūkstošā vērtība vektorā x_i, kura tuvākie kaimiņi ir x_a un x_b, tiek aprēķināta pēc sekojošas formulas:

$$A_n^{x_i} = \frac{A_n^{x_a} + A_n^{x_b}}{2} \quad (1)$$

Klašu dekompozīcija

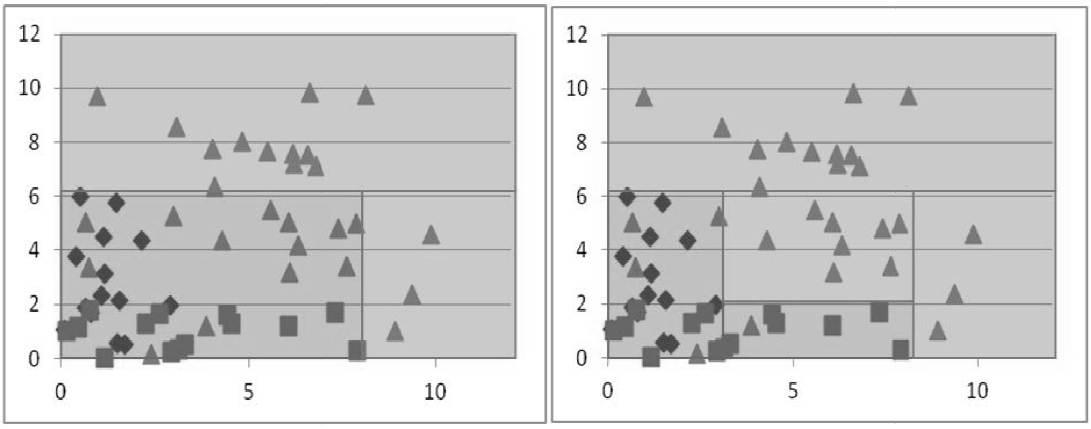
Vienai slimībai var būt apakštipi ar atšķirīgiem gēnu vai antivielu profiliem, kas norāda uz šo slimību. Tāpēc tiek analizēta pozitīvo klašu iekšējā struktūra ar mērķi atrast tajā augsta blīvuma apgabalus (izmantojot klasteru analīzi), kas var tikt izmantoti par sākotnējo klašu apakšklasēm. Darba uzdevums nav atrast īstus pārbaudītus slimību apakštipus, bet gan atvieglot klasifikācijas uzdevuma risināšanas procesu un tādā veidā uzlabot klasifikatoru precizitāti, izmantojot datos dabiski pastāvošas objektu grupas. Slimību apakštipu atrašanas process, izmantojot klasteru analīzi, šajā darbā tiek saukts par klašu dekompozīciju.

6. attēlā attēlota hipotētiska situācija, kurā pastāv pozitīvā klase (rombi un kvadrāti), kuru uzdevuma atvieglošanai var sadalīt divās apakšklasēs, un negatīvā klase (trijstūri), kuras pastāv divu dimensiju telpā (*x* un *y* ass).



6. attēls. Klasterizācijas rezultāts divu atribūtu telpā un atbilstošā dendrogramma

Situācija, kad tiek veidots viens vienkāršs klasifikators divām klasēm (pozitīvajai un negatīvajai) parādīta 7. attēla kreisajā pusē, bet klasifikators, kas izveidots, ņemot vērā pozitīvās klases apakšklases, ir parādīts labajā pusē. Otrajā grafikā ir mazāk kļūdaino pozitīvo gadījumu un līdz ar to – labāka klasifikācijas precizitāte.



7. attēls. Klasifikatora hiperplaknes, kas norāda uz klasēm:
divu klašu gadījums kreisajā pusē un klašu dekompozīcijas gadījums labajā pusē

Tā kā klašu dekompozīcijā tiek izmantoti dabiski datos pastāvoši pozitīvās klases augsta blīvuma apgabali, tie tiek noteikti, izmantojot klasteru analīzi. Klašu dekompozīcijas soļi ir sekojoši:

1. Datu kopas sagatavošana;
2. Datu sadalīšana atbilstoši klasēm;
3. Pozitīvās klases apakškopas klasterizācija, nosakot augsta blīvuma apgabalus, kas norāda biomedicīniskās slimības apakštipus, kuru profili atšķiras vienas klases ietvaros;
4. Iezīmju piešķiršana augsta blīvuma apgabaliem, kuras varēs tālāk izmantot klasifikācijas procesā;
5. Datu kopas apvienošana;

6. Atšķirīgo klašu apakštipu izmantošana tālākajā datu analīzē (klasifikācijā, prognostikā utt.) un to interpretācija, atbilstoši sākotnējām klasēm, analīzes procesa beigās.

Klasteru analīze tiek veikta, izmantojot hierarhisko aglomeratīvo klasterizāciju. Citu pētījumu analīze ļāva noteikt, ka k vidējo klasterizācija un hierarhiskā klasterizācija ir populārākās un precīzākās metodes bioinformātikā, taču metožu analīze darba ietvaros norādīja, ka k vidējo algoritmam piemīt īpašība noteikt dažus stipri atšķirīgus punktus, kuri tika apvienoti atsevišķos klasteros [52]. Hierarhiskās klasterizācijas metode savukārt atrada lielākas objektu grupas, kas var tikt interpretētas kā apakšklases.

Datu kopa, kas sagatavota pirmapstrādes procesā, satur $n = n_a + n_b$ antivielu/gēnu testu rezultātus, kur n_a ir pacientu testi, bet n_b ir veselo donoru testi. Viena testa rezultātus attēlo vektors $\vec{x} = \{x_1, x_2, \dots, x_m\}$, kas satur *m* gēnu vai antivielu rādījumus. Datu kopa n_a tiek sadalīta objektu grupas, kas atbilst augsta blīvuma apgabaliem, kas noteikti klasteru analīzes procesā. Klasterizācijas rezultātā katrs n_a saturētais vektors $\vec{x}$ pieder vienai apakšklasei C_i tā, ka vienas grupas objekti savā starpā ir līdzīgāki nekā atšķirīgām grupām piederoši objekti. Šajā gadījumā izmantotais īdzīguma mērs ir Eiklīda attālums. Atšķirība starp objektu grupām tiek noteikta atbilstoši Varda metodei [56].

Lai noteiktu klasteru skaitu, tiek izmantoti attālumi starp objektiem un klasteriem. Skaits tiek izvēlēts tā, lai klasteru kopai *S* ar *m* klasteriem C_{m}, kas satur x_{n} ierakstus, izpildās sekojoši nosacījumi:

$$\max_{C_i, C_j \in S} (d(C_i, C_j)) \tag{2}$$

$$min_{x_a, x_b \in C_i} (d(x_a, x_b)) \tag{3}$$

Noteikto klasteru kvalitāte darba ietvaros tika novērtēta, izmantojot *Gap statistic* mēru [55] un klasteru stabilitāti (robustumu) 20 iterācijās [53].

Klasifikācijas metode

Tā kā pētījumā izmantotajiem datiem ir augsta dimensionalitāte, datu analīzē izmantojamajai ir jābūt viegli mērogojamai. Tai ir jābūt arī gana caurspīdīgai, lai rezultāti būtu viegli interpretējami un izmantojami medicīnā.

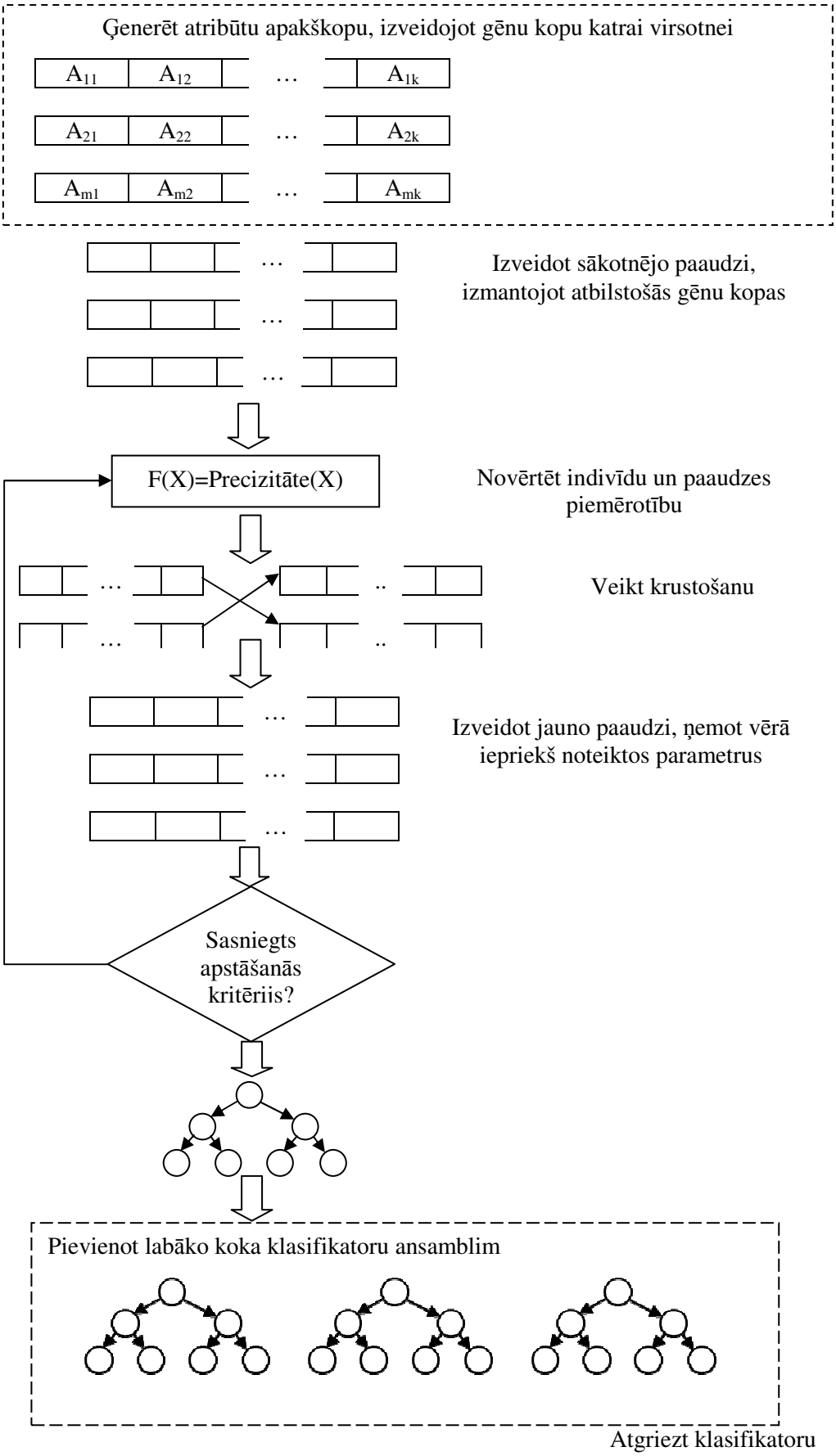
Uz lēmumu koku klasifikatoriem balstītās metodes ir mērogojamas un uzrāda augstu precizitāti (skatīt iepriekšējās nodaļas), kā arī tās veido klasifikācijas modeļus, kurus ir viegli attēlot un interpretēt, nosakot svarīgu biomarķieru paneļus un attiecības starp biomarķieriem. Tāpat lēmumu koku klasifikatori ir robusti pret troksni atribūtu vērtību skalām, kā arī citiem tos aprakstošiem parametriem.

Izstrādātā metodoloģija izmanto gadījuma apakštelpas metodi (Random subspace method), kas arī izmantota Random Forests metodē, taču tās implementācija ir

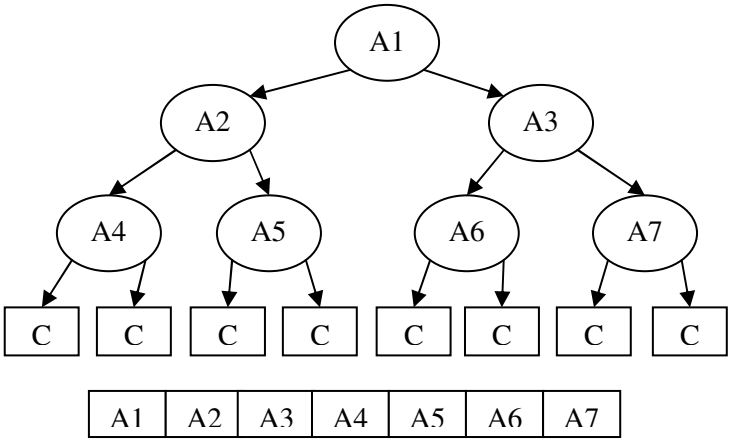
atšķirīga. Šī metode ir modificēta konkrētajam uzdevumam, veidojot gēnu kopu katram hromosomas gēnam ģenētiskajā algoritmā, kas izmantots lēmumu koku klasifikatoru ansambļa klasifikatora indukcijai, samazinot pārmeklējamo risinājumu kopu [56]. Izmantojot atšķirīgas gēnu kopas atšķirīgiem klasifikatoriem, ir iespējams izmantot informatīvākos atribūtus, nepakļaujot klasifikatorus pārapmācībai. Cits mehānisms, kas iebūvēts izstrādātajā metodoloģijā, ir atsevišķo ansamblī ietilpstošo lēmumu koku klasifikatoru lieluma ierobežošana (nosakot dziļuma ierobežojumus). Šis mehānisms ir balstīts uz Okama asmens principa – vienkāršākais (mazākais) klasifikators ir visiespējamākais patiesais datu atspoguļojums.

Izstrādātā klasifikācijas metode, kas balstīta uz ģenētiskā algoritma un lēmumu koku klasifikatoru ansambļiem ir parādīta 8. attēlā.

Izstrādātajā metodoloģijā ģenētiskais algoritms tiek izmantots, lai atrastu labākos (precīzākos) lēmumu koku klasifikatoru ansambļus. Katrs klasifikators tiek iekodēts hromosomā, izmantojot katras virsotnes atribūtus un tad nosakot šķelšanas vērtību, balstoties uz entropijas mēru (skatīt 9. attēlu). Lēmumu koka klasifikatora dziļums ir parametrs, kuru lietotājs uzstāda pirms klasifikācijas procesa. Tāpat arī lēmumu koku klasifikatoru skaits ansamblī un ģenētiskā algoritma parametri, kā paaudzes lielums, mutācijas un krustošanas varbūtības, ir iepriekš nosakāmi parametri.



8. attēls. Izstrādātā klasifikācijas metode



9. attēls. Lēmumu koka klasifikatora kodēšana

Lēmumu koku klasifikatoru ansambļi tiek izmantoti, jo tiem piemīt spēja attēlot sarežģītus datus, taču kodēt visu klasifikatoru ansambli un ar to darboties ir sarežģīti un resursietilpīgi milzīgās pārmeklējamās telpas dēļ. Tāpēc ansambļi tiek veidoti no atsevišķiem precīziem klasifikatoriem, kas veidoti atšķirīgās gadījuma veida apakštelpās, tāpēc tie papildina citos kokos ietvertās zināšanas. Kad jānosaka jauna ieraksta klase, katrs koks „balso" un balsij tiek piešķirts svars, kas atbilst atsevišķā koka klasifikatora precizitātei apmācības kopā.

Ģenētisko algoritmu pozitīvā iezīme, kas ir vienlaikus arī to negatīvā īpašība, ir tā, ka tie izmanto gadījuma veida izmaiņas klasifikatorā, padarot to sniegumu nestabilu, palaižot to vairākkārt. Lai novērtētu metodoloģiju un izstrādāto algoritmu, tika veikti 100 piegājieni ar katru datu kopu, nosakot optimālākos parametrus.

Arī rezultātu interpretācija metodoloģijai ir atšķirīga. Tā kā slimību apakštipi jāklasificē kā pozitīvā klase, klasiskā neskaidrības matrica jāpārveido. Neskaidrības matrica trīs pozitīvām apakšklasēm (+1, +2, +3) un negatīvai klasei parādīta 2. tabulā.

2. tabula

Neskaidrības matrica klašu dekompozīcijas gadījumā

		Piešķirtā klase			
		+1	+2	+3	-
Patiesā klase	+1	$\bar{I}P\|\bar{I}+1$*	$\bar{I}P\|K+2$**	$\bar{I}P\|K+3$	KN***
	+2	$\bar{I}P\|K+1$	$\bar{I}P\|\bar{I}+2$	$\bar{I}P\|K+3$	KN
	+3	$\bar{I}P\|K+1$	$\bar{I}P\|K+2$	$\bar{I}P\|\bar{I}+3$	KN
	-	KP#	KP	KP	$\bar{I}N$##

*Īstie pozitīvie, īstie +1; **Īstie pozitīvie, kļūdainie +2;
***Kļūdainie negatīvie; #Kļūdainie pozitīvie; ##Īstie negatīvie

Šajā gadījumā, lai aprēķinātu kopējo klasifikatora precizitāti, ir jāsaskaita:

$$\frac{\bar{I}P|\bar{I}^{+1} + \bar{I}P|K^{+1} + \bar{I}P|\bar{I}^{+2} + \bar{I}P|K^{+2} + \bar{I}P|\bar{I}^{+3} + \bar{I}P|K^{+3} + \bar{I}N}{\bar{I}P|\bar{I}^{+1} + \bar{I}P|K^{+1} + \bar{I}P|\bar{I}^{+2} + \bar{I}P|K^{+2} + \bar{I}P|\bar{I}^{+3} + \bar{I}P|K^{+3} + \bar{I}N + KP + KN}$$

Klasifikatora jutīgums jeb tas, cik bieži klasifikators atpazīst slimos pacientus:

$$\frac{\bar{I}P|\bar{I}^{+1} + \bar{I}P|K^{+1} + \bar{I}P|\bar{I}^{+2} + \bar{I}P|K^{+2} + \bar{I}P|\bar{I}^{+3} + \bar{I}P|K^{+3}}{\bar{I}P|\bar{I}^{+1} + \bar{I}P|K^{+1} + \bar{I}P|\bar{I}^{+2} + \bar{I}P|K^{+2} + \bar{I}P|\bar{I}^{+3} + \bar{I}P|K^{+3} + KN}$$

Klasifikatora specifiskums jeb tas, cik bieži klasifikators atpazīst veselos cilvēkus:

$$\frac{\bar{I}N}{\bar{I}N + KP}$$

5. EKSPERIMENTĀLĀ ANALĪZE

Lai pārbaudītu pirmo hipotēzi, ka nelielas atribūtu apakškopas var tikt izmantotas diagnostikai tikpat efektīvi vai pat efektīvāk nekā pilnās datu kopas, populārāko klasifikācijas metožu precizitāte tika pārbaudīta pilnās un samazinātās datu kopās. Dimensionalitāte tika samazināta līdz 10, 20, 50, 100 un 200 informatīvākajiem atribūtiem, izmantojot atribūtu apakškopas atlases un ranžēšanas metodes. Atšķirība starp precizitāti pilnajā un labāko precizitāti samazinātajās datu kopās ir parādīta 3. tabulā. Klasifikācijas metožu nosaukumi saīsināti sekojoši: NB – naivais Baijesa klasifikators, SVM – atbalsta vektoru mašīnas (Sequential Minimal Optimization algoritms), C4.5 – lēmumu koku klasifikatoru indukcijas metodes C4.5 algoritma realizācija J48, RF – Random Forest. Katrai datu kopai aprēķināts arī vidējais rezultāts, kas iegūts ar visām četrām klasifikācijas metodēm.

3. tabula

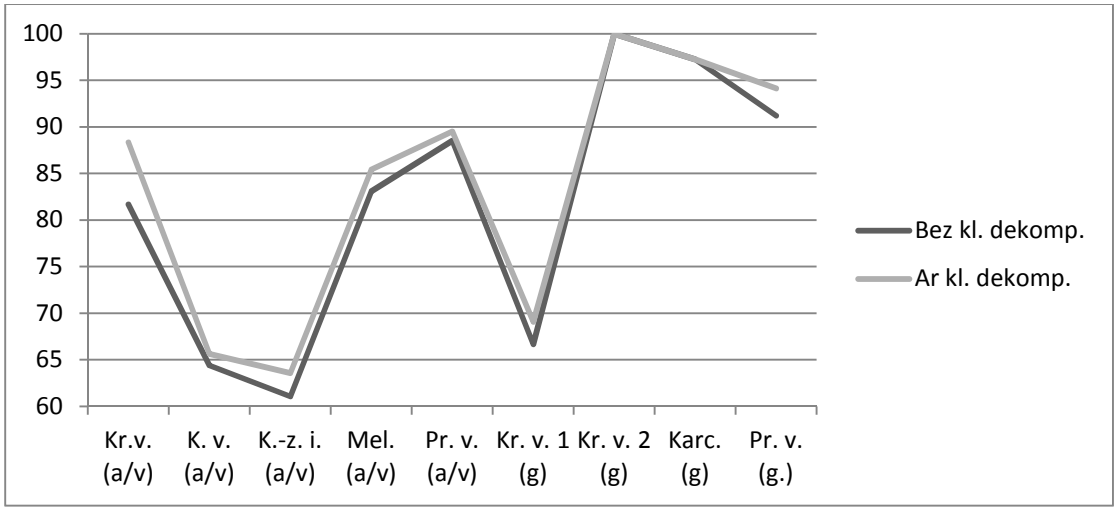
Klasifikācijas precizitātes dinamika, mainoties atribūtu kopas lielumam

Datu kopa(-s)	Atribūtu skaits samazinātajā datu kopā				
	10	20	50	100	200
Krūts vēzis (a/v)NB	**1,79**	**2,5**	**2,38**	**2,26**	**2,62**
Krūts vēzis (a/v)SVM	-0,71	0,48	0,36	**1,43**	0,83
Krūts vēzis (a/v)C45	**1,67**	**1,19**	0,83	-0,71	-1,67
Krūts vēzis (a/v)RF	-0,24	0,24	0	0,48	0,12
Krūts vēzis (a/v) vidēji	0,63	1,1	0,89	0,86	0,48
Kuņģa vēzis (a/v)NB	-4,7	-3,66	-3,66	0,98	0,12
Kuņģa vēzis (a/v)SVM	-11,65	-6,65	-6,77	-6,22	-3,48
Kuņģa vēzis (a/v)C45	-4,76	-0,3	-0,61	-2,13	-0,3
Kuņģa vēzis (a/v)RF	**3,35**	**4,88**	**5,06**	**4,27**	**6,65**
Kuņģa vēzis (a/v) vidēji	-4,44	-1,43	-1,49	-0,78	0,75
Kuņģa iekaisums (a/v) NB	0,5	**2,06**	**5,2**	**2,35**	**3,77**
Kuņģa iekaisums (a/v) SVM	-3,84	0	**2,49**	**2,92**	**2,35**
Kuņģa iekaisums (a/v) C45	0,43	**1,71**	0,64	0,14	**1,21**
Kuņģa iekaisums (a/v) RF	0,28	**2,56**	**4,56**	0,85	**6,55**
Kuņģa iekaisums (a/v) vidēji	-0,66	1,58	3,22	1,57	3,47

Datu kopa(-s)	Atribūtu skaits samazinātajā datu kopā				
	10	20	50	100	200
Melanoma NB	-7,11	-7,99	-7,52	-4,49	-1,17
Melanoma SVM	-16,03	-13,18	-6,65	-2,68	-2,62
Melanoma C45	**1,98**	**3,21**	**1,69**	**2,86**	0
Melanoma RF	-6,65	-4,55	-1,17	**1,87**	**1,92**
Melanoma vidēji	-6,95	-5,63	-3,41	-0,61	-0,47
Prostatas vēzis NB	-7,54	-2,9	-1,64	-2,42	0
Prostatas vēzis SVM	-11,4	-8,7	-5,99	-5,41	-1,84
Prostatas vēzis C45	**4,93**	**2,9**	**4,73**	**5,12**	**1,35**
Prostatas vēzis RF	-3,77	-0,58	**5,12**	**3,48**	**2,8**
Prostatas vēzis vidēji	-4,44	-2,32	0,56	0,19	0,58
Kopā vidēji	-3,17	-1,34	-0,05	0,25	0,96

Vidējie klasifikācijas rezultāti visām klasifikācijas metodēm uzrāda klasifikācijas precizitātes pieaugumu izņemot, melanomas datu kopu, kas, samazinot dimensionalitāti, zaudē daļu svarīgas informācijas. Taču arī šajā datu kopā atribūtu kopas samazināšana no 1229 atribūtiem līdz 200 rada vidēji 0,47% kritumu klasifikācijas precizitātē, kas ir salīdzinoši maz tāda mēroga samazināšanai. Lai gan dažos gadījumos precizitāte samazinās (dažas datu kopas un klasifikācijas metodes kombinācijas), lielākoties klasifikācijas kļūda nepalielinās, kas pierāda pirmo hipotēzi.

Galvenā metode, kas šajā darbā piedāvāta klasifikācijas precizitātes paaugstināšanai ir klašu dekompozīcija. Šo metodi var izmantot arī ar citām klasifikācijas metodēm, ne tikai tām, kas iekļautas izstrādātajā metodoloģijā, tāpēc tā tika pārbaudīta, izmantojot tās pašas klasifikācijas metodes, kas izmantotas salīdzinošajai analīzei. Rezultāti dažādās datu kopās ar un bez klašu dekompozīcijas parādīti 10. attēlā.



10. attēls. Labākā klasifikācijas precizitāte katrā datu kopā ar un bez klašu dekompozīcijas izmantošanas

Kopumā rezultāti uzrāda klasifikācijas precizitātes uzlabošanos visās datu kopās, izņemot iekaisuma krūts vēža datu kopu, kur abos gadījumos tiek uzrādīts 100% rezultāts, kā arī nelielajā karcinomas datu kopā, kur sniegums, izmantojot klašu dekompozīciju, nemainījās. Tas pierāda otro hipotēzi.

Tāpat klašu dekompozīcijas efektivitāte ir lineāri saistīta ar apakšklašu stabilitāti (klasteru robustumu). Apakšklašu nestabilitātes un atbilstošā maksimālā uzlabojuma klasifikācijas precizitātē rezultāti doti 4. tabulā („a/v" apzīmē antivielu datu kopas, bet „g" norāda, ka datu kopa satur datus par gēnu ekspresijām). Pīrsona korelācijas koeficients, apskatot apakšklašu nestabilitāti un maksimālo precizitātes uzlabojumu, ir -0,76 pie p vērtības $p<0,05$, kas norāda uz izteiktu lineāro korelāciju. Tas norāda, ka stabilākas apakšklases ļauj sasniegt lielāku klasifikācijas precizitātes pieaugumu.

4. tabula

Klasifikācijas precizitātes pieaugums un atbilstošā klasterizācijas nestabilitāte (vidējais pārvietoto objektu skaits)

Datu kopa	Maksimālais precizitātes pieaugums	Vidējais pārvietoto objektu skaits
Krūts vēzis (a/v)	15,39	0,00
Kuņģa vēzis (a/v)	2,50	0,04
Kuņģa iekaisuma (a/v)	13,93	0,01
Melanoma (a/v)	2,62	0,04
Prostatas vēzis (a/v)	1,00	0,33
Krūts vēzis1 (g)	4,77	0,01
Krūts vēzis2 (g)	3,12	0,02
Karcinoma (g)	8,33	0,00
Prostatas vēzis (g)	-	0,03

Lai pārbaudītu izstrādāto klasifikācijas metodi, kas izmanto ģenētisko algoritmu lēmumu koku klasifikatoru ansambļu indukcijai, sākotnējās datu kopas tika sadalītas apmācības un testa kopās. Tas veikts, lai nodrošinātu vienādas datu kopas visām metodēm atbilstoši desmitkārtīgajai šķērsvalidācijai, lai novērstu datu kopas izmaiņu ietekmi uz klasifikācijas rezultātiem. Klasifikācijas rezultāti (vidējā precizitāte visās 10 šķērsvalidācijas kārtās) ir doti 5. tabulā. Klasifikācijas metožu nosaukumi saīsināti sekojoši: GARF – izstrādātā metode, kas izmanto ģenētisko algoritmu lēmumu koku klasifikatoru ansambļu indukcijai (Genetic Algorithm generated Random Forest), GACT – GARF metodes īpašais gadījums, kad ansamblī ir tikai viens koks, NB – naivā Baijesa metode, SVM – atbalsta vektoru mašīnas, RF – Random Forest metode.

5. tabula

Izstrādātās un populārāko klasisko metožu klasifikācijas precizitātes

Datu kopa	GACT	GARF	NB	SVM	C4.5	RF
Kr.v.(g)	**78,57%**	**80,95%**	**78,57%**	69,05%	66,67%	64,29%
I.kr.v.(g)	98,89%	**100,00%**	84,44%	54,44%	72,22%	**100,00%**
Kr.v.(a/v)	**92%**	**93%**	88%	88%	**92%**	83%
Karc.(g)	80,56%	**94,44%**	86,11%	**100,00%**	86,11%	83,33%
K.v.(a/v)	59,33%	63,00%	**65,33%**	**66,00%**	59,00%	55,00%
K.-z.i. (a/v)	60,00%	**67,14%**	55,36%	59,64%	55,36%	**85,00%**
Mel.(a/v)	78,82%	**81,18%**	73,24%	79,41%	**81,18%**	**97,35%**
Pr.v.(g)	82%	**93%**	66%	**93%**	83%	82%
Pr.v.(a/v)	79%	84%	83%	**87%**	78%	**94%**

Ar treknrakstu izceltās šūnas (labākie rezultāti) norāda uz diviem labākajiem klasifikācijas rezultātiem katrā datu kopā (ja otrais un trešais rezultāts sakrīt, atzīmēti visi trīs rezultāti).

GACT metode, kurā ar ģenētisko algoritmu tiek ģenerēts viena lēmumu koka klasifikators, uzrāda labāko rezultātu tikai vienā no datu kopām, jo lielākoties datu kopas to dimensionalitātes un mazā ierakstu skaita dēļ ir pārāk sarežģītas, lai klasifikācijas modeli aprakstītu tikai ar vienu lēmumu koku ar ierobežotu līmeņu skaitu (šis parametrs tika izvēlēts no 2 līdz 10 līmeņiem). Metode, kurā ģenētiskais algoritms izmantots lēmumu koku klasifikatoru ansambļa ģenerēšanai (GARF) sasniedz vienu no labākajiem rezultātiem septiņās no deviņām datu kopām (labākais rezultāts četrās no tām). Tas parāda, ka izstrādātās klasifikācijas metodes precizitāte uzrāda līdzvērtīgi augstu rezultātu, taču saglabā klasifikatoru interpretējamu un klasifikatora veidošanas procesā atlasa svarīgāko biomarķieru paneli. Tas pierāda trešo un ceturto hipotēzi.

Izstrādātās metodoloģijas (izstrādātās GARF klasifikācijas metodes, tās īpašgadījuma GACT un klašu dekompozīcijas apvienojuma) rezultāti ir parādīti 6. tabulā. Eksperimentu plāns ir līdzīgs iepriekšējai eksperimentu sērijai, salīdzinot visas metodoloģijas precizitāti, nevis tikai klasifikācijas metodes precizitāti.

6. tabula

Izstrādātās metodoloģijas un populārāko klasisko metožu klasifikācijas precizitātes

Datu kopa	GACT	GARF	NB	SVM	C4.5	RF
Kr.v.(g)	**85,71%**	**85,71%**	78,57%	69,05%	66,67%	64,29%
I.kr.v.(g)	98,89%	**100,00%**	84,44%	54,44%	72,22%	**100,00%**
Kr.v.(a/v)	**92%**	**97%**	88%	88%	**92%**	83%
Karc.(g)	97,22%	**100,00%**	86,11%	**100,00%**	86,11%	83,33%
K.v.(a/v)	60,67%	**66,00%**	65,33%	**66,00%**	59,00%	55,00%
K.-z.i. (a/v)	61,43%	**69,29%**	55,36%	59,64%	55,36%	**85,00%**
Mel.(a/v)	81,47%	**82,65%**	73,24%	79,41%	81,18%	**97,35%**
Pr.v.(g)	87%	**94%**	66%	**93%**	83%	82%
Pr.v.(a/v)	83,00%	**90,50%**	82,50%	87,00%	78,00%	**94,00%**

Tabulā ir acīmredzami uzlabojumi pirmajās divās kolonnās, kas radušies, papildus izstrādātajai klasifikācijas metodei izmantojot metodoloģijā iekļauto klašu dekompozīciju. Arī šajā tabulā ar treknrakstu izcelti divi labākie rezultāti. Un var redzēt, ka metodoloģijai ar GARF metodes izmantošanu ir viens no diviem labākajiem rezultātiem visās datu kopās, turklāt visaugstākais rezultāts sešās no deviņām datu kopām. Neviena cita metode neuzrāda šādu dominanci, kas pierāda izstrādātās metodoloģijas efektivitāti.

REZULTĀTI UN SECINĀJUMI

Promocijas darba mērķis bija izstrādāt bioinformātikas metodoloģiju, kas izmanto datu struktūras aprakstu un ģenētiskos algoritmus klasifikācijas modeļu konstruēšanai. Mērķis tika veiksmīgi sasniegts, un tajā procesā tika veikti sekojoši soļi, sasniedzot norādītos rezultātus:

- Tika veikta līdzīgu pētījumu analītiskā izpēte, atklājot populārākās un efektīvākās metodes šajā sfērā:
    - Naivais Baijesa klasifikators, atbalsta vektoru mašīnas, C4.5 un Random Forests klasifikācijā;
    - *K* vidējo un hierarhiskā metode klasterizācijā.
- Tika izstrādāta pieeja, kas apraksta klases iekšējo struktūru un kuru var izmantot klasifikācijas procesā, bioinformātikas klasifikācijas (piemēram, diagnostikas) uzdevumā; tā tika eksperimentāli testēta, novērtējot tās parametrus un ietekmi uz klasifikācijas rezultātiem;
- Tika izstrādāta hibrīda klasifikācijas metode, kas balstās uz lēmumu koku klasifikatoru ansambļiem un ģenētiskajiem algoritmiem; tā tika eksperimentāli testēta bioinformātikas datos ar augstu dimensionalitāti un salīdzinoši mazu ierakstu skaitu;
- Tika izstrādāta vienota metodoloģija, kas pielieto izstrādātās metodes un pieejas; tā tika testēta bioinformātikas datos ar augstu dimensionalitāti un salīdzinoši zemu ierakstu skaitu;
- Tika veikta salīdzinošā analīze un izdarīti secinājumi par izstrādāto pieeju, metožu un metodoloģijas precizitāti un pielietojamību.

Visas izstrādātās metodes un metodoloģija tika eksperimentāli analizētas, lai pārbaudītu izvirzītās hipotēzes un to rezultāti ir sekojoši:

- Pirmā hipotēze tika pierādīta, samazinot atribūtu kopu un pārbaudot datos saglabāto zināšanu daudzumu, veicot klasifikatoru indukciju; klasifikatori, kas tika izveidoti, izmantojot samazinātās atribūtu kopas, bija tikpat precīzi (nav svarīgas informācijas zuduma) vai precīzāki (samazināts troksnis un atkārtošanās), kas nozīmē, ka nepieciešama tikai neliela daļa no pilnās atribūtu kopas (gēnu vai antivielu panelis), lai aprakstītu svarīgākos paraugus datos;
- Otrā hipotēze tika pierādīta, veicot klasifikatoru indukciju sākotnējā datu kopā un tad datu kopā, kurā aprakstīta klašu iekšējā struktūra; rezultāti uzrādīja precizitātes uzlabojumus septiņās datu kopās no deviņām (divās datu kopās abos gadījumos tika uzrādītas vienādas precizitātes);

- Trešā hipotēze tika pierādīta, salīdzinot klasisko klasifikācijas metožu precizitāti ar rezultātu, kuru uzrādīja izstrādātā metode, kas balstās uz lēmumu koku klasifikatoru ansambļu indukciju, izmantojot ģenētiskos algoritmus; rezultāti rāda, ka septiņos gadījumos no deviņiem izstrādātā metode sasniedza vienu no diviem augstākajiem rezultātiem, kas pierāda, ka izstrādātā metode ir līdzvērtīga precīzākajām klasiskajām metodēm, taču veido vieglāk uztveramus klasifikācijas modeļus;
- Ceturtā hipotēze tika pierādīta līdzīgā eksperimentu sērijā, un rezultāti uzrādīja, ka visos gadījumos ģenētisko algoritmu ģenerētie lēmumu koku ansambļi (10 koki) bija precīzāki par metodes īpašgadījumu, kurā ansambļi saturēja tikai vienu koku.

Izstrādātās metodoloģijas efektivitāte tika pārbaudīt salīdzinošajā analīzē, kurā tika analizēti metodoloģijas pielietojuma un klasisko klasifikācijas metožu rezultāti. No tiem secināms, ka metodoloģijas precizitāte ir augstāka par citu metožu rezultātiem sešās no deviņām datu kopām un otra labākā precizitāte pārējās datu kopās. Šis rezultāts ir vissvarīgākais, jo neviena cita metode neuzrādīja līdzvērtīgi stabilu rezultātu – citus labākos rezultātus uzrādīja klasiskās klasifikācijas metodes, kuru precizitāte citās datu kopās bija sliktāka. Kopumā par darbā veikto pētījumu ir izdarīti sekojoši secinājumi:

- Klasifikācijas metožu precizitātes uzlabojās septiņās datu kopās no deviņām, kad tika pielietota klašu dekompozīcija, kas nozīmē, ka klašu dekompozīcijas pielietošana (klašu iekšējās struktūras apraksta izmantošana klasifikācijā) uzlabo klasifikācijas precizitāti;
- Klasteru analīze augstu blīvuma apgabalu noteikšanas procesā uzrāda, ka klašu dekompozīcijai labākā klasterizācijas metode ir hierarhiskā aglomeratīvā klasterizācija;
- Tā kā Pīrsona korelācijas koeficients apakšklašu stabilitātes un klasifikācijas precizitātes korelācijai ir 0,76, lai sasniegtu maksimālo precizitātes pieaugumu, klašu dekompozīcijai ir jābūt stabilai;
- Izstrādātā klasifikācijas metode uzrāda līdzvērtīgu rezultātu precīzākajām klasiskajām klasifikācijas metodēm, saglabājot klasifikatorus vienkāršus un saprotamus, kas nozīmē, ka izstrādātā metode ir piemērotāka biomedicīnas uzdevumiem;
- Izstrādātā klasifikācijas metode labāk darbojas ar lēmumu koku klasifikatoru ansambļiem (sākot no 10 kokiem), jo metodes īpašgadījums, kurā ansamblis sastāv no viena koka uzrādīja sliktākus rezultātus nekā 10 koku ansambļi visās datu kopās;

- Izstrādātā metodoloģija, kas izmanto izstrādātās metodes, uzrādīja labāko klasifikācijas precizitāti sešās no deviņām datu kopām un otru labāko precizitāti pārējās datu kopās, kas nozīmē, ka izstrādātā metodoloģija veido precīzākus klasifikatorus nekā klasiskās klasifikācijas metodes;
- Lēmumu koku ansambļi, kurus inducē, izmantojot izstrādāto metodoloģiju, saturēja līdz desmit kokiem, kur katrs koks saturēja līdz sešiem (bet biežāk – mazāk) līmeņiem, kas nozīmē, ka izveidotie klasteri ir caurspīdīgi un viegli uztverami;
- Tā kā klasifikatoru lielums izstrādātajā metodē ir ierobežots, to indukcijas laikā tiek atlasīta informatīvāko atribūtu apakškopa (biomarķieru panelis);
- Izstrādātā metodoloģija uzrādīja vienu no diviem labākajiem rezultātiem visās izmantotajās datu kopās, kas ļauj secināt, ka izstrādātā metode ir arī stabila dažādās datu kopas, kas ir unikāls rezultāts, jo citas metodes, kas uzrādīja labus rezultātus, uzrādīja zemu precizitāti citās datu kopās.

IZMANTOTĀS LITERATŪRAS SARAKSTS

1. Aitkenhead M. J. A co-evolving decision tree classification method// Expert System Applications. – 2008. – Vol. 34, No. 1. – 18.-25. lpp.
2. Alizadeh A. A., Eisen M. B., Davis R. E., Ma C., Lossos I. S., Rosenwald A., Boldrick J. C., Sabet H., Tran T., Yu X., Powell J. I., Yang L., Marti G. E., Moore T., Hudson J. J., Lu L., Lewis D. B., Tibshirani R., Sherlock G., Chan W. C., Greiner T. C., Weisenburger D. D., Armitage J. O., Warnke R., Staudt L. M. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling// Nature. – 2000. – Vol. 403, Issue 6769. – 503.-511. lpp.
3. Barros R. C., Basgalupp M. P., de Carvalho A. C. P. L. F., Freitas A. A. A Survey of Evolutionary Algorithms for Decision Tree Induction// IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews. – 2012. – Vol. 42, Issue 3. – 291.-312. lpp.
4. Basgalupp M., de Carvalho A., Barros R. C., Ruiz D., Freitas A. Lexicographic multi-objective evolutionary induction of decision trees// International Journal of Bio-Inspired Computation. – 2009. – Vol. 1, No. 1/2. – 105.-117.
5. Bellman R. E. Adaptive Control Processes: A Guided Tour. – Prnceton, NJ: Princeton University Press, 1961. – 255 lpp.
6. Boser B. E., Guyon I. M., Vapnik V. N. A training algorithm for optimal margin classifiers// In Proceedings of the 5th Annual Workshop on Computational Learning Theory, Pittsburgh, PA, USA, July 27-29, 1992. – New York, NY: ACM Press, 1992. – 144.-152. lpp.
7. Breiman L. Bagging Predictors// Machine Learning. – 1996. – Vol. 24, No. 2. – 123.-140. lpp.
8. Breiman L., Friedman J., Olshen R., Stone C. Classification and Regression Trees. – Belmont: Wadsworth Int. Group, 1984. – 368 lpp.
9. Brown G. Encyclopedia of Machine Learning// Sammut C., Webb G.I., Eds. – Berlin, Heidelberg, Springer-Verlag, 2010. – 312.-320. lpp.
10. Büssow K., Konthur Z., Lueking A., Lehrach H., Walter G. Protein Array Technology: Potential Use in Medical Diagnostics// American Journal of PharmacoGenomics. – 2001. – Vol. 1, Issue 1. – 37.-43. lpp.
11. Carugo, O., Eisenhaber F., Eds. Data Mining Techniques for the Life Sciences (Methods in Molecular Biology). – Totowa, NJ: Humana Press, 2010. – 420 lpp.
12. Dietterich T. G. Ensemble Methods in Machine Learning// In Proceedings of the 1st International Workshop on Multiple Classifier Systems, Cagliari, Italy, June 21-23, 2000. – Lecture Notes in Computer Science. – Vol. 1857. – New York: Springer Verlag, 2000. – 1.-15. lpp.
13. Dudoit S., Fridlyand J., Speed T. P. Comparison of discrimination methods for the classification of tumors using gene expression data// Journal of the American Statistical Association. – 2002. – Vol. 97, Issue 457. – 77.–87. lpp.
14. Freund Y., Schapire R. E.. Experiments with a new boosting algorithm// In Proceedings of the 13th International Conference on Machine Learning, Bari, Italy, July 3-6, 1996. - San Francisco: Morgan Kaufmann Pub., 1996. – 148.-156. lpp.
15. Fu Z., Golden B. L., Lele S., Raghavan S., Wasil E. Diversification for better classification trees// Comput. Oper. Res. – 2006. – Vol. 33, No. 11. – 3185.-3202. lpp.
16. 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. 386. – 531.-537. lpp.

17. Gan G., Ma C., Wu J. Data clustering - theory, algorithms, and applications. – Philadelphia, USA: Society for Industrial and Applied Mathematics, 2007. – 489 lpp.

18. Guyon I., Weston J., Barnhill S., Vapnik V. Gene selection for cancer classification using support vector machines// Machine Learning. – 2002. – Vol. 46. – 389.-422. lpp.

19. Haizhou D., Chong M. Study on constructing generalized decision tree by using dna coding genetic algorithm// In Proceedings of the International Conference on Web Information Systems and Mining, Shanghai, China, November 7-8, 2009. – Washington, DC, USA: IEEE Computer Society, 2009. – 163.–167. lpp.

20. Hall D. A., Ptacek J., Snyder M. Protein Microarray Technology// Mech Ageing Dev. – 2007. – Vol. 128, Issue 1. – 161.–167. lpp.

21. Hall M. A. Correlation based Feature Subset Selection for Machine Learning. – 1998. – Disertācija Vaikato universitātē (Hamilton, Jaunzēlande) – 198 lpp.

22. Han J., Kamber M. Data Mining: Concepts and Techniques. The Morgan Kaufmann Series in Data Management Systems. – Gray J., Ed. - San Mateo: Morgan Kaufmann Pub., 2000. – 550 lpp.

23. Harrington C. A., Rosenow C., Retief J. Monitoring gene expression using DNA microarrays// Current Opinion in Microbiology. – 2000. – Vol. 3, Issue 3. – 285.-291. lpp.

24. Hinneburg E., Keim D. A. Clustering Techniques for Large Data Sets From the Past to the Future// In Proceedings of the International Conference on Knowledge Discovery and Data Mining, Sandiego, California, USA, August 15-18, 1999. – New York, NY: ACM Press, 1999. – 141.-181. lpp.

25. Ho T. K. The Random subspace method for constructing decision forests// IEEE Trans Pattern Analysis and Machine Intelligence. – 1998. – Vol. 20, Issue 8. – 832.-844. lpp.

26. Holland J. Adaptation in Natural and Artificial Systems. Ann Arbor, MI: University of Michigan Press, 1975. – 183 lpp.

27. Holte R. C. Very Simple Classification Rules Perform Well on Most Commonly Used Datasets// Machine Learning. – 1993. – Vol. 11. – 63.-91. lpp.

28. Horng J., Wu L., Liu B., Kuo J., Kuo W., Zhang J. An expert system to classify microarray gene expression data using gene selection by decision tree// Expert Systems Applications. – 2009. – Vol. 36, Issue 5. – 9072.-9081. lpp.

29. Hu H. Mining patterns in disease classification forests// Journal of Biomedical Informatics – 2010. – Vol. 43, Issue 5. – 820.-827. lpp.

30. Huang J., Fang H., Fan X. Decision forest for classification of gene expression data// Computers in Biology and Medicine. – 2010. – Vol. 40, Issue 8. – 698.-704. lpp.

31. Huber W., von Heydebreck A., Vingron M. Analysis of Microarray Gene Expression Data// In Handbook of Statistical Genetics. – Hoboken, NJ: John Wiley & Sons, 2004. – 203.-230. lpp.

32. John G. H., Langley P. Estimating Continuous Distributions in Bayesian Classifiers// In Proceedings of the Eleventh Conference on Uncertainty in Artificial Intelligence, Montreal, Quebec, Canada, August 18-20, 1995. – San Mateo: Morgan Kaufmann Pub., 1995. – 338.-345. lpp.

33. Jones B. R., Crossley W. A., Lyrintzis A. S. Aerodynamic and Aeroacoustic Optimization of Airfoils via a Parallel Genetic Algorithm// In Proceedings of the 7th Symposium on Multidisciplinary Analysis and Optimization, St. Louis, MO, USA, September 2-4, 1998. – Reston: AIAA, 1998. – 1088.-1096. lpp.

34. Kalles D, Papagelis A. Lossless fitness inheritance in genetic algorithms for decision trees// Soft Computing. – 2010. – Vol. 14. – 973.-993. lpp.

35. Kantardzic M. Data Mining: Concepts, Models, Methods, and Algorithms, Second Edition. – Hoboken, NJ: John Wiley & Sons, Inc., 2011. – 552 lpp.

36. Kohavi R., John G. H. Wrappers for feature subset selection// Artificial Intelligence. – 1997. – 1.-2. lpp.

37. 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. – 267.-276. lpp.

38. Kohavi R., Provost F. Glossary of terms// Applications of Machine Learning and the Knowledge Discovery Process. – 1998. – Vol. 30, No. 2/3. – 2.-3. lpp.

39. Kononenko, I. Estimating attributes: analysis and extensions of RELIEF// In Proceedings of the European Conference on Machine Learning on Machine Learning, Catania, Italy, April 6-8, 1994. – New York: Springer-Verlag, 1994. – 171.-182. lpp.

40. Kretowski M., Grzes M. Evolutionary induction of cost-sensitive decision trees// In Proceedings of the 16th international conference on Foundations of Intelligent Systems, Bari, Italy, September 27-29, 2006. – Berlin, Heidelberg: Springe-Verlag, 2006. – 121.-126. lpp.

41. Lee G., Rodriguez C., Madabhushi A. An Empirical Comparison of Dimensionality Reduction Methods for Classifying Gene and Protein Expression Datasets// In Proceedings of the Bioinformatics Research and Applications: Third International Symposium, Atlanta, GA, USA, May 7-10, 2007. – Berlin, Heidelberg: Springer-Verlag, 2007. – 170.-181. lpp.

42. Lee J. W., Lee J. B., Park M.. Song S. H. An extensive comparison of recent classification tools applied to microarray data// Computational Statistics & Data Analysis. – 2005. – Vol. 48, Issue 4. – 869.-885. lpp.

43. Lu Y., Han J. Cancer classification using gene expression data// Information Systems. – 2003. – Vol. 28, Issue 4. – 243.-268. lpp.

44. MacQueen J. Some methods for classification and analysis of multivariate observations// In Proceedings of the Fifth Berkeley Symp. on Math. Statist. and Prob., Berkeley, CA, USA, June 21-July 18, 1965, Vol. 1. – Berkeley, CA: University of California Press, 1967. – 281.-297.lpp.

45. Mingers J. An Empirical Comparison of Pruning Methods for Decision Tree Induction// Machine learning. – 1989. – Vol. 2, Issue 4. – 227.-243. lpp.

46. Mishra D., Shaw K., Mishra S., Rath A. K., Acharya M. Hash based biclustering for class discovery from gene expression data: A pattern similarity approach// In Proceedings of the 3rd International Conference on Electronics Computer Technology (ICECT), Kanyakumari, India, April 8-10, 2011, Vol. 2. –- Washington, DC: IEEE Computer Society, 2011. – 137.-141. lpp.

47. Monti S., Tamayo P., Mesirov J., Golub T. Consensus clustering - A resampling-based method for class discovery and visualization of gene expression microarray data// Machine Learning. – 2003. – Vol. 52, Issue 1-2. – 91.-118. lpp.

48. Papagelis A., Kalles D. GA Tree: genetically evolved decision trees// In Proceedings of the 12th IEEE International Conference on Tools with Artificial Intelligence (ICTAI '00), Vancouver, Canada, November 13-15, 2000. – Washington, DC: IEEE Computer Society, 2000. – 203.-207. lpp.

49. Poļaka, I., Borisovs, A. Genethic Algorithm and Tree Based Classification in Bioinformatics. No: *Studies in Classification, Data Analysis, and Knowledge Organization*. Heidelberg: Springer, 2014, (iesniegts).

50. Poļaka, I., Borisovs, A. The Application of Class Structure to Classification Tasks. *Informācijas tehnoloģija un vadības zinātne*. Nr.16, 2013, (apstiprināts).

51. Poļaka, I., Borisovs, A. Genethic Algorithm and Tree Based Classification in Bioinformatics. No: European Conference on Data Analysis 2013: Book of Abstracts, Luksemburga, Luksemburga, 10.-12.jūlijs, 2013. Luksemburga: 2013. 107. lpp.

52. Poļaka, I. Clustering Algorithm Specifics in Class Decomposition. No: *Applied Information and Communication Technology 2013 (AICT2013): Proceedings of the 6th International Scientific Conference*, Latvija, Jelgava, 25.-26. aprīlis, 2013. Jelgava: 2013, 29.-36.lpp.

53. Poļaka, I., Borisovs, A. The Impact of Cluster Stability on Class Decomposition in Antibody Display Data. *Information Technology and Management Science*. Nr.15, 2012, 70.-75.lpp. ISSN 22559086.

54. Poļaka, I., Borisovs, A. Class Decomposition in Bioinformatics Analyzing Omics Data. No: Proceedings of Workshop on Data Mining in Life Sciences (DMLS'2012): Workshop on Data Mining in Life Sciences (DMLS'2012), Vācija, Berlin, 20.-20. jūlijs, 2012. Berlin: Springer-Verlag Berlin Heidelberg, 2012, 158.-167.lpp.

55. Poļaka I., Borisovs A. Robust Dimensionality Reduction in Bioinformatics Data // *21st European Meeting on Cybernetics and Systems Research (EMCSR 2012): Book of Abstracts*, Austria, Vīne, 10.-13. April, 2012. - pp 286-289.

56. Poļaka I. Genetic Algorithm for Random Tree Generation in Bioinformatics Data // *Proceedings of the 5th International Scientific Conference on Applied Information and Communication Technologies (AICT2012)*, Latvija, Jelgava, 26.-27. aprīlis, 2012. - 335.-340. lpp.

57. Poļaka I., Borisovs A. Impact of Antibody Panel Size on Classification Accuracy // RTU zinātniskie raksti. 5. sēr., Datorzinātne. - 49. sēj. (2011), 85.-90. lpp.

58. Grabusts P., Poļaka I. Estimation of the Efficiency of Knowledge Acquisition Techniques Using Clustering // Proceedings of the Ninth International Scientific School MA SR - 2011, Krievija, Sanktpēterburga, 28.jūnijs-2. jūlijs, 2011. - 131.-137. lpp.

59. Poļaka I., Borisovs A. Impact of Feature Selection on Classifier Testing Validity // Proceedings of the 17th International Conference on Soft Computing MENDEL, Čehija, Brno, 15.-17. jūnijs, 2011. - 411.-418. lpp.

60. Poļaka I. Feature Selection Approaches in Antibody Display Data Analysis // Proceedings of the 8th International and Practical Conference, June 20-22, 2011, Volume II, Latvija, Rēzekne, 20.-22. jūnijs, 2011. - 16.-23. lpp.

61. Poļaka I., Borisovs A. Using Data Structure Properties in Decision Tree Classifier Design // RTU zinātniskie raksti. 5. sēr., Datorzinātne. - 44. sēj. (2010), 111.-117. lpp.

62. Poļaka I., Tom I., Borisovs A. Decision Tree Classifiers in Bioinformatics // Scientific Journal of RTU. 5. series., Datorzinātne. - 44. vol. (2010), pp 118-123.

63. Poļaka, I., Borisovs, A. Clustering-Based Decision Tree Classifier Construction. *Technological and Economic Development of Economy*, 2010, Vol.16, Iss.4, 765.-781.lpp. Pieejams: doi:10.3846/tede.2010.47

64. Quinlan J. R. C4.5: Programs for Machine Learning. – San Mateo: Morgan Kaufmann Pub., 1993. – 302 lpp.

65. Quinlan J. R. Simplifying decision trees// International Journal of Man-Machine Studies. – 1987. - Vol. 27. – 221.-248. lpp.

66. Slonim D. K., Tamayo P., Mesirov J. P., Golub T. R., Lander E. S. Class prediction and discovery using gene expression data// In Proceedings of Fourth

Annual International Conference on Computational Molecular Biology, Tokyo, Japan, April 8-11, 2000. – New York, NY: ACM, 2000. – 263.-272. lpp.

67. Sreekumar A., Nyati M. K., Varambally S., Barrette T. R., Ghosh D., Lawrence T. S., Chinnaiyan A. M. Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteīns// Cancer Res. – 2001. – Vol. 61. – 7585.-7593. lpp.

68. Steinfeld I., Navon R., Ardigò D., Zavaroni I., Yakhini Z. Clinically driven semi-supervised class discovery in gene expression data// *Bioinformatics. – 2008. – Vol.* 24, Issue 16. – i90.-i97. lpp.

69. Tan C. P., Lim K. S., Lai W. K. Multi Dimensional Features Reduction of Consistency Subset Evaluator on Unsupervised Expectation Maximization Classifier for Imaging Surveillance Application// International Journal of Image Processing. – 2008. – Vol. 2, Issue 1. – 18.-26. lpp.

70. Tan P. N., Steinbach M., Kumar V. Introduction to Data Mining. – Boston: Pearson Addison-Wesley, 2006. – 769 lpp.

71. Tibshirani R., Walther G., Hastie T. Estimating the Number of Clusters in a Dataset via the Gap Statistic// Journal of the Royal Statistical Society, Series B. – 2000. – Vol. 63. – 411.-423. lpp.

72. Ward J. H., Jr. Hierarchical Grouping to Optimize an Objective Function// Journal of the American Statistical Association. – 1963. – Vol. 58. – 236.-244. lpp.

73. Witten I. H., Frank E. Data mining: Practical machine learning tools and techniques (Second edition). – San Francisco, CA: Morgan Kaufmann, 2005. – 560 lpp.

74. Yu Z., Wong H. S. Class discovery from gene expression data based on perturbation and cluster ensemble// IEEE Trans Nanobioscience. – 2009. – Vol. 8, Issue 2. – 147.-160. lpp.

75. Zhao H. A multi-objective genetic programming approach to developing pareto optimal decision trees// Decis. Support Syst. – 2007. – Vol. 43, No. 3. – 809.-826. lpp.

76. Zintzaras E., Kowald A. Forest classification trees and forest support vector machines algorithms: Demonstration using microarray data// Comput. Biol. Med. – 2010. – Vol. 40, Issue 5. – 519.-524. lpp.