RĪGAS TEHNISKĀ UNIVERSITĀTE
Datorzinātnes un informācijas tehnoloģijas fakultāte
Informācijas tehnoloģijas institūts

Madara GASPAROVIČA-ASĪTE
Doktora studiju programmas «Informācijas tehnoloģija» doktorante

IZPLŪDUŠĀS KLASIFIKĀCIJAS METODOLOĢIJA BIOINFORMĀTIKAS DATU APSTRĀDEI UN ANALĪZEI

Promocijas darba kopsavilkums

Zinātniskā vadītāja
asoc. prof. *Dr. sc. ing.*
L. ALEKSEJEVA

RTU Izdevniecība
Rīga 2015

Gasparoviča-Asīte M. Izplūdušās klasifikācijas metodoloģija bioinformātikas datu apstrādei un analīzei. – Rīga: RTU Izdevniecība, 2015. – 36 lpp.

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

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

ISBN 978-9934-10-762-7

PROMOCIJAS DARBS
IZVIRZĪTS INŽENIERZINĀTŅU DOKTORA GRĀDA IEGŪŠANAI
RĪGAS TEHNISKAJĀ UNIVERSITĀTĒ

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

OFICIĀLIE RECENZENTI

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

Profesors *Dr. sc. ing.* Pēteris Grabusts
Rēzeknes Augstskola, Latvija

Asociētais profesors *Dr. comp. Vytautas Čyras*
Viļņas Universitāte, Lietuva

APSTIPRINĀJUMS

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

Madara Gasparoviča-Asīte …….(Paraksts)

Datums: ……

Promocijas darbs ir uzrakstīts latviešu valodā, tajā ir ievads, piecas nodaļas, rezultātu analīze un secinājumi, 31 tabula, 47 attēli, seši pielikumi, kopā 160 lappušu, t. sk. seši pielikumi. Literatūras sarakstā ir 133 vienības.

Saturs

Vispārējs darba raksturojums

2003. gadā ar cilvēka genoma atšifrēšanu tika sākta strauja informācijas meklēšana tajā. Biologi un mediķi pārsvarā lieto statistikas metodes, tomēr datu ieguves tehnoloģiju lietošana tikpat veiksmīgi vai pat veiksmīgāk var palīdzēt atklāt zināšanas liela apjoma bioinformātikas datos. Vēl 2005. gadā bioinformātika un datu ieguve saskarās tikai atsevišķu autoru pētījumos, tomēr mūsdienās arvien vairāk biologu izvēlas sadarboties ar datu ieguves speciālistiem, un tiek veidoti kopīgi pētījumi.

Problēmas nostādne

Zināšanu atklāšana bioinformātikas datos var palīdzēt atklāt būtisku informāciju par to, kas cilvēka organismā var liecināt par kādu saslimšanu. Sarežģītie bioinformātikas dati ar lielo dimensionalitāti pēdējā desmitgadē ir vākti un krāti aizvien plašāk, kā arī ir pieejami elektroniski, un radusies arī nepieciešamība tos elektroniski apstrādāt. Lielie datu apjomi cilvēkam nav pārskatāmi un analizējami, tajos nav iespējams atrast kādas likumsakarības, neizmantojot papildu analīzes metodes. Bioinformātikas speciālisti visbiežāk izmanto statistikas metodes, kas lielākoties ļauj pārbaudīt zināmas sakarības, turklāt tās ir prasīgas pret datiem (sadalījums, izkliede, ierakstu skaits). Tomēr šādas informācijas nav, jo salīdzinoši jaunajā genomikas un proteomikas jomā vēl ir daudz nezināmā.

Datu ieguve ar savām plašajām tehnoloģijām un to lietošanas iespējām ir sevi pierādījusi dažādās problēmsfērās. Tāpēc tās lietošana tieši bioinformātikā pēdējās desmitgadēs paver iespējas plašiem pētījumiem. It īpaši piedāvājot iespējas meklēt likumsakarības, ņemot vērā klases jēdzienu. Tas var palīdzēt noteikt dažādu atribūtu dažādas vērtības un to kombinācijas, kas ietekmē datu kopas klasi. Ja klases ir slimie/veselie cilvēki, tad, lietojot datu ieguves metodes, iespējams iegūt dažādas vienkāršas, cilvēkam ērti uztveramas likumsakarības (likumus), kas norāda uz dažādām datu īpašībām. Izplūdušās tehnoloģijas, kas ir vistuvākās cilvēka uztverei, paver plašākas iespējas imitēt reāla cilvēka viedokli, uzskatot, ka viena vērtība var piederēt vairākām grupām.

Tikai kopš 2000. gada veikti pētījumi par datu ieguves lietošanu bioinformātikā. Lielākā daļa no sastopamiem pētījumiem koncentrējas uz atsevišķu datu ieguves tehnoloģiju izmantošanu bioinformātikas datu apstrādē, taču vispārīgu un plašu pētījumu par izplūdušo klasifikācijas sistēmu lietošanu, ietverot datu pirmapstrādi, klasifikāciju un likumu novērtēšanu, nav izdevies atrast. Šādas metodoloģijas izstrāde paver plašas iespējas, izmantojot tās pozitīvās puses – to, ka tā īpaši paredzēta bioinformātikas daudzdimensionālajiem datiem, kā arī izplūduma iespēju, kas nodrošina piederību vairākām vērtībām vienlaikus, rezultātā iegūstot cilvēka domāšanai līdzīgā procesā iegūtus likumus.

Nepieciešams izstrādāt izplūdušās klasifikācijas metodoloģiju, kas, izmantojot dažādas datu ieguves metodes un algoritmus, veiks specifisko bioinformātikas datu pirmapstrādi un klasificēšanu, lai atklātu datos paslēptās sakarības, kas raksturo atsevišķas pacientu grupas, slimības u. c., kas jāuztver kā klases.

Mērķis un uzdevumi

Darba **mērķis** ir izplūdušās klasifikācijas metodoloģijas izstrādāšana, kas paredzēta bioinformātikas datu apstrādei un analīzei. Lai sasniegtu izvirzīto mērķi, definēti šādi **uzdevumi**.

1. Bioinformātikas datu izpētes gaitā jādefinē prasības klasifikācijas algoritmiem, kas īpaši paredzēti šādu datu apstrādei.
2. Jāveic datu pirmapstrādes metožu izpēte, lai noskaidrotu tās metodes, ko iespējams izmantot un kas ir efektīvas bioinformātikas datu pirmapstrādē.
3. Jāveic praktiski pētījumi, lai noskaidrotu algoritmus un metodes, ko jāizmanto izplūdušās klasifikācijas sistēmas izstrādē.
4. Jāizstrādā piederības funkciju konstruēšanas metode, izmantojot klasterizāciju.
5. Jāizstrādā likumu fazifikācijas metode, kas paplašinātu likumu nosacījumus.
6. Jāizstrādā izplūdusī klasifikācijas metodoloģija, ņemot vērā teorētiskajos un praktiskajos pētījumos noskaidroto.
7. Uz izveidotās metodoloģijas pamata jāizstrādā izplūdusī klasifikācijas sistēma.
8. Izplūdušās klasifikācijas metodoloģijas un sistēmas jāizmanto reāliem bioinformātikas datiem, jāizdara secinājumi par metodoloģijas un sistēmas efektivitāti.

Objekts un subjekts

Pētījuma objekts ir mašīnapmācības un datu ieguves algoritmi. Pētījuma subjekts ir bioinformātikas dati, kuru vajadzībām un apstrādei atbilstoši tiek izstrādāta izplūdusī klasifikācijas metodoloģija, izmantojot piemērotus mašīnapmācības un datu ieguves algoritmus.

Hipotēzes

Darba izstrādes gaitā tika formulētas un aizstāvēšanai izvirzītas šādas ar izstrādājamo izplūdušās klasifikācijas metodoloģiju saistītas hipotēzes.

- Piederības funkciju konstruēšana, lietojot klasterizāciju, izmantojot informāciju un zināšanas par datiem, kas uzlabo klasifikācijas precizitāti.
- Likumu stiepšanas (angļu val. *streching*) un fazifikācijas lietošana nodrošina iespēju pārklāt jaunus ierakstus, kuru vērtības ir tuvas tām, kas piedalījušās apmācībā, bet nav identiskas.

Pirmā hipotēze balstās idejā, ka klasterizācijas procesa laikā tiek meklētas likumsakarības datos un dati tiek grupēti atbilstoši šīm likumsakarībām. Izmantojot šo iegūto

informāciju (klasteru centrus), notiek piederības funkciju konstruēšana. Līdz ar to piederības funkcijas netiek konstruētas, tikai skaitliski un proporcionāli sadalot visus datus izvēlētā intervālu skaitā, bet gan tiek ņemtas klasterizācijas gaitā iegūtās zināšanas. Šī hipotēze tiks pārbaudīta, salīdzinot uz klasterizāciju balstītu piederības funkciju konstruēšanas metožu iegūtos klasifikācijas rezultātus ar matemātiski izrēķinātu piederības funkciju klasifikācijas rezultātiem. Ja uz klasterizāciju balstītu piederības funkciju konstruēšanas metožu iegūtais klasifikācijas rezultāts būs precīzāks, hipotēze tiks pierādīta kā patiesa.

Otrā hipotēze balstīta likumu fazifikācijā un likumu stiepšanas būtībā. Likuma fazifikācijas gaitā no piederības funkciju konstruēšanas darbības tiek iegūts ne tikai intervāls, kurā iegūtais likums ir spēkā, bet arī intervāls, kurā iegūtais likums ir daļēji spēkā, tādējādi paplašinot vērtību skaitu, kas pieder likumam. Ja likumu fazifikācijas ieviešana paplašinās likuma veidojošo nosacījumu vērtību skaitu un testa nolūkos radīto līdzīgo, bet ne identisko ierakstu klasifikācija, izmantojot fazificētos likumus, būs veiksmīga, hipotēze tiks pierādīta kā patiesa.

Metodes

Promocijas darbā tiek izmantota izplūdušo kopu teorija, matemātiskās un datu ieguves metodes – datu pirmapstrādes metodes (trūkstošu vērtību apstrādāšana, atribūtu atlase, piederības funkciju konstruēšana), kā arī klasifikācija un klasterizācija.

Darba aktualitāte un zinātniskā novitāte

Parasti bioinformātikas dati tiek pētīti ar statistikas metodēm, tomēr pēdējos gados arvien vairāk tiek lietotas arī datu ieguves un mašīnapmācības tehnoloģijas un algoritmi. Šo algoritmu lietojums var palīdzēt atklāt dažādas likumsakarības bioinformātikas datos, tiem ir plašas datu pirmapstrādes iespējas.

Izplūdušo algoritmu lietojums bioinformātikas datu analīzē paver vēl plašākas iespējas, jo tiek skatīta iespēja, kad konkrēts ieraksts pieder vairākām vērtībām vienlaikus, tikai – ar atbilstošu piederības līmeni. Jo medicīnas kvantitatīvie dati nav viennozīmīgi un dažādiem pacientiem (pat ar vienu diagnozi) var būt ļoti atšķirīgi izmērīto faktoru līmeņi. Šī izplūdušo algoritmu īpašība ir atbilstoša reālas pasaules spriedumiem un cilvēka uztverei un izpratnei, jo nav iespējams precīzi noteikt to, kur, piemēram, atdalās viens lingvistisks jēdziens no otra.

Šī promocijas darba zinātniskā novitāte saistīta ar šādiem aspektiem.

- Šī darba izstrādes gaitā izveidota izplūdusī klasifikācijas metodoloģija, kas nodrošina datu pirmapstrādi, datu klasifikāciju, likumu bāzes izveidi un jaunu piemēru klasifikāciju, kā arī rezultātu novērtēšanu.

- Izstrādāta piederības funkciju konstruēšanas metode, balstoties uz klasterizācijas algoritmiem.
- Pielāgota likumu stiepšanas un fazifikācijas metode.

Praktiskais nozīmīgums

Galvenais darba praktiskais nozīmīgums ir izstrādātā izplūdusī klasifikācijas metodoloģija, kas var apstrādāt, klasificēt un analizēt lielas dimensionalitātes datus (daudz atribūtu un maz ierakstu), tādus kā bioinformātikas dati. Uz metodoloģijas bāzes izveidotā izplūdusī klasifikācijas sistēma ar eksperimentāli noskaidrotajiem labākajiem katrā solī izvēlētajiem algoritmiem un metodēm uzrāda konkurētspējīgus klasifikācijas rezultātus, tai pat laikā nodrošinot labi uztveramus klasifikācijas rezultātus «Ja–tad» likumu veidā.

Darba izstrādes gaitā apkopotie teorētiskie materiāli un veiktie praktiskie eksperimenti ir izmantojami kā pētījums par bioinformātikā izmantojamām datu ieguves tehnoloģijām.

Aprobācija

Promocijas darba izstrādes gaitā veiktie pētījumi ir prezentēti 13 starptautiskās konferencēs.

1. RTU 55th International Scientific Conference, Rīga, Latvija, 14–17 Oktobris, 2014 (kopā ar L. Aleksejevu).
2. RTU 54th International Scientific Conference, Rīga, Latvija, 14–16 Oktobris, 2013 (kopā ar L. Aleksejevu).
3. 6th Conference Applied Information and Communication Technology, Jelgava, Latvija, 25–26 Aprīlis, 2013 (kopā ar L. Aleksejevu un V. Nazaruku).
4. RTU 53th International Scientific Conference, Rīga, Latvija, 10–12 Oktobris, 2012 (kopā ar L. Aleksejevu un V. Gersonu).
5. Workshop on Data Mining in Life Sciences DMLS'2012, Berlīne, Vācija, 20–22 Jūlijs, 2012 (kopā ar G. Krieviņu un L. Aleksejevu).
6. 5th Conference Applied Information and Communication Technology, Jelgava, Latvija, 26–27 Aprīlis, 2012 (kopā ar L. Aleksejevu).
7. EMCSR 2012 (European Meetings on Cybernetics and Systems Research) Vīne, Austrija, l0– 13 Aprīlis, 2012 (kopā ar L. Aleksejevu).
8. RTU 52th International Scientific Conference, Rīga, Latvija, 13–16 Oktobris, 2011 (kopā ar L. Aleksejevu un I. Tuleiko).
9. 8th International Scientific and Practical Conference «Environment. Technology. Resources», Rēzekne, Latvija, 20–22 Jūnijs, 2011 (kopā ar L. Aleksejevu).
10. Mendel 17th International Conference on Soft Computing, Brno, Čehija, 15–17 Jūnijs, 2011 (kopā ar L. Aleksejevu un I. Tuleiko).
11. RTU 51th International Scientific Conference, Rīga, Latvija, 11–15 Oktobris, 2010 (kopā ar L. Aleksejevu un N. Novoselovu).
12. Mendel 16th International Conference on Soft Computing, Brno, Čehija, 23–25 Jūnijs, 2010 (kopā ar L. Aleksejevu).
13. RTU 50th International Scientific Conference, Rīga, Latvija, 12–16 Oktobris, 2009 (kopā ar L. Aleksejevu).

Publikācijas

Promocijas darbā veikto pētījumu rezultāti aprobēti projektā «Medicīnisko un bioloģisko datu intelektuālo metožu un apstrādes algoritmu kompleksa izstrāde onkoloģisko slimību diagnostikas pilnveidošanai» un atspoguļoti 15 zinātniskajās publikācijās.

1. Gasparovica-Asite M., Aleksejeva L. Fuzzy Classification Systems for Bioinformatics Data Analysis // *Scientific Journal of Riga Technical University. Computer science. Information Technology and Management Science*. – 2014. – Vol.17. – P. 92–97. **Cited by EBSCO, CSA/ProQuest, CNPIEC, Ulrich's Periodical Directory / ulrichsweb, WorldCat (OCLC), VINITI.**
2. Gasparovica M., Aleksejeva L., Nazaruks V. Using Fuzzy Clustering with Bioinformatics Data // Proceedings of the 6th International Conference on Applied Information and Communication Technologies (AICT2013), Latvia, Jelgava, 25–26 April 2013. – Jelgava: Latvia University of Agriculture, Faculty of Information Technologies, 2013. – P. 62–70.
3. Gasparovica M., Aleksejeva L., Gersons V. The Use of BEXA Family Algorithms in Bioinformatics Data Classification // Scientific Journal of Riga Technical University. Computer science. Information Technology and Management Science. – 2012. – Vol.15. – P.120–126. **Cited by EBSCO, CSA/ProQuest, Versita, VINITI.**
4. Gasparovica M., Aleksejeva L., Gersons V. Use of BEXA Family Algorithms in Bioinformatics Data Classification // Riga Technical University 53rd International Scientific Conference: Dedicated to the 150th Anniversary and the 1st Congress of World Engineers and Riga Polytechnical Institute / RTU Alumni: Digest, Latvija, Riga, 10–12 October, 2012. – Riga: RTU, 2012. – P. 89.
5. Gasparovica M., Krievina G., Aleksejeva L. Biological Interpretation of Metabolic Syndrome Data Missing Value Imputation and Classification // Proceedings of Workshop on Data Mining in Life Sciences DMLS'2012, Germany, Berlin, July 20, 2012. – Fockendorf: Ibai-Publishing, 2012. – P. 167–176.
6. Gasparovica M., Aleksejeva L. Feature Selection for Bioinformatics Data Sets – Is It Recommended? // Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012), Latvia, Jelgava, 26–27 April 2012. – Jelgava: Latvia University of Agriculture, Faculty of Information Technologies, 2012. – P. 325–335.
7. Gasparovica M., Aleksejeva L. Rule Weight Use in Bioinformatics Data Classification // European Meetings on Cybernetics and Systems Research: Book of Abstracts, Austria, Vienna, 11–13 April, 2012. – Vienna: Bertalanffy Center for the Study of Systems Science. – P. 229–231.
8. Gasparovica M., Tuleiko I., Aleksejeva L. Influence of Membership Functions on Classification of Multi-Dimensional Data // Scientific Journal of Riga Technical

University. Series 5. Computer science. Information Technology and Management Science. – 2011. – Vol. 49. – P. 78–84. **Cited by EBSCO, CSA/ProQuest, Versita, VINITI.**

9. Gasparovica M., Aleksejeva L. Brain Cancer Antibody Display Classification // Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. Latvia, Rezekne, 20–22 June, 2011. Vol. II. – Rezekne: Rezekne Higher Education Institution, 2011. – P. 9–15. **Cited by SCOPUS.**
10. Gasparovica M., Aleksejeva L., Tuleiko I. Finding Membership Functions for Bioinformatics Data // Proceedings of 17th International Conference on Soft Computing – MENDEL 2011, Czech Republic, Brno, 15–17 June, 2011. – Brno: Brno University of Technology, 2011. – P. 133–140. **Cited by Thomson Reuters Web of Science and SCOPUS.**
11. Gasparovica M., Aleksejeva L. Using Fuzzy Unordered Rule Induction Algorithm for Cancer Data Classification // Proceedings of 17th International Conference on Soft Computing – MENDEL 2011, Czech Republic, Brno, 15–17 June, 2011. – Brno: Brno University of Technology, 2011. – P. 141–147. **Cited by Thomson Reuters Web of Science and SCOPUS.**
12. Gasparovica M., Aleksejeva L. Using Fuzzy Algorithms for Modular Rules Induction // Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science. – 2010. – Vol. 44. – P. 94–98. **Cited by EBSCO, CSA/ProQuest, VINITI.**
13. Gasparovica M., Novoselova N., Aleksejeva L. Using Fuzzy Logic to Solve Bioinformatics Tasks // Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science. – 2010. – Vol. 44. – P. 99–105. **Cited by EBSCO, CSA/ProQuest, VINITI.**
14. Gasparovica M., Aleksejeva L. A Comparative Analysis of Prism and MDTF Algorithms // Proceedings of 16th International Conference on Soft Computing – MENDEL 2010, Czech Republic, Brno, 23–25 June 2010. – Brno: Brno University of Technology, 2010. – P. 191–197. **Cited by Thomson Reuters Web of Science and SCOPUS.**
15. Gasparoviča M., Aleksejeva L. A Study on the Behaviour of the Algorithm for Finding Relevant Attributes and Membership Functions // Scientific Journal of Riga Technical University. Series 5. Computer Science. Information Technology and Management Science.– 2009. – Vol. 40. – P.75–80. **Cited by EBSCO, CSA/ProQuest, VINITI.**

Galvenie rezultāti

Darba izstrādes gaitā iegūtie galvenie rezultāti, sasniegumi un pētījumu apkopojums.

- Definētas prasības klasifikācijas algoritmiem, kas īpaši paredzēti bioinformātikas datu apstrādei.

- Veikts pētījums par datu pirmapstrādes metožu lietojumu bioinformātikas datu analīzē, lai noskaidrotu tās metodes, ko iespējams izmantot un kas ir efektīvas bioinformātikas datu pirmapstrādē.
- Praktisku eksperimentu gaitā noskaidrotas veiksmīgākās metodes un algoritmi, jāizmanto izplūdušās klasifikācijas sistēmas izstrādē.
- Izstrādāta klasterizācijā balstīta piederības funkciju konstruēšanas metode.
- Izvēlētajam klasifikācijas algoritmam pievienota likumu stiepšanas un izstrādātā likumu fazifikācijas metode.
- Uz praktisko eksperimentu un izstrādāto metožu bāzes izveidota izplūdusī klasifikācijas metodoloģija, kas piemērota tieši bioinformātikas datu klasifikācijai.
- Uz izstrādātās metodoloģijas pamata izveidota izplūdusī klasifikācijas sistēma.
- Izstrādātā klasifikācijas sistēma pārbaudīta eksperimentos ar reāliem bioloģijas datiem, kas ļāva izdarīt secinājumus par izstrādātās sistēmas un metodoloģijas iespējām.

Promocijas darba struktūra un apraksts

Pirmajā nodaļā ir dots darbā risināmā uzdevuma definējums, kā arī bioinformātikas, datu ieguves un izplūduma jēdzienu vispārējs definējums un galvenās risināmās problēmas.

Otrā nodaļa veltīta promocijas darbā izmantojamo algoritmu aprakstam, izpētei un salīdzinošai analīzei un to iepriekšējam lietojumam bioinformātikā.

Trešajā nodaļā aprakstīta izstrādātā izplūdušās klasifikācijas metodoloģija, apkopojot informāciju un dodot pamatojumu par katrā metodoloģijas daļā izmantoto metodi/algoritmu, balstoties uz praktisko eksperimentu rezultātiem

Ceturtajā nodaļā aprakstītas promocijas darba gaitā izstrādātas izplūdušās klasifikācijas metodoloģijas modifikācijas –klasterizācijā balstīta piederības funkciju konstruēšana, *FURIA* likumu stiepšana un izstrādātā likumu fazifikācijas metode. Dota arī izstrādātās izplūdušās klasifikācijas metodoloģijas arhitektūra.

Piektajā nodaļā dots izstrādātās sistēmas apraksts, sistēmas praktiskas lietošanas rezultāti, dots eksperimentos izmantojamo datu kopu apraksts, kā arī eksperimentāli noteikto labāko sistēmas parametru apraksts.

Darbu noslēdz **rezultātu analīzes un secinājumu nodaļa**, kurā apkopoti iegūtie rezultāti un secinājumi par paredzētas bioinformātikas datu analīzei izplūdušās klasifikācijas metodoloģijas izstrādi un praktisku lietojumu.

1. Bioinformātika, datu ieguve un izplūduma lietošana

Vairāku dažādu zinātņu nozaru sadarbība problēmu risināšanā pēdējos gados ir kļuvusi par veiksmīgu praksi. Arī biologu sadarbība ar informācijas tehnoloģiju speciālistiem ir perspektīva pētījumu joma. Bioinformātikā tiek iegūtas zināšanas no bioloģiskiem datiem datoranalīzes ceļā. Tā var būt informācija, kas tiek glabāta ģenētiskajā kodā, arī eksperimentu rezultāti, kas iegūti no dažādiem avotiem, izmantojot pacientu statistiku un zinātnisko literatūru. Kopš 1999. gada Goluba raksta [94], bioinformātikas jautājumu risināšanā tiek lietotas arī datu ieguves tehnoloģijas. Bioinformātikas trīs galvenie risināmie jautājumi – gēnu un proteīnu struktūras un funkcijas pētījumi, liels datu un zināšanu apjoms, kā arī jaunu zināšanu iegūšana. Šajā darbā aprakstītā metodoloģija atbilst lielā datu un zināšanu apjoma analīzei, un rezultāts ir predefinētu grupu iegūšana jaunu ierakstu grupu piederības noteikšanai, lietojot antivielu, dažādu pētījumos populāru datu kopu un metabolisma datu klasifikācijai [81].

Ņemot vērā bioinformātikas datu sarežģīto dabu – lielo atribūtu (pat vairāki desmiti tūkstoši) un salīdzinoši nelielo ierakstu skaitu (līdz simts), nepieciešamas metodes, kas spētu šajos datos atrast likumsakarības, kas būtu arī bioloģiski interpretējamas. Datu ieguves lietošana šādas likumsakarības palīdz atrast, apstrādāt un attēlot arī biologiem saprotamā veidā. Izplūduma klasifikācijas algoritmu lietošana nodrošinās biologiem un mediķiem viegli uztveramu klasifikācijas likumu lietošanu.

1.1. Uzdevuma definējums

Šajā darbā tiek risināta klasifikācijas problēma – jaunu ierakstu klasificēšana pēc iepriekš apmācībā iegūtiem likumiem (jo izmantotas gēnu un proteīnu datu kopas, kurās doti gēnu un proteīnu līmeņi un mērķa klase), kā arī notiek apmācība ar skolotāju (jo zināma klases vērtība). Klasifikācija tiek veikta, lietojot izplūdumu. Veicot klasifikāciju, jāņem vērā datu specifiskās īpašības [85]:

1. ļoti liels atribūtu skaits (pat vairāki desmiti tūkstošu) un salīdzinoši neliels ierakstu skaits (mazāks nekā 100, nereti pat mazāks nekā 50);
2. liels atribūtu skaits (pamatā bioinformātikā atribūti ir gēni), kas nenes nekādu klasifikācijā nozīmīgu informāciju;
3. troksnis sākuma datos;
4. klasifikācijas rezultātu bioloģiskā interpretējamība;
5. iespēja apstrādāt datus no dažādiem avotiem.

Formālais uzdevuma apraksts var tikt standartizēts ar datu ieguves jēdzieniem. Dota datu kopa X, kurā ietilpst m objekti $x_i = x_i^{A1}, \dots, x_i^{An}, x_i^C$, kur $i \in m$, kas ir ierakstu atbilstošo

vērtību vektori un klases atribūta C vērtība. Klasifikācija tiek veikta ar izplūdušo algoritmu palīdzību, tāpēc nepieciešams no sākotnējiem datiem pāriet uz izplūdušiem datiem, kas darāms ar piederības funkciju konstruēšanu, kur katra x_i vērtība kļūst par fazificētu vektoru $u_i = (\mu_{Ai1}(x_i))$, kur $\mu_{Ai1} \in [0,1]$. Izplūdušais klasifikators katrai klasei aprēķina izplūdušo piederības vērtību $\mu C_k(x_i)$, kas atbilst sakarībai, ka x_i pieder klasei C_k. Rezultātā iegūti likumi: ja x_i ir vērtība A_{ij}, tad klase ir C.

1.2. Datu ieguves jēdziens un lietojums

Datubāzu pirmsākumi meklējami 20. gs. 60. gados ar vienkāršu datu apstrādi, tās gadu gaitā kļuvušas arvien sarežģītākas, un datubāzēs mūsdienās tiek glabāts ievērojams informācijas daudzums [59]. Cilvēkam šādu informācijas daudzumu apstrādāt ir neiespējami, tāpēc tehnoloģijas, ar kurām to iespējams darīt, kļūst arvien populārākas. Datu ieguve ir relatīvi jauns zinātnes virziens, jo tā kļuvusi aktuāla līdz ar datoru evolūciju un informācijas glabāšanu elektroniski. Datu ieguve piedāvā tehnoloģijas un metodes, ar kuru palīdzību iespējams apstrādāt un izgūt informāciju no datubāzēm.

Datu ieguves procesu var iedalīt trīs daļās [57]: **Pirmapstrādes posms**, kurā notiek datu apstrāde un sagatavošana datu ieguves algoritmu lietošanai; **Modeļa veidošanas un validācijas posms**, kurā notiek modeļu veidošana, izmantojot dažādus datu ieguves algoritmus, un veiksmīgākais no tiem tiek izvēlēts tālākajai izmantošanai; **Modeļa lietošanas posms** paredzēts iegūtā modeļa lietošanai jaunu datu apstrādē, lai iegūtu pareizu prognozi aplēses problēmas atrisināšanai.

Datu pirmapstrāde ir viens no datu ieguves posmiem, kas var būt pat 80 % no visa datu ieguves procesa. Šis process ir jebkura datu ieguves projekta pamats, un no tā veiksmīgas realizācijas atkarīgi visa datu ieguves projekta rezultāti. Ja dati netiek pienācīgi sagatavoti, arī no tiem iegūtā informācija un zināšanas var būt maldinošas vai neprecīzas [132]. Šajā darbā tiek pētītas trūkstošo datu apstrādes metodes, atribūtu atlases metodes un piederības funkciju konstruēšanas metodes.

Ja aplūko **trūkstošo datu apstrādes** tehnoloģiju klasifikāciju pēc darbībām, kas jāveic to lietošanai [120], tās tiek iedalītas:

- trūkstošo datu ignorēšana vai atribūtu/ierakstu ar trūkstošām vērtībām izdzēšana no sākotnējās datu kopas;
- atribūtu, kritēriju novērtēšana – īpaši algoritmi, kas var novērtēt trūkstošo datu nozīmi;
- trūkstošo vērtību aizvietošana.

Atribūtu/ierakstu ar trūkstošām vērtībām izdzēšana no sākotnējās datu kopas nav piemērota, jo bioinformātikas ierakstu skaits tāpat ir proporcionāli neliels, lai kādu izdzēstu, un atribūtu izdzēšana nav vēlama, jo nav iegūta informācija par katra atribūta derīgumu.

Novērtēšanas algoritmu izmantošana nav lietderīga, jo ierakstu skaits netiks samazināts un atribūtu atlasei ir atsevišķa pirmapstrādes procesa daļa, līdz ar to tālākiem pētījumiem derīga trūkstošo datu aizvietošanas pieeja.

Datu kopas **nenozīmīgo un lieko dimensiju samazināšana** var tikt sadalīta divās apakšproblēmās [27].

- Atribūtu atlase – tiek atrasti atribūti, kas ir nozīmīgi, un no sākotnējās datu kopas izslēgti atribūti ar lieku un/vai nenozīmīgu informāciju.
- Ierakstu atlase – tieši tāpat kā daži atribūti ir nozīmīgāki (derīgāki) nekā citi, tā ir arī ar ierakstiem (piemēriem).

Ņemot vērā datu specifiku – salīdzinoši mazo ierakstu skaitu pret ievērojami lielāko atribūtu skaitu, ierakstu atlasi lietot nav lietderīgi, tāpēc tiks pētīta tikai atribūtu atlase.

Piederības funkciju konstruēšana ir datu pirmapstrādes metode, kas raksturīga tikai izplūdušiem datiem. Pareizas piederības funkcijas konstruēšanas metodes izvēle var ievērojami ietekmēt iegūto rezultātu, jo, izmantojot piederības funkciju konstruēšanu, notiek datu transformācija un pāreja uz stāvokli, kad viena ieraksta viena vērtība vienlaikus pieder vairākām klasēm. Dota datu kopa X, izplūdusī apakškopa F no X ir definēta ar tās piederības funkciju $\mu_F: X \rightarrow [0,1]$. Apkopojot dažādos avotos pausto, lai konstruētu piederības funkcijas, ir iespējams izmantot [16, 74] divas pieejas (abas tiks pētītas praktiski):

- pieeja, kurā tiek izmantotas ekspertu zināšanas;
- datos balstītā pieeja, kur notiek automātiska piederības funkciju ģenerēšana.

Modeļa veidošanas un validācijas posmā tiek lietota klasifikācija, kas ir viens no datu ieguves algoritmu veidiem, kas palīdz atklāt datos esošas zināšanas, izmantojot informāciju par klasi, un tās vēlāk izmanto jaunu, iepriekš nezināmu piemēru klasificēšanā [59]. Šā darba kontekstā izplūduma jēdziens attiecas uz situācijām, kurās neprecizitātes avots ir nevis nejauši mainīgie vai procesi, bet jēdzienu klases, kurām nav stingri definēti ierobežojumi. Izplūduma lietošana datu ieguvē, konkrēti – klasifikācijas algoritmā, paver plašas iespējas operēt ar cilvēkam viegli uztveramiem vērtējumiem, nodrošinot labāku izpratni par konkrētu atribūta vērtību ietekmi uz klasifikācijas rezultātu.

2. Datu ieguves metodes un to lietojums bioinformātikā

Ir pieejamas daudz un dažādas trūkstošo datu aizvietošanas metodes. Lai noskaidrotu veiksmīgāko lietojamo metodi, nepieciešams veikt eksperimentālus pētījumus, jo veiksmīgākās metodes izvēle ir atkarīga no datu kopas, kurā ir trūkstošas vērtības un trūkstošo vērtību skaita datu kopā.

Pēdējās desmitgades laikā motivācija lietot atribūtu atlases metodes bioinformātikā mainījusies no ilustratīviem piemēriem uz priekšnoteikumu modeļu veidošanai. It īpaši mikrorežģu analīzē, kur atribūtu atlases metodes kļuvušas par *de facto* standartu. Neskatoties uz to, ka atribūtu atlases metodes var tikt lietotas apmācībā gan ar skolotāju, gan bez, tomēr biežāk tās pētītas gadījumos ar skolotāju, t. i., klasifikācijā, kad klases vērtība ir zināma iepriekš.

Izpētīti klasiskie datu ieguves klasifikācijas algoritmi *JRIP* (*Repeated Incremental Pruning to Produce Error Reduction* algoritma versija), *FURIA* (*Fuzzy Unordered Rule Induction Algorithm*), *SVM* (*Support Vector machines*), *KNN* (*K–Nearest Neighbor*), *NB* (*Naive Bayes*), *CART* (*Classification and regressian trees*) un C4.5 (lēmumu koku algoritms) un to lietojums bioinformātikā. Secināts, ka visi apskatītie algoritmi tiek izmantoti dažādos bioinformātikas pētījumos [3, 8, 17, 28, 60, 70, 86, 110, 114, 116, 122, 126], tāpēc tos iespējams lietot arī šajā darbā veiktajos eksperimentos izstrādātās metodoloģijas un sistēmas iegūto rezultātu salīdzināšanai.

Analizēti K–vidējo sadalošais algoritms [15, 33] un X–vidējo klasterizācijas algoritms [102], un to darbības galvenie principi. Aplūkots to lietojums bioinformātikas datu apstrādē. Secināts, ka iespējams lietot klasterizācijas algoritmus piederības funkciju konstruēšanā.

Aplūkotas klasifikācijas rezultātu novērtējuma metodes, no kā secināts, ka nepieciešams lietot šķērsvalidāciju tālākiem eksperimentiem, lai novērstu subjektivitāti datu apstrādē, un klasifikācijas rezultātus novērtēt, izmantojot neskaidrības matricu.

Tā kā bioinformātikas datu klasifikācijas problēma definē specifiskas īpašības – liels atribūtu skaits (vairāki tūkstoši) – un nav nekādas informācijas par to, kurš atribūts varētu būt nozīmīgs, vislielāko perspektīvu no aplūkotajiem algoritmiem uzrāda *FuzzyBEXA* [123] algoritms, kura galvenais mīnuss ir tas, ka algoritma klasifikācijas rezultāti ir atkarīgi no algoritma uzstādījumiem. Tomēr, lietojot atbilstošus algoritma uzstādījumus, klasifikācijas precizitāte var pieaugt, tādējādi iespējams iegūt labākus klasifikācijas likumus. *FuzzyBEXA* algoritmam ir arī vairāki plusi – iespējams izmantot jebkuru piederības funkciju konstruēšanas metodi, algoritms darbojas gan ar skaitliskiem, gan kategoriskiem datiem, kā arī nav nekādu specifisku ierobežojumu attiecībā uz datu kopas lielumu (atribūti un ieraksti).

3. Izplūdušās klasifikācijas metodoloģijas daļu eksperimentāla izvēle

Bioinformātikas datu specifiskās dabas dēļ metodoloģijai tika izvirzītas prasības, kas apkopotas 3.1. tabulā. Lai novērtētu izstrādātās izplūdušās klasifikācijas sistēmas iespējas reālu klasifikācijas uzdevumu atrisināšanā, tika veikta eksperimentu sērija ar 25 reālām bioinformātikas datu kopām: antivielu datiem no Latvijas biomedicīnas pētījumu un studiju

centra (BMC); internetā publiski pieejami dažādu pētnieku apkopoti gēnu dati un metabolisma pētījumu dati no Latvijas Universitātes Medicīnas fakultātes.

3.1. tabula

Prasību pamatojums un atbilstošs iekļaujamais metodoloģijas solis/metode

Prasība	Pamatojums	Iekļaujamais solis/metode
Datu attīrīšana	Trūkstoši, kļūdaini sākuma dati, ņemot vērā to iegūšanas veidu, t. i., medicīniski pētījumi.	Trūkstošo vērtību apstrāde
Informācijas no dažādiem datu avotiem, apstrāde	Dati var būt iegūti no vairākiem avotiem, ārstu pierakstiem, analīžu rezultātiem.	Normalizācija
Datu apstrāde, kur ierakstu skaits ir vairākas reizes mazāks nekā atribūtu skaits	Bioinformātikas datu specifiskā daba – ierakstu skaits visbiežāk ir pacientu skaits (var būt pat < 50) un atribūti ir gēni, kuru skaits ir mērāms pat vairākos desmitos tūkstošu.	Atribūtu atlase
Atribūtu skaita samazināšana	Samazinot nesvarīgo atribūtu skaitu, tiek uzlabots klasifikācijas laiks, tai pat laikā būtiski nesamazinot klasifikācijas precizitāti.	Atribūtu atlase
Bioloģiski interpretējamu klasifikācijas likumu ģenerēšana	Iegūtajiem klasifikācijas likumiem jābūt tādā formā, ka tos viegli iespējams saprast un interpretēt arī biologiem un mediķiem.	Ja–tad likumu lietošana

Teorētiskā pētījuma apraksta gaitā tika noskaidrots, ka izstrādātai klasifikācijas metodoloģijai – pēc pirmās un otrās nodaļas pētījumiem un pēc [39] – jābūt četrām galvenajām daļām.

1. Datu pirmapstrāde:
    a. trūkstošo vērtību apstrādāšana;
    b. atribūtu skaita samazināšanas;
    c. piederības funkciju konstruēšanas.
2. Klasifikatora apmācīšana un likumu bāzes izveidošana.
3. Jaunu piemēru klasificēšana (klasifikatora pārbaudes jeb testēšanas fāzes).
4. Rezultātu novērtēšana.

Trūkstošo vērtību praktiskiem eksperimentiem izvēlētas Vienas klases visbiežāk iespējamās atribūta vērtības ievietošana [58], Globālā visbiežāk iespējamā atribūta vērtības ievietošana [58], K–tuvāko kaimiņu ievietošana [10], K–vidējo klasterizācijas algoritma pielietošana trūkstošo vērtību aizpildīšanai [120], Izplūdušā K–vidējo klasterizācijas algoritma lietošana trūkstošo vērtību aizpildīšanai [120], kā arī Atbalsta vektoru mašīnu uz regresiju balstītās trūkstošo vērtību aizpildīšanas pieeja [6]. Praktiski eksperimenti veikti ar datu kopām, kam ir trūkstošās vērtības, t. i., BrCa, GaCa, PrCa, GIS, Mel un Meta, veicot atribūtu atlasi ar *FastCoreletionBasedFilterSolution* [130] metodi. Pēc iegūtajiem rezultātiem secināts, *in vivo*

pētījumu gadījumā (gan klīnisko, gan preklīnisko) nav iespējams lietot trūkstošo datu apstrādes metodes, bet *in vitro* pētījumus veic labi kontrolētos laboratorijas apstākļos, līdz ar to vērtības var aizvietot. Gadījumā, kad sākotnējos datus (ar trūkstošajām vērtībām) iespējams veiksmīgi apstrādāt arī bez datu aizvietošanas, no bioloģijas viedokļa datus labāk neaizvietot. Ja tomēr tiek nolemts datus aizvietot, tad atbilstoši eksperimentiem labākos klasifikācijas rezultātus uzrāda K–vidējo, K–tuvāko kaimiņu un Svērtā K–tuvāko kaimiņu ievietošana.

Atribūtu atlase veikta ar 74 *Weka* [118] programmatūrā esošajām piemērotām atribūtu meklēšanas un novērtēšanas metožu kombinācijām ar trim datu kopām – *MLL* (jauktās leikēmijas datu kopa), *Ch.ALL_2* (bērnu akūtas limfoblastiskās leikēmijas datu kopa) un *GastricCancer3* (kuņģa vēža datu kopa). Veikts arī katras atribūtu atlases metodes darbības novērtējums, skatot rezultātā iegūto datu kopu klasifikācijas precizitāti, lai varētu noteikt, vai atribūtu skaita samazinājums ietekmē klasifikācijas rezultātu, izmantojot – *JRIP, FURIA, SVM, KNN, NB, CART, C4.5* un *FuzzyBexa. FuzzyBexa* klasifikācijas algoritms uzrāda vienu no trim labākajiem klasifikācijas rezultātiem gandrīz katrā gadījumā, līdz ar to tas pierāda šā algoritma konkurētspēju ar citiem bioinformātikā izmantotiem datu ieguves algoritmiem. Apkopojot visu trīs pilno eksperimentu iegūtos klasifikācijas rezultātus ar atribūtu atlases un vērtēšanas metodēm, ņemot vērā arī klasifikācijas laiku un samazinātās datu kopas atribūtu skaitu, tika secināts, ka turpmākai izmantošanai piemērotas atribūtu atlases un meklēšanas metožu kombinācijas: *FilteredSubsetEval+LinearForwardSelection, WrapperSubsetEval+IWSSembeddedNB* (strādā ātrāk nekā citas metožu kombinācijas), *CfsSubsetEval+ LinearForwardSelection, Symmetrical UncertAttributeSetEval+FCBF* (strādā ātrāk nekā citas metožu kombinācijas), *Consistency SubsetEval+ LinearForwardSelection* un *ConsistencySubsetEval+ RerankingSearch.*

Veicot literatūras pētījumu, netika iegūta informāciju par to, kāda būtu **pirmapstrādes metožu izmantošanas rekomendētā secība**, tāpēc tika veikta eksperimentu sērija, lai to noskaidrotu. Vislabākie klasifikācijas rezultāti iegūti, vispirms veicot trūkstošo vērtību aizvietošanu, tad atribūtu atlasi un beigās klasifikāciju, līdz ar to arī izstrādātajā izplūdušās klasifikācijas metodoloģijā lietota šāda secība.

Piederības funkciju konstruēšanai lietot eksperta metodes nav lietderīgi, jo nav pieejams problēmsfēras eksperts, tāpēc tiek nolemts izmantot matemātiskās piederības funkciju konstruēšanas metodes. Vienkāršā un ērti izmantojamā trijstūra piederības funkciju konstruēšanas metode [19] neizmanto nekādu papildu informāciju par datu kopu, tikai katra konkrētā atribūta vērtības sadala proporcionālos intervālos, tāpēc būtu vajadzīga kāda piederības funkciju konstruēšanas metode, kas izmantotu arī zināšanas par datiem.

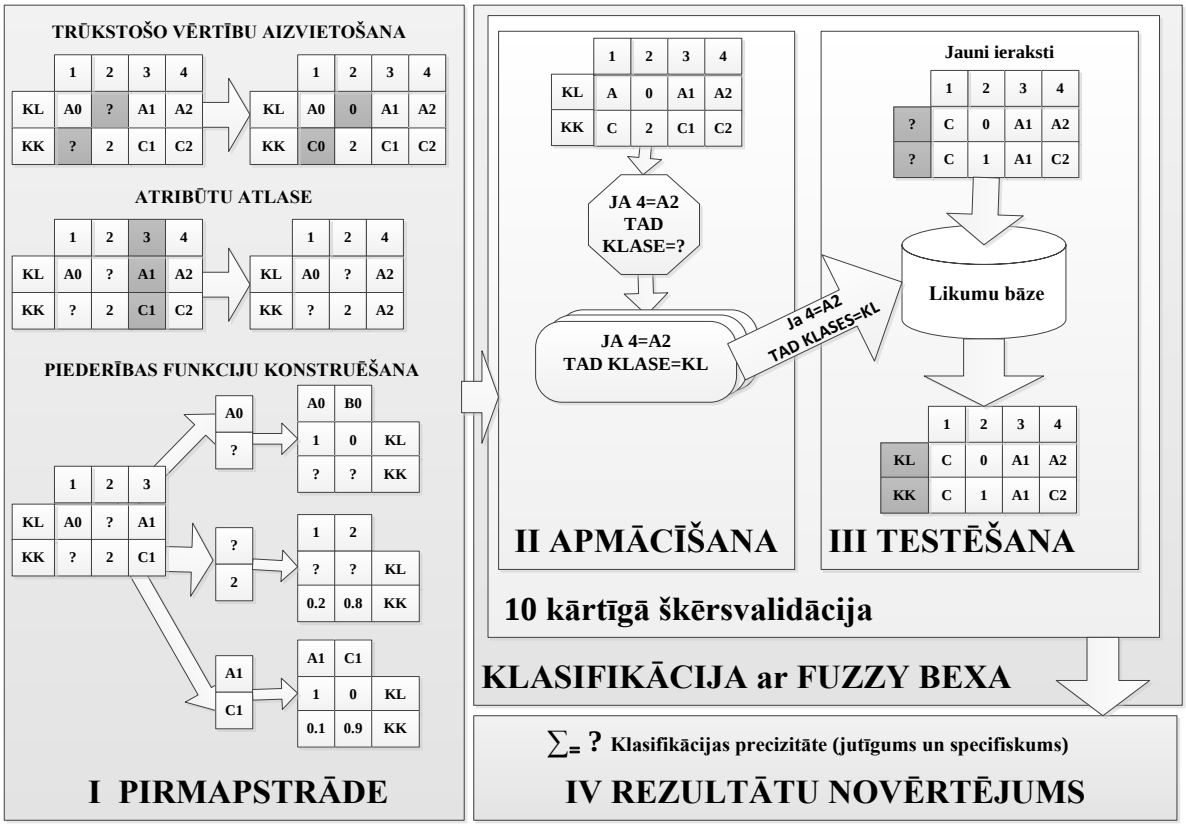
Klasifikācija veikta ar ***FuzzyBexa*** klasifikācijas algoritmu, tomēr šajā algoritmā nav pieejama neviena likumu pēcapstrādes metode, tāpēc tiek nolemts uzlabot klasifikācijas

metodoloģiju, tajā iekļaujot likumu apstrādes metodes, tas ļautu pārklāt visus piemērus, kas ir sākotnējā datu kopā.

Lai klasifikācijas rezultātu novērtējumi būtu objektīvāki, un, ņemot vērā bioinformātikas datu jau iepriekš aprakstīto specifiku, kas izpaužas mazajā ierakstu skaitā, atbilstošākā metode klasifikatora veiktspējas vērtēšanai ir **šķērsvalidācija**, konkrētāk – desmitkārtīgā šķērsvalidācija, līdz ar to apmācības un testēšanas fāzes nav atdalāmas un tiks apskatītas kopā. Diagnostiskos testos visbiežāk tiek lietots tieši jutīgums un specifiskums, lai raksturotu iegūtos rezultātus [113], līdz ar to no bioloģiskās puses klasifikācijas rezultātu novērtēšanai jālieto **jutīgums** un **specifiskums**, bet no datu ieguves puses – **klasifikatora kopējā precizitāte**.

4. Izplūdušās klasifikācijas metodoloģija un izstrādātie uzlabojumi

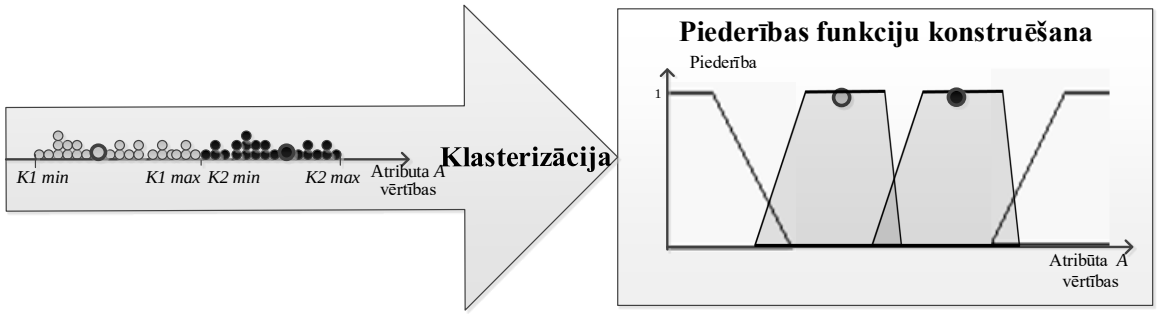
Apkopojot iepriekšējās nodaļās aprakstīto, rezultātā ir izveidota izplūdušās klasifikācijas metodoloģija, kuras soļu detalizētu aprakstu iespējams redzēt 4.1. attēlā. Metodoloģijas pirmajā daļā tiek veikta pirmapstrāde: trūkstošo vērtību aizvietošana, atribūtu atlase, piederības funkciju konstruēšana.



4.1. att. **Izplūdušās klasifikācijas metodoloģijas daļu detalizēts apraksts**

Otrajā – klasifikācijas – daļā tiek veikta klasifikācija, klasifikatora veiktspējas novērtēšanai tiek lietota desmitkārtīga šķērsvalidācija. Iegūtie klasifikācijas likumi tiek saglabāti likumu bāzē un piedalās jaunu piemēru klasifikācijā. Klasifikācijas rezultāti tiek novērtēti ar klasifikatora kopējo precizitāti, jutīgumu un specifiskumu. Piederības funkciju konstruēšanai lietot eksperta metodes nav lietderīgi, jo nav pieejams problēmsfēras eksperts, tāpēc tiek nolemts izmantot matemātiskās piederības funkciju konstruēšanas metodes. Vienkāršā un ērti izmantojamā trijstūra piederības funkciju konstruēšanas metode neizmanto nekādu papildu informāciju par datu kopu, tikai katra konkrētā atribūta vērtības sadala proporcionālos intervālos, tāpēc būtu vajadzīga kāda piederības funkciju konstruēšanas metode, kas izmantotu arī zināšanas par datiem.

Tika nolemts izstrādāt **jaunu piederības funkciju konstruēšanas metodi**, kas ņemtu vērā papildu informāciju par datiem, t. i., izmantot klasterizācijā iegūto informāciju par klasteru centroīdiem un klasteru minimālās un maksimālās vērtības (skat. grafisku attēlojumu 4.2. attēlā). Balstoties uz piederības funkciju konstruēšanas metodi, kas tiek izmantota izplūdušās klasterizācijas uzdevumos [74], un ņemot vērā tās universālo iespēju metodi izmantot ar jebkuru klasterizācijas algoritmu, tika nolemts metodi uzlabot, lietošanai tieši klasifikācijas algoritmā.

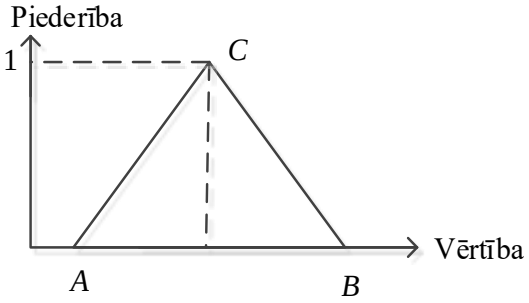


4.2. att. **Klasterizācijas informācijas izmantošana piederību konstruēšanā**

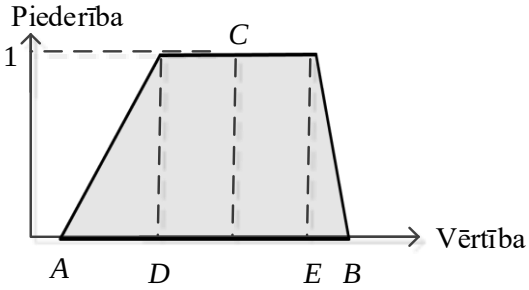
Metodei pievienota trapeces veida piederības funkciju konstruēšanas iespēja. Izstrādātās metodes sākumam – klasterizācijai – ērti pieslēgt jebkuru klasterizācijas algoritmu, jo metodes darbam nepieciešama tikai informācija par klasteru skaitu (vai tas ir uzdots klasterizācijas algoritmam, vai izrēķināts), klasteru centroīdu vērtībām un katra klastera minimālajām un maksimālajām vērtībām. Ņemot vērā bioinformātikas datu sarežģītību ar lielo atribūtu skaitu un mazo ierakstu skaitu, tika nolemts piederības funkciju konstruēšanai izmantot gaballineārās funkcijas.

Šajā metodē k ir pārklāšanās koeficients, kas raksturo to, cik viena lingvistiskā vērtība pārklāj citu. Uz klasterizāciju balstītās piederības funkciju konstruēšanas metodes darbības algoritms darbojas katram atribūtam no sākotnējās datu kopas secīgi:

- tiek veikta datu kopas klasterizācija katram atribūtam secīgi. No klasterizācijas algoritma tiek iegūts klasteru skaits K_{sk}, katra atribūta klastera minimālā un maksimālā vērtība – K_{min} un K_{max}, kā arī katra atribūta klastera centroīda vērtība K_c;
- tiek konstruētas gaballineāras piederības funkcijas, piemēram, trijstūra (skat. formulu (4.1.)) vai trapeces veida (skat. formulu (4.2.));

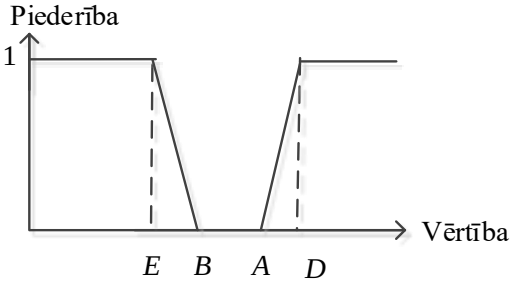


$$\begin{cases} C = K_c; \\ A = K_{min} - (K_{min} * k); \\ B = K_{max} + (K_{max} * k). \end{cases} \quad (4.1.)$$



$$\begin{cases} A = K_{min} - (K_{min} * k_1); \\ B = K_{max} + (K_{max} * k_2); \\ D = K_{min}; \\ E = K_{max.} \end{cases} \quad (4.2.)$$

- Malējiem intervāliem vērtības tiek konstruētas šādi:



$$\begin{cases} A = K_{min} - (K_{min} * k_1); \\ B = K_{max} + (K_{max} * k_2); \\ D = K_{min}; \\ E = K_{max}. \end{cases} \quad (4.3.)$$

- katras vērtības piederība tiek noteikta, nolasot to no aprēķinātajām vērtībām; pēc sākotnējo piederības funkciju noteikšanas tiek veikta piederības funkciju normalizācija, iegūstot piederības funkcijas, kas atbilst formulai:

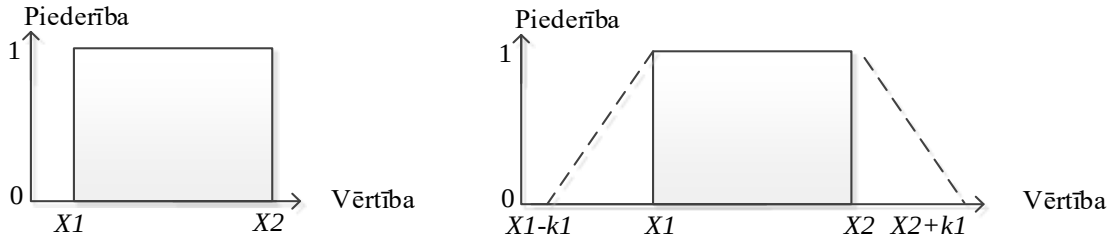
$$\sum_{i=1}^{n} \mu_s(x_i) = 1. \quad (4.4.)$$

Neskatoties uz to, ka piederības funkciju konstruēšanai iespējams lietot jebkuru klasterizācijas algoritmu, no kura iespējams iegūt klasteru skaitu, centroīdu vērtības un katra

klastera minimālo un maksimālo vērtību, ņemot vērā pasaules populārākos un biežāk lietotos algoritmus [119], nolemts iekļaut *K–means* klasterizācijas algoritmu un tā modifikāciju *X–means* klasterizāciju.

FuzzyBexa klasifikācijas algoritmam tiek pievienota ***FURIA* likumu stiepšanas stratēģija** (to definējot kā ieslēdzamu funkcionalitāti), lai pārklātu visus sākotnējos piemērus datu kopā, kas nav pārklāti ar sākotnējiem klasifikācijas likumiem. Likuma vispārināšana vai «stiepšana» tiek realizēta, izdzēšot vienu vai vairākus tās nosacījumus. Līdz ar to minimālā likumu vispārināšana ir realizējama, izdzēšot likuma nosacījumus, kas neatbilst klasificējamam piemēram. Šī stratēģija tiek iedarbināta tikai gadījumā, kad kāds piemērs nav pārklāts ar atbilstošu likumu.

Izmantojot ideju par likumu fazifikāciju, tika izveidota **likumu fazifikācijas stratēģija**, kas balstīta uz piederības funkciju konstruēšanas mehānismu, t. i., izmanto piederības funkciju konstruēšanas gaitā iegūtās vērtības, paplešot likumus atbilstoši k pārklāšanās koeficientam. Ja salīdzina likumu bez fazificēšanas un likumu ar fazificēšanu, tad (skatīt 4.3. attēlu) uzskatāmi redzamas atšķirības gan likumu nosacījumu daļā, gan arī vērtību piederībā. Attēla (a) daļā redzams likums bez fazifikācijas, bet (b) daļā likums ar fazifikāciju.



Ja vērtība pieder [*X1, X2*], tad klase *K*

(a)

Ja vērtība pieder [(*X1-k1*), (*X2+k1*)], tad klase *K*

(b)

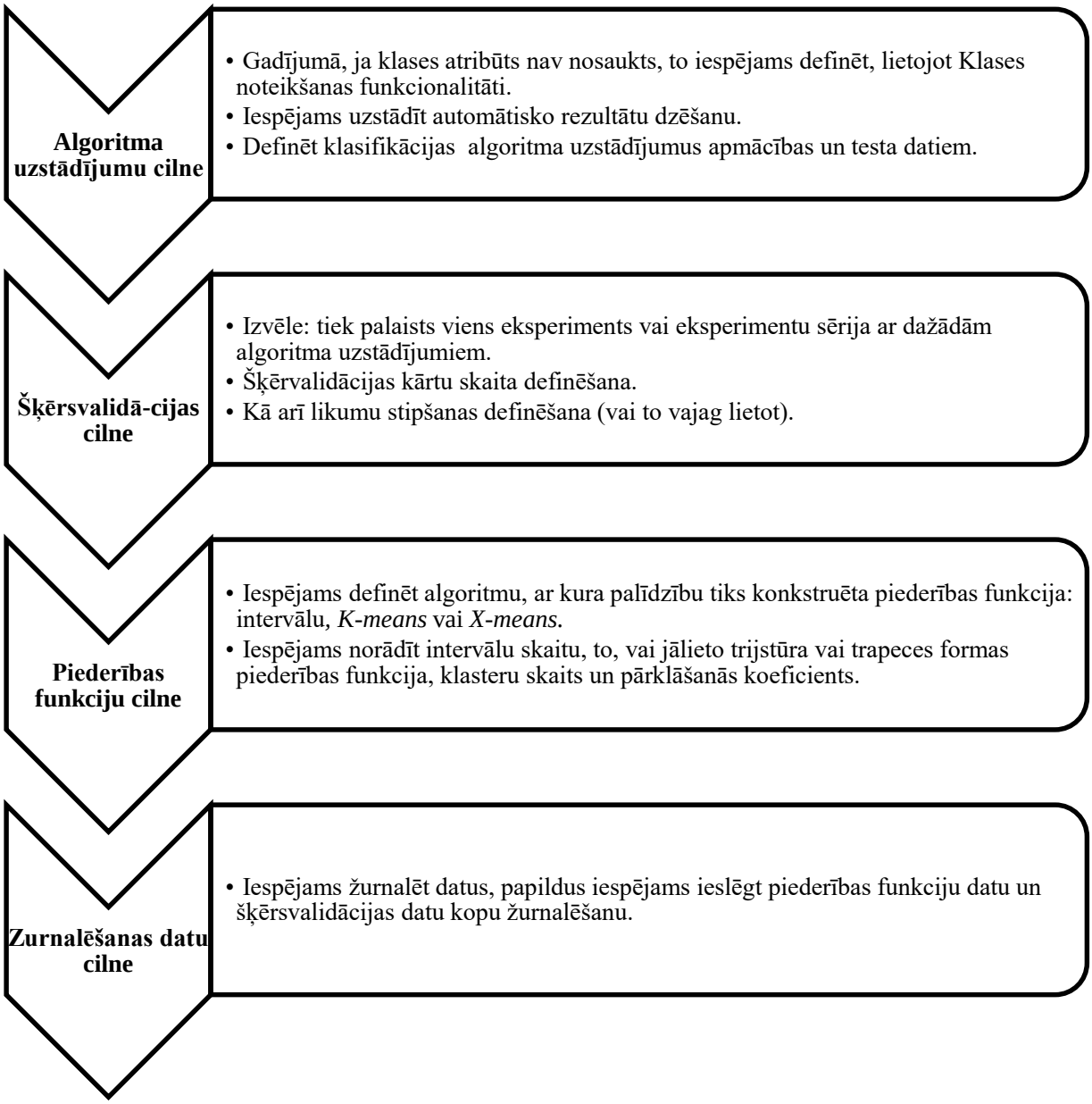
4.3. att. **Piederības funkcijas stingram un fazificētam likumam**

FuzzyBexa klasifikācijas algoritms tiek papildināts ar *FURIA* likumu stiepšanas stratēģiju, lai varētu pārklāt visus sākotnējās kopas piemērus. Izstrādāta likumu fazifikācijas stratēģija, kas paplašina iegūto klasifikācijas likumu darbības apmērus, t. i., ar fazificēto likumu palīdzību iespējams klasificēt arī ierakstu, kuru vērtības ir līdzīgas ierakstu vērtībām, kas piedalījušās apmācībā, bet nav identiskas.

5. Izplūdušās klasifikācijas sistēmas lietojumi

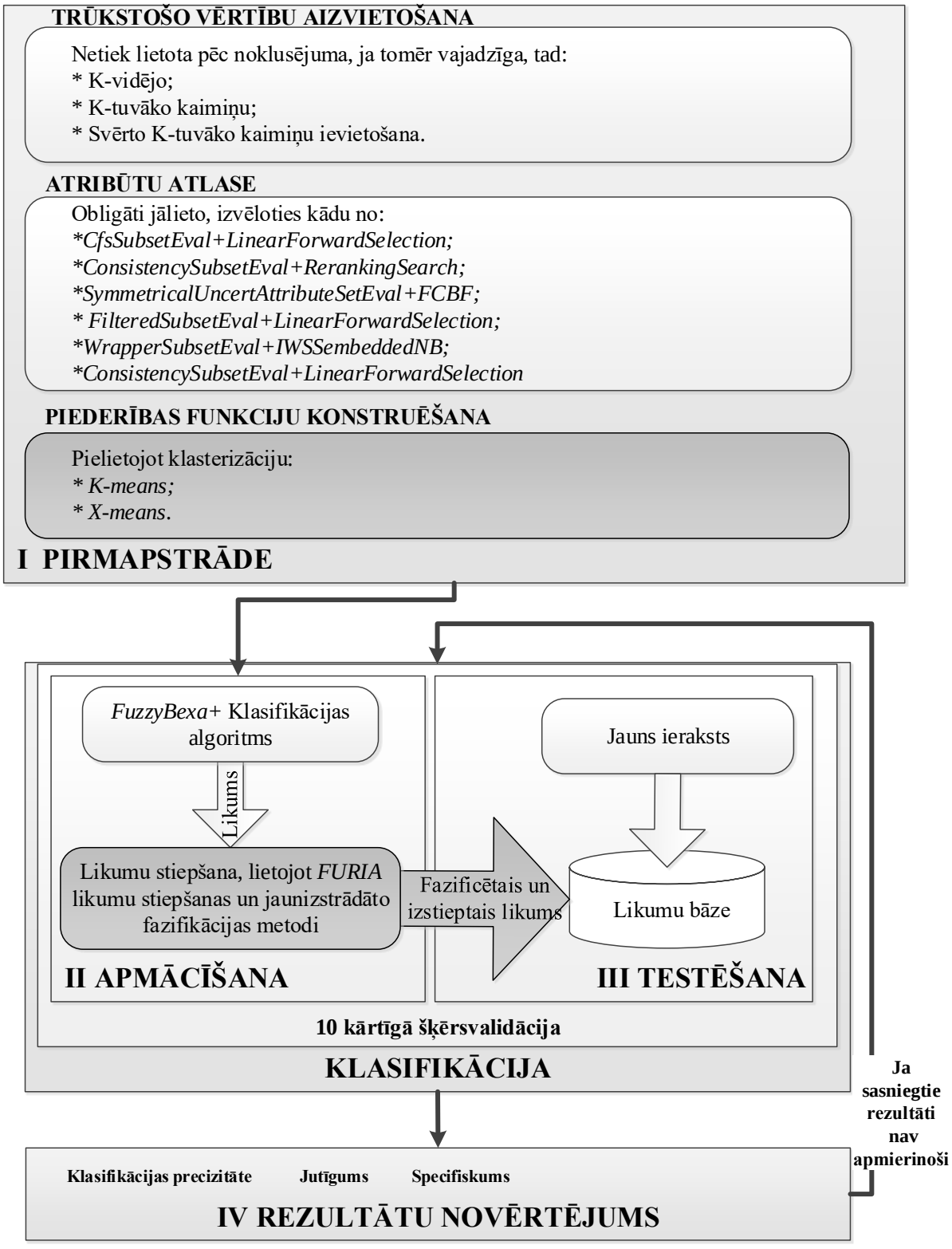
Izstrādātā klasifikācijas algoritma *FuzzyBexa+* realizācija ar papildu funkcionalitāti ir sistēma, kas ir lokāla *Java* lietotne. Tajā tiek veikta piederības funkciju konstruēšana un

klasificēšana. Pārējie datu pirmapstrādes soļi notiek ārpus sistēmas. Lietotnei ir četras cilnes, to apkopojums redzams 5.1. attēlā.



5.1. att. ***FuzzyBexa+* lietotnes cilņu funkcionalitāte**

Katrā solī veiksmīgākās, izplūdušās klasifikācijas sistēmā iekļaujamās metodes apkopotas 5.2. attēlā. Ieteikums būtu paralēli lietot visas metodes, tādējādi nodrošinot objektīvu rezultātu, bez nejaušības, ka vienas metodes izvēle varētu pārāk ietekmēt klasifikācijas rezultātu.



5.2. att. **Izplūdušās klasifikācijas sistēmas shēma**

Tika veikta eksperimentu sērija ar *DLBCL*, *Prostate*, *Medulloblast* un *Glioblast* datu kopām, lietojot visas sešas rekomendētās atribūtu atlases metodes, un apkopoti klasifikācijas rezultāti, skat. 5.1. tabulā (tikai *Medulloblast* kopai), ar pelēku iezīmēti labākie trīs katras rindas

rezultāti, bet ar I – atzīmēts *FuzzyBaxa+* algoritms ar intervālu piederības funkciju konstruēšanas metodei, ar X-piederības funkciju konstruēšanas metodi, kas balstīta uz *X-means* klasterizāciju, bet ar K-piederības funkciju konstruēšanas metodi, kas balstīta uz *K-means* klasterizāciju. Pēc tabulā apkopotajiem rezultātiem var secināt, ka sešu dažādu atribūtu atlases metožu kombināciju izmantošana ir attaisnojusies, jo var redzēt, cik klasifikācijas rezultāti ir atšķirīgi, t. i., vienai datu kopai izvēloties atbilstošu metodi, kopējās klasifikācijas precizitātes rezultāti variē no 87 % līdz pat 65 % neveiksmīgākajai metodei. *Medulloblast* un *DLBCL* datu kopām *FuzzyBexa* klasifikācijas rezultāti ir konkurētspējīgi ar citām populārajām datu ieguves metodēm, bet *Prostate* un *Glioblast* datu kopām tikai, izvēloties veiksmīgāko un atbilstošāko atribūtu atlases metodi. Tas skaidrojams ar sākotnējo atribūtu skaitu *Medulloblast* un *DLBCL* datu kopām tas ir zem 8000, bet – *Prostate* un *Glioblast* – virs 10000. Līdz ar to – jo lielāka datu kopa, jo svarīgāka ir atribūtu atlases metode, kas no lielā apjoma izvēlas klasifikācijai derīgus atribūtus.

5.1. tabula

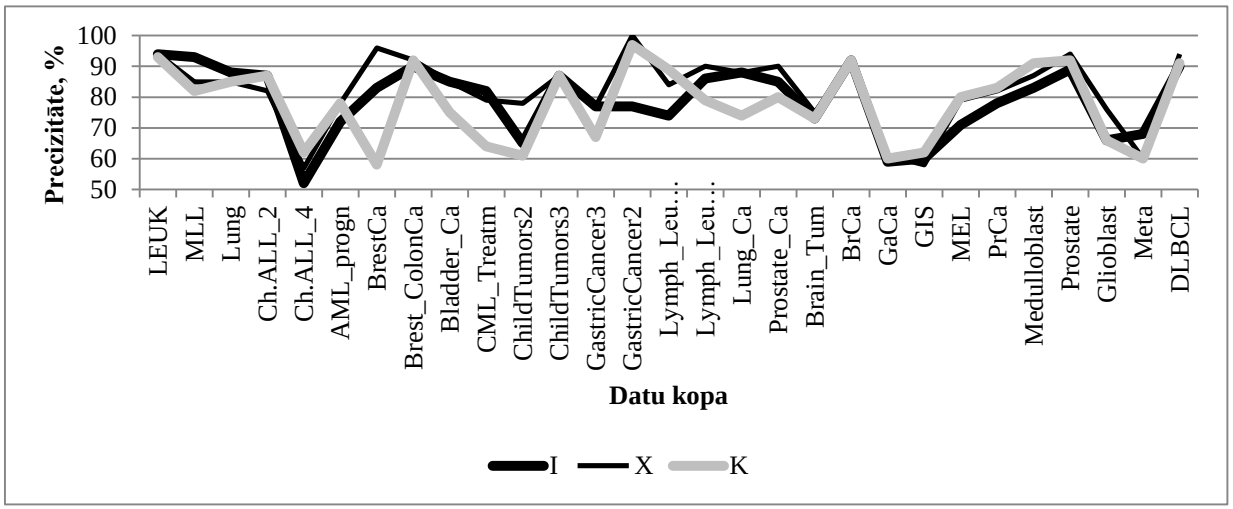
Kopējā klasifikācijas precizitāte, %

Datu kopa	Atribūtu atlases meklēšanas un novērtēšanas metode	Top 10 datu ieguves algoritmi							FuzzyBexa+		
		CART	C4.5	JRip	Furia	KNN	SVM	NB	I.	X	K
Medullo blast	*CfsSubsetEval+ LinearForwardSelection*	65	83	57	78	91	91	91	83	87	83
	ConsistencySubsetEval+ RerankingSearch	74	87	87	83	74	57	74	52	65	78
	SymmetricalUncertAttributeSetEval+FCBFSearch	78	83	74	83	91	96	91	74	83	78
	FilteredSubsetEval+ LinearForwardSelection	*78*	*87*	*78*	*91*	*91*	*91*	*87*	*83*	*87*	*91*
	WrapperSubsetEval+ IWSSembeddedNB	91	87	87	96	91	65	96	74	87	83
	ConsistencySubsetEval+ LinearForwardSelection	87	91	83	96	91	65	96	70	78	78
...											

Tuvāk apskatot *FuzzyBexa+* klasifikācijas algoritma rezultātus ar visām piederības funkciju konstruēšanas metodēm (skat. 5.3. attēlu), izvēloties labāko no sešām atribūtu atlases pieeju kombinācijām ar noklusējuma uzstādījumiem, redzams, ka uz klasterizāciju balstītās piederības funkciju konstruēšanas metodes uzrāda labākus vai vienādus klasifikācijas rezultātus ar matemātisko intervālu metodi gandrīz visos gadījumos. Kopumā 86 % gadījumu uz klasterizāciju balstīto piederības funkciju iegūtie klasifikācijas rezultāti ir labāki vai tādā pašā līmenī nekā ar intervālu metodi iegūtie.

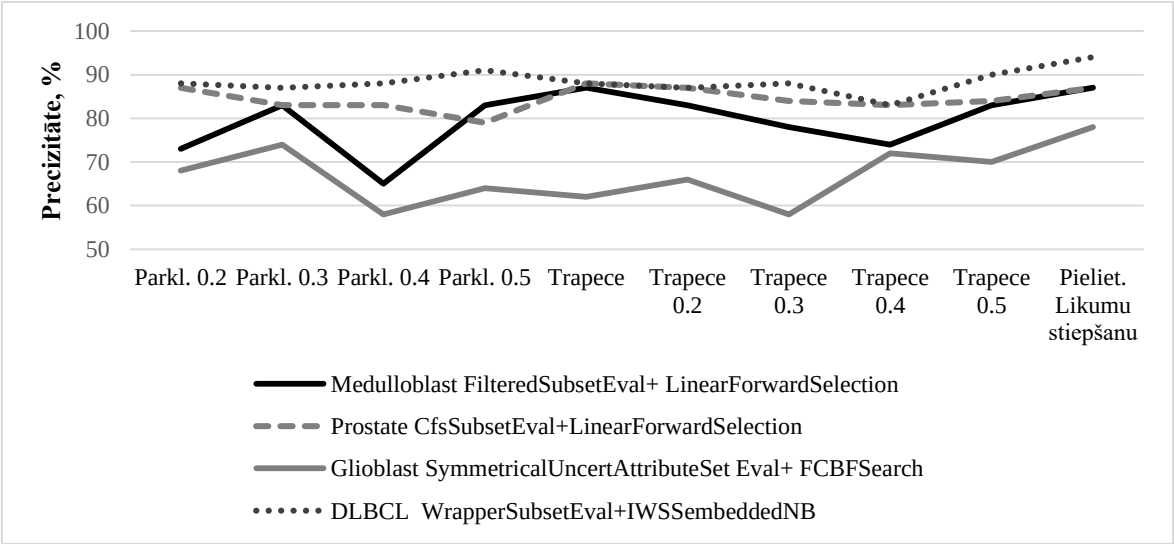
Ja aplūko jutīgumu un specifiskumu, jāsecina, ka katrai datu kopai tas uzrāda citādus rādītājus. *Medulloblast* datu kopai labākais rezultāts ir pie *C4.5, FURIA, KNN, SVM, NB* un

FuzzyBexa K-means algoritmiem, *Prostate* datu kopai – *JRIP, KNN, SVM, NB* un *FuzzyBexa X-means*. Toties *DLBCL* datu kopai *KNN, NB* un *FuzzyBexa+ X-means*.



5.3. att. **Kopējā klasifikatora precizitāte ar piederības funkciju konstuēšanas metodēm**

Kā noskaidrots iepriekšējos eksperimentos, izvēloties intervālu skaitu – pieci, un jebkuru α_I vērtību no 0,1–0,8, eksperimenta rezultāti nemainīsies, tāpēc tos var uzskatīt par universāliem algoritma uzstādījumiem. Skatot no eksperimentu rezultātiem, datu kopām, kam ir līdz 8000 atribūtiem, rekomendēts *X-means* algoritmam izmantot α_T vērtību 0,7 un 0,8, bet *K-means* gadījumā 0,3 un 0,4. Toties datu kopām, kam atribūtu skaits ir lielāks par 10000, rekomendēts abām piederības funkciju konstruēšanas metodēm izmantot α_T vērtības 0,3 un 0,8.



5.4. att. **Klasifikācijas precizitātes saistība ar piederības funkciju parametriem**

Tika veikta eksperimentu sērija ar dažādiem piederības funkciju konstruēšanas parametriem, iegūtie klasifikācijas rezultāti visām četrām datu kopām redzami 5.4. attēlā. Skatītas dažādas pārklāšanās koeficienta vērtības, trapeces formas piederības funkciju konstruēšanas metodes, kā arī – lietojot likumu stiepšanu. Pēc attēla ver secināt, ka labākie *FuzzyBexa+* parametri piederības funkciju konstruēšanai – likumu stiepšana, kā arī jālieto visas kombinācijas, izvēloties veiksmīgāko: pārklāšanās koeficients 0,3; trapeces formas piederības funkcijas; trapeces formas piederības funkcijas ar pārklāšanās koeficientu 0,4.

Rezultāti un secinājumi

Promocijas darbā ir izstrādāta izplūdušās klasifikācijas metodoloģija bioinformātikas datu apstrādei, kā arī piedāvāta uz klasterizāciju balstīta piederības funkciju konstruēšanas metode un likumu fazifikācijas metode. Darba izstrādes gaitā sasniegti šādi rezultāti:

- izpētīti bioinformātikas dati un definētas prasības klasifikācijas algoritmam, kas ar šādiem datiem strādā;
- izpētītas datu pirmapstrādes metodes un izvēlētas piemērotākās bioinformātikas datu apstrādei;
- veikti apjomīgi praktiski eksperimenti ar 26 datu kopām, lai noskaidrotu algoritmus un metodes, kas jāizmanto uz izstrādātās metodoloģijas balstītās izplūdušās klasifikācijas sistēmas izstrādē;
- izveidota uz klasterizāciju balstītā piederības funkciju konstruēšanas metode;
- izstrādāta likumu fazifikācijas metode, kas paplašina likumu nosacījumus;
- izveidota izplūdusī klasifikācijas metodoloģija bioinformātikas datu apstrādei un analīzei;
- izstrādāta izplūdušās klasifikācijas sistēma bioinformātikas datu apstrādei un analīzei;
- veikta izstrādātās metodoloģijas un sistēmas lietošana reāliem bioinformātikas datiem.

Darba izstrādes gaitā tika veikta literatūras analīze, lai noskaidrotu prasības, kas jāņem vērā, izstrādājot metodoloģiju, kas paredzēta neprecīziem datiem un balstās uz izplūdušo kopu teoriju. Skatītas dažādas pirmapstrādes metodes un to izmantošanas iespējamību bioinformātikas datu apstrādē un analīzē. Ar perspektīvākajām metodēm veikti plaši praktiski eksperimenti ar 26 reālām bioinformātikas datu kopām un noskaidrotas katrā pirmapstrādes solī piemērotākās metodes. Izpētīta arī veiksmīgākā pirmapstrādes soļu secība. Izstrādāta uz klasterizāciju balstīta piederības funkciju konstruēšanas metode, kas izmanto *X-means* un *K-means* klasterizācijas algoritmus. Izvēlētajam klasifikācijas algoritmam *FuzzyBexa* pievienota *FURIA* likumu stiepšanas metode. Izstrādāta arī likumu fazifikācijas metode, kas izmanto

piederības funkciju konstruēšanas gaitā iegūto informāciju. Visas atrastās un izstrādātās metodes apvienotas izplūdušās klasifikācijas metodoloģijā. Uz metodoloģijas pamata izstrādāta izplūdusī klasifikācijas sistēma. Izstrādātā metodoloģija un uz tās pamata izveidotā klasifikācijas sistēma ir lietojamas arī citu bioinformātikai līdzīgu datu klasifikācijā, kam ir neliels ierakstu skaits (aptuveni 100) un proporcionāli ļoti liels atribūtu skaits (pat vairāki desmiti tūkstoši).

Veikta sistēmas eksperimentālā validācija ar četrām jaunām, iepriekš eksperimentos neizmantotām bioinformātikas datu kopām. Tika pārbaudītas abas izvirzītās hipotēzes, un to rezultāti ir šādi:

- pirmā hipotēze tika pierādīta eksperimentāli, salīdzinot uz klasterizāciju balstītu piederības funkciju konstruēšanas metožu iegūtos klasifikācijas rezultātus ar matemātiski izrēķinātu piederības funkciju klasifikācijas rezultātiem. 25 no 29 gadījumiem uz klasterizāciju balstītās metodes uzrāda labākus klasifikācijas rezultātus;
- otrā hipotēze tika pierādīta eksperimentāli, testa nolūkos radot līdzīgu, bet ne identisku, ierakstu sākotnējā datubāzē esošajam. Jaunais ieraksts tika veiksmīgi klasificēts, izmantojot fazificētos likumus. Ja likums nebūtu fazificēts, šis ieraksts netiktu klasificēts veiksmīgi.

Darba izstrādes gaitā un eksperimentālā validācijā iegūtie secinājumi par izplūdušo klasifikācijas metodoloģiju un uz to balstīto sistēmu:

- salīdzinot oriģinālo datu kopu klasifikācijas precizitāti ar trūkstošo datu aizvietoto datu kopu klasifikācijas precizitāti, labāki rezultāti uzrādīti aizvietoto datu gadījumā, līdz ar to trūkstošo datu apstrādes metožu izmantošana uzrāda labākus rezultātus, tomēr šo metožu izmantošana jāapsver no bioloģiskā viedokļa. *In vivo* pētījumu gadījumā (gan klīnisko, gan preklīnisko) nav iespējams lietot trūkstošo datu apstrādes metodes, bet *in vitro* pētījumus veic labi kontrolētos laboratorijas apstākļos, visbiežāk izmantojot standartizētu šūnu, mikroorganismu vai vīrusa kultūru, līdz ar to var lietot trūkstošo datu aizvietošanas metodes;
- klasifikācijas rezultāti, izmantojot atribūtu atlases metodes, salīdzinot ar oriģinālo datu kopu, nepasliktinās, bet gan pat uzlabojas, tādējādi šo pirmapstrādes soli noteikti var rekomendēt izmantošanai arī sarežģītajiem bioinformātikas datiem;
- vislabākie klasifikācijas rezultāti iegūti, vispirms veicot trūkstošo vērtību aizvietošanu, tad atribūtu atlasi un beigās klasifikāciju, līdz ar to arī šāda secība tiek rekomendēta praktiskiem eksperimentiem;

- ja sākotnējā datu kopā ir vairāk nekā 10000 atribūtu, svarīga ir atribūtu atlases metodes izvēle, lai sasniegtu labāku klasifikācijas rezultātu. Tāpēc rekomendēts izmantot visas sešas atribūtu atlases metodes;
- eksperimentāli izvēloties labāko no sešām atribūtu atlases metodēm, arī *FuzzyBexa* klasifikācijas rezultāti ir konkurētspējīgi ar citām populārām datu ieguves metodēm un visu četru datu kopu gadījumā ir starp trim labākajiem. Līdz ar to iespējams pielietot *FuzzyBexa+* izplūdušo klasifikācijas algoritmu, nezaudējot precizitāti;
- uz klasterizāciju balstītās piederības funkciju konstruēšanas metodes uzrāda labākus klasifikācijas rezultātus vairumā gadījumu, t. i., 86 % gadījumu (25 datu kopām no 29) uz klasterizāciju balstīto piederības funkciju iegūtie klasifikācijas rezultāti ir labāki vai tādā pašā līmenī nekā ar intervālu metodi iegūtie;
- ja jāizvēlas viena uz klasterizāciju balstītā piederības funkciju konstruēšanas metode, rekomendēts izmantot *X–means* metodi, jo tai nav iepriekš jāzina klasteru skaits un iegūtie rezultāti tādi paši un labāki nekā *K-means*;
- *FuzzyBexa+* optimālie parametri piederības funkciju konstruēšanai – likumu stiepšana, kā arī jālieto visas kombinācijas, izvēloties veiksmīgāko: pārklāšanās koeficients 0,3; trapeces formas piederības funkcijas; trapeces formas piederības funkcijas ar pārklāšanās koeficientu 0,4;
- optimālie *FuzzyBexa+* uzstādījumi – intervālu skaits – pieci, un jebkuru α_I vērtību no 0,1–0,8 eksperimenta rezultāti nemainīsies. Datu kopām, kam ir līdz 8000 atribūtiem, rekomendēts *X-means* algoritmam izmantot α_T vērtību 0,7 un 0,8, bet *K-means* gadījumā 0,3 un 0,4. Toties datu kopām, kam atribūtu skaits ir lielāks par 10000, rekomendēts abām piederības funkciju konstruēšanas metodēm izmantot α_T vērtības 0,3 un 0,8.

Nākamie pētījuma attīstības virzieni varētu būt saistīti ar citu klasterizācijas algoritmu lietošanu uz klasterizāciju balstītajā piederības funkciju konstruēšanas procesā, kā arī pašas uz klasterizāciju balstītās metodes uzlabošanā.

Bibliogrāfijas saraksts

1. A Fuzzy Inductive Learning Strategy for Modular Rules / C. H. Wang, J. F. Liu, T. P. Hong, et al. // *Fuzzy Sets and Systems*. – 1999. – Vol. 103. – P. 91–105.
2. Almuallim H., Dietterich T. G. Efficient Algorithms for Identifying Relevant Features // *Proceedings of 9th Canadian Conference on Artificial Intelligence* (11–15 May 1992). – Vancouver, BC: Morgan Kaufmann, 1992. – P. 38–45.
3. Analysis of Mass Spectrometry Data from the Secretome of an Explant Model of Articular Cartilage Exposed to Pro-inflammatory and Anti-inflammatory Stimuli using Machine Learning

/ A. L. Swan, K. L. Hillier, J. R. Smith et al. // *BMC Musculoskeletal Disorders*. – 2013. – Vol. 14:349.– P. 1–12. – Available from Internet: http://www.biomedcentral.com/1471-2474/14/349.

4. Ang K. K, Quek C. Supervised Pseudo Self-Evolving Cerebellar algorithm for generating fuzzy membership functions // *Expert Systems with Applications*. – 2012. – Vol. 39, Issue 3. – P. 2279–2287.
5. A Quantifier-based Fuzzy Classification System for Breast Cancer Patients / D. Soria, J. M. Garibaldi, A. R. Green, et.all. // *Artificial* Intelligence in *Medicine*. – 2013. – Vol. 58 (3). – P. 175–184.
6. A SVM Regression Based Approach to Filling in Missing Values / H. A. B. Feng, G. Chen, C. Yin et al. // *Proceedings of 9th International Conference on Knowledge-Based and Intelligent Information and Engineering Systems (KES2005)*. – Berlin Heidelberg: Springer–Verlag, 2005. – P. 581–587. (LNAI 3683).
7. Auesukaree C. cDNA Microarray Technology for the Analysis of Gene Expression // *KMITL Science and Technology Journal*. – 2006. – Vol. 6, No. 1. – P. 29–34.
8. Automatic Diagnosis of Melanoma Using Machine Learning Methods on a Spectroscopic System / L. Li, Q. Zhang, Y. Ding, et al. // *BMC Medical Imaging*. – 2014. – Vol. 14:36. – P. 1–12. – Available from Internet: http://www.biomedcentral.com/1471-2342/14/36.
9. Babu M. M. An Introduction to Microarray Data Analysis // *Computational Genomics: Theory and Applications* / ed. by R. Grant. –Taylor & Francis, 2004. – P. 225–249.
10. Batista G. E. A. P. A., Monard M. C. An Analysis of Four Missing Data Treatment Methods for Supervised Learning // *Applied Artificial Intelligence*. – 2003. – Vol. 175, No. 5. – P. 519–533.
11. Bilgic T., Turksen I. B. Measurement of Membership Functions: Theoretical and Empirical Work // *Fundamentals of Fuzzy Sets: The Handbooks of Fuzzy Sets Series* / Ed. L. Zadeh, D. Dubois. Vol. 7. – Kluver Academy Publishers, 2000. – P. 195–227.
12. *Bioinformatics for Geneticists: A Bioinformatics Primer for the Analysis of Genetic Data* / Ed. M.R. Barnes. 2nd edition. – England: John Wiley & Sons, 2007. – 576 p.
13. *Bioinformatics*: Introduction [Internet]. – Portugal: Instituto Superior Técnico, 2007. – Available from internet: https://fenix.tecnico.ulisboa.pt/downloadFile/3779571263334/intro_bioinformatica_55.pdf.
14. Bioinformatics Laboratory home page [Internet]. – University of Ljubljana, Faculty of Computer and Information Science, 2008. – Available from Internet: http://www.biolab.si/supp/bi-cancer/projections/index.htm.
15. Bishop C.M. *Neural Networks for Pattern Recognition*. – New York: Oxford University Press, 1995. – 504 p.
16. Borisovs A. N., Krumbergs O. A., Fjodorovs I.P. Decision – Making based on Fuzzy Models. – Riga: Zinatne, 1990. –184 p. (in Russian).
17. BrainCheck – a Very Brief Tool to detect Incipient Cognitive Decline: Optimized Case-Finding Combining Patient- and Informant-Based Data / M.M. Ehrensperger, **K.I. Taylor, M. Berres** et al. // *Alzheimer's* Research & *Therapy*. – 2014. – Vol. 6:69. – P. 1–12. – Available from Internet: http://alzres.com/content/6/9/69.
18. Casillas J. Genetic Feature Selection in a Fuzzy Rule – Based Classification System Learning Process for High – Dimensional Problems // *International Jurnal of Latest Trend in Computing*. – 2010. – Vol. 16. – P. 69–72.
19. Chutia R., Mahata S., Baruah H.K. An Alternative Method of Finding the Membersip of Fuzzy Number // *Information Sciences*. – 2000. – Vol. 136. – P. 135–157.
20. Chen C. H., Hong T. P., Tseng V. S. A Cluster-Based Fuzzy-Genetic Mining Approach for Association Rules and Membership Functions // *2006 IEEE International Conference on Fuzzy Systems*. – Vancouver, BC: IEEE, 2006. – P. 1411–1416.
21. Chen C. H., Hong T. P., Tseng V. S. A Comparison of Different Fitness Functions for Extracting Membership Functions Used in Fuzzy Data Mining // *Proceedings of the 1st IEEE Symposium on Foundations of Computational Intelligence (FOCI'07), 1–5 April 2007, Gonolulu, USA*. – IEEE,

2007. – P. 550–555.
22. Chromosomal Aberrations and Gene Expression Profiles in Non-Small Cell Lung Cancer / E. Dehan, A. Ben-Dor, W. Liao, et al. // *Lung Cancer*. – 2007. – Vol. 56, Issue 2. – P. 175–184.
23. *Classification and Regression Trees* / L. Breiman, J. H. Fridman, L. A. Olshen et al. – Washington, DC: Chapman & Hall / CRC, 1984. – 358 p. (Series: Wadsworth Statistics/Probability).
24. Cohen W. W. Fast Effective Rule Induction // *Machine Learning: Proceedings of the 12th International Conference (ML'95)*. – Morgan Kaufmann, 1995. – P. 115–123.
25. Combined Optimization of Feature Selection and Algorithm Parameters in Machine Learning of Language / W. Daelemans, V. Hoste, F. D. Meulder et al. // *Machine Learning: ECML 2003, Proceedings of 14th European Conference on Machine Learning, Cavtat-Dubrovnik, Croatia, September 22–26, 2003*. – Berlin Heidelberg: Springer-Verlag, 2003. – P. 84–95.
26. Cortes C., Vapnik V. Support-Vector Network // *Machine Learning*. – 1995. – Vol. 20. – P. 273–297.
27. *Data Mining and Knowledge Discovery Handbook* / Ed. O. Maimon, L. Rokach. –Berlin Heidelberg: Springer, 2010. – 1285 p.
28. Data Mining Techniques in a CGH-based Breast Cancer Subtype Profiling: An Immune Perspective with Comparative Study / F. Menolascina, S. Tomassi, P. Chiarappa et al. // *BMC Systems Biology*. – 2007. – Vol. 1 (Suppl 1): P 70. – Available from Internet: http://www.biomedcentral.com/1752-0509/1/S1/P70.
29. Data Preprocessing Techniques for Data Mining // *Winter School on «Data Mining Techniques and Tools for Knowledge Discovery in Agricultural Datasets»*. – India: Indian Agricultural Statistics Research Institute, 2002. – P. 139–144. – Available from internet: http://www.iasri.res.in/ebook/win_school_aa/notes/Data_Preprocessing.pdf.
30. *Datu ieguve. Pamati: Metodiskais līdzeklis* / A. Sukovs, L. Aleksejeva, K. Makejeva, A. Borisovs. – Rīga: Rīgas Tehniskā universitāte, 2007. – 130 lpp.
31. Diffuse Large B-Cell Lymphoma Outcome Prediction by Gene-Expression Profiling and Supervised Machine Learning / M. A. Shipp, K. N. Ross, P. Tamayo, et al. // *Nature Medicine*. – 2002. – Vol. 8, No. 1. – P. 68–74.
32. Dubois D., Prade H. What Are Fuzzy Rules and How to Use Them // *Fuzzy Sets and Systems*. – 1996. – Vol. 84. – P. 169–185.
33. Duda R. O., Hart P. E. *Pattern Classification and Scene Analysis*. – California: John Wiley&Sons, 1973. – 512 p.
34. Eineborg M., Boström H. Classifying uncovered examples by rule stretching // *ILP '01: Proceedings of the 11th International Conference on Inductive Logic Programming, Strasbourg, France, 9–11 September 2011*. – Berlin etc.: Springer-Verlag, 2001. – P. 41–50.
35. Expression Profiling of Medulloblastoma: PDGFRA and the RAS/MAPK Pathway as Therapeutic Targets for Metastatic Disease / T. J. MacDonald, K. M. Brown, B. LaFleur, et al. // *Nature Genetics*. – 2001. – Vol. 29 (2). – P. 143–152.
36. Fuzzy-based Classification of Breast Lesions using Ultrasound Echography and Elastography / S. Selvan, M. Kavitha, S. S. Devi, et al. // *Ultrasound Q*. – 2012. – Vol. 28 (3). – P. 159–167.
37. Fuzzy Inductive Learning Strategies / C. H. Wang, C. I. Tsai, T. P. Hong et al. // *Applied Intelligence*. – 2003. – Vol. 18, Issue 2. – P. 179–193.
38. *Fuzzy Logic for the Management of Uncertainty*. 1st edition / Ed. by L. A. Zadeh, J. Kacprzyk. – USA: John Wiley & Sons, 1992. – 676 p.
39. Gasparovica–Asite M., Aleksejeva L. Fuzzy Classification Systems for Bioinformatics Data Analysis // *Scientific Journal of Riga Technical University. Computer science. Information Technology and Management Science*. – 2014. – Vol. 17. – P. 92–97.
40. Gasparovica M., Aleksejeva L. A Comparative Analysis of Prism and MDTF Algorithms // *Proceedings of 16th International Conference on Soft Computing – MENDEL 2010, Czech Republic, Brno, 23–25 June 2010*. – Brno: Brno University of Technology, 2010. – P. 191–197.

41. Gasparoviča M., Aleksejeva L. A Study on the Behaviour of the Algorithm for Finding Relevant Attributes and Membership Functions // *Scientific Journal of Riga Technical University. Series 5. Computer Science. Information Technology and Management Science*.– 2009. – Vol. 40. – P. 75–80.
42. Gasparovica M., Aleksejeva L. Brain Cancer Antibody Display Classification // *Environment. Technology. Resources: Proceedings of the 8th International Scientific and Practical Conference. Latvia, Rezekne, 20–22 June, 2011.* Vol. II. – Rezekne: Rezekne Higher Education Institution, 2011. – P. 9–15.
43. Gasparovica M., Aleksejeva L. Feature Selection for Bioinformatics Data Sets – Is It Recommended? // *Proceedings of the 5th International Conference on Applied Information and Communication Technologies (AICT2012), Latvia, Jelgava, 26–27 April 2012*. – Jelgava: Latvia University of Agriculture, Faculty of Information Technologies, 2012. – P. 325–335.
44. Gasparovica M., Aleksejeva L. Rule Weight Use in Bioinformatics Data Classification // *European Meetings on Cybernetics and Systems Research: Book of Abstracts, Austria, Vienna, 11–13 April, 2012*. – Vienna: Bertalanffy Center for the Study of Systems Science. – P. 229–231.
45. Gasparovica M., Aleksejeva L., Gersons V. The Use of BEXA Family Algorithms in Bioinformatics Data Classification // *Scientific Journal of Riga Technical University. Computer science. Information Technology and Management Science*. – 2012. – Vol.15. – P. 120–126.
46. Gasparovica M., Aleksejeva L., Gersons V. Use of BEXA Family Algorithms in Bioinformatics Data Classification // *Riga Technical University 53rd International Scientific Conference: Dedicated to the 150th Anniversary and the 1st Congress of World Engineers and Riga Polytechnical Institute / RTU Alumni: Digest, Latvija, Riga, 10–12 October, 2012.* – Riga: RTU, 2012. – P. 89.
47. Gasparovica M., Aleksejeva L., Tuleiko I. Finding Membership Functions for Bioinformatics Data // *Proceedings of 17th International Conference on Soft Computing – MENDEL 2011, Czech Republic, Brno, 15–17 June, 2011*. – Brno: Brno University of Technology, 2011. – P. 133–140.
48. Gasparovica M., Aleksejeva L. Using Fuzzy Algorithms for Modular Rules Induction // *Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science*. – 2010. – Vol. 44. – P. 94–98.
49. Gasparovica M., Aleksejeva L. Using Fuzzy Unordered Rule Induction Algorithm for Cancer Data Classification // *Proceedings of 17th International Conference on Soft Computin – MENDEL 2011, Czech Republic, Brno, 15–17 June, 2011*. – Brno: Brno University of Technology, 2011. – P. 141–147.
50. Gasparovica M., Aleksejeva L., Nazaruks V. Using Fuzzy Clustering with Bioinformatics Data // *Proceedings of the 6th International Conference on Applied Information and Communication Technologies (AICT2013), Latvia, Jelgava, 25–26 April 2013.* – Jelgava: Latvia University of Agriculture, Faculty of Information Technologies, 2013. – P. 62–70.
51. Gasparovica M., Krievina G., Aleksejeva L. Biological Interpretation of Metabolic Syndrome Data Missing Value Imputation and Classification // *Proceedings of Workshop on Data Mining in Life Sciences DMLS'2012, Germany, Berlin, July 20, 2012*. – Fockendorf: Ibai-Publishing, 2012. – P. 167–176.
52. Gasparovica M., Novoselova N., Aleksejeva L. Using Fuzzy Logic to Solve Bioinformatics Tasks // *Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science*. – 2010. – Vol. 44. – P. 99–105.
53. Gasparovica M., Tuleiko I., Aleksejeva L. Influence of Membership Functions on Classification of Multi-Dimensional Data // *Scientific Journal of Riga Technical University. Series 5. Computer science. Information Technology and Management Science.* – 2011. – Vol. 49. – P. 78–84.
54. Gene Expression Profiling For The Prediction of Therapeutic Response to Docetaxel in Patients with Breast Cancer / J.C. Chang, E.C. Wooten, A. Tsimelzon, et al. // *Lancet*. – 2003. – Vol. 362 (9381). – P. 362–369.
55. Gene Expression Profiling Reveals Intrinsic Differences between T-cell Acute Lymphoblastic Leukemia and T-cell Lymphoblastic Lymphoma / E. A. Raetz, S. L. Perkins, D. Bhojwani, et al. // *Pediatric Blood & Cancer*. – 2006. – Vol. 47 (2). – P. 130–140.

56. Genetic Learning of Membership Functions for Mining Fuzzy Association Rules / R. Alcala, J. Alcala-Fdez, M. J. Gasto, et al. // *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'2007), 23–26 July, 2007, London*. – IEEE, 2007. – P. 1–6.
57. Gorunescu F. *Data Mining: Concepts, Models and Technologies*. – Springer–Verlag, 2011. – 360 p.
58. Handling Missing Attribute Values in Preterm Birth Data Sets / J. W. Grzymala-Busse, L. K. Goodwin, W. J. Grzymala-Busse et al. // *Proceedings of 10th International Conference of Rough Sets, Fuzzy Sets, Data Mining and Granular Computing (RSFDGrC'05), Regina, Canada, August 31–September 3, 2005*. Part II. – Vol. 4182. – Berlin Heidelberg: Springer-Verlag, 2005. – P. 342–351. (Lecture Notes in Computer Science, 3642).
59. Han J., Kamber M., Pie J. *Data Mining: Concepts and Techniques*. 2nd Edition. – San Francisco: Morgan Kaufmann Publishers, 2005. – 743 p.
60. Hatamikia S., Maghooli K., Nasrabadi A. L. The Emotion Recognition System Based on Autoregressive Model and Sequential Forward Feature Selection of Electroencephalogram Signals // *Journal of Medical Signals and Sensors*. – 2014. – Vol. 4 (3). – P. 194–201. – Available from Internet: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC4187354/.
61. Hippo Y. Global Gene Expression Analysis of Gastric Cancer by Oligonucleotide Microarrays // *Cancer Research*. – 2002. – Vol. 62 (1). – P. 233–240
62. Hogeweg P. The Roots of Bioinformatics in Theoretical Biology // *PLOS Computional Biology*. – 2011. – Vol. 7 (3). – P. 1–5. – Available from internet: http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002021.
63. Hong T. P., Chen J. B. Finding Relevant Attributes and Membership Functions // *Fuzzy Sets and Systems*. – 1999. – Vol. 103, No. 3. – P. 389–404.
64. Hong T. P., Chen J. B. Processing Individual Fuzzy Attributes for Fuzzy Rule Induction // *Fuzzy Sets and Systems*. – 2000. – Vol. 112. – P. 127–140.
65. Ho S.-Y. Interpretable Gene Expression Classifier with an Accurate and Compact Fuzzy Rule Base for Microarray Data Analysis // *BioSystems*. – 2006. – Vol. 85. – P. 165–176.
66. Huerta E. B., Duval B., Hao J.-K. Fuzzy Logic Elimination of Redundant Information of Microarray data // *Genomics, Proteomics & Bioinformatics*. – 2008. – Vol. 6, No. 2. – P. 61–73.
67. Hühn J., Hüllermeier E. FURIA: An Algorithm for Unordered Fuzzy Rule Induction // *Data Mining and Knowledge Discovery*. – 2009. – Vol. 19, No. 3. – P. 293–319.
68. Identification of a Gene Expression Signature Associated with Pediatric AML Prognosis / T. Yagi, A. Morimoto, M. Eguchi, et al. // *Blood*. – 2003. – Vol. 102 (5). – P. 1849–1856.
69. Identifying Distinct Classes of Bladder Carcinoma Using Microarrays / L. Dyrskjøt, T. Thykjaer, M. Kruhøffer, et al. // *Nature Genetics*. – 2003. – Vol. 33 (1). – P. 90–96.
70. Improve Discrimination Power of Serum Markers for Diagnosis of Cholangiocarcinoma using Data Mining-based Approach / S. Pattanapairoj, A. Silsirivanit, K. Muisik et al. // *Clinical Biochemistry*. – 2015. – Vol. 15. – Available from Internet: doi:10.1016/j.clinbiochem.2015.03.022.
71. In Chronic Myeloid Leukemia White Cells from Cytogenetic Responders and Non-responders to Imatinib Have Very Similar Gene Expression Signatures / L. C. Crossman, M. Mori, Y. C. Hsieh, et al. // *Haematologica*. – 2005. – Vol. 90 (4). – P. 459–464.
72. *Izplūdusī loģika, iespējamību teorija un to pielietojumi: Metodiskais līdzeklis* / A. Borisovs, L. Dubrovskis, L. Aleksejeva, et al. – Rīga: RTU, 1995. – 135 lpp.
73. Jafari P., Azuaje F. An Assessment of Recently Published Gene Expression Data Analyses: Reporting Experimental Design and Statistical Factors // *BMC Medical Information and Decision Making*. – 2006. – Vol. 6:27. – Available from Internet: http://www.biomedcentral.com/1472-6947/6/27.
74. Jamsandekar S. S. Mudholkar R. R. Self Generated Fuzzy Membership Function using ANN Clustering Technique // *International Journal of Latest Trends in Engineering and Technology (IJLTET)*. – 2013. – Special Issue - IDEAS-2013. – P. 142–152.

75. Jezewski M., Leski J. Cardiotocographic Signals Classification Based on Clustering and Fuzzy If-Then Rules // *5th European Conference of the International Federation for Medical and Biological Engineering IFMBE Proceedings. 14–18 September 2011, Budapest, Hungary*. Vol. 37. – Berlin Heidelberg: Springer-Verlag, 2012. – P. 121–124.
76. Kataria A., Singh M.D. A Review of Data Classification Using K-Nearest Neighbour Algorithm // *International Journal of Emerging Technology and Advanced Engineering*. – 2013. – Vol. 3, Issue 6. – P. 354–360.
77. KEEL Data-Mining Software Tool: Data Set Repository, Integration of Algorithms and Experimental Analysis Framework / J. Alcalá-Fdez, A. Fernandez, J. Luengo, et al. // *Journal of Multiple-Valued Logic and Soft Computing*. – 2011. – Vol. 17, No. 2-3. – P. 255–287.
78. Kira K., Rendell L.A. A Practical Approach to Feature Selection // *Proceedings of the 9th International Conference on Machine Learning* (D. Sleeman and P. Edwards, Eds.). – San Francisco, CA: Morgan Kaufman, 1992. – P. 249–256.
79. Kohavi R. A study of Cross–validation and Bootstrap for Accuracy Estimation and Model Selection // *Proceedings of the 14th International Joint Conference on Artifical Intelligence*. – San Mateo, CA: Morgan Kaufmann, 1995. – P. 1137–1143.
80. Koller D., Sahami M. Towards Optional Feature Selection // *Proceedings of 13th Conference on Machine Learning*. – San Francisco, CA: Morgan Kaufmann, 1996. – P. 284–292.
81. Li L., Wong L., Yang Q. Data Mining in Bioinformatics // *IEEE Intelligent Systems*. – IEEE Computer Society, 2005. – P. 16–18.
82. Li X. L., Tan Y. C., Ng S. K. Systematic Gene Function Prediction from Gene Expression Data by Using a Fuzzy Nearest-Cluster Method // *BMC Bioinformatics*. – 2006. – Vol. 7 (Suppl4):S23. – P. 1-11. – Available from Internet: http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1780124/.
83. Liu H., Setiono R. A Probabilistic Approach to Feature Selection: A Filter Solution // *Proceedings of the 13th International Conference on Machine Learning*. – San Francisco, CA: Morgan Kaufmann, 1996. – P. 319–327.
84. Luscombe N. M., Greenbaum D., Gerstein M. What is bioinformatics? A proposed definition and overview of the field // *Methods of Information in Medicine*. – 2001. Vol. 40, Issue 4. – P. 346–358. – Available from internet: http://www.ncbi.nlm.nih.gov/pubmed/11552348.
85. Lu Y., Han J. Cancer Classification using Gene Expression Data // *Information Systems*. – 2003. – Vol. 28. – P. 243–268.
86. Mapping the Sex Determination Locus in the Atlantic Halibut (Hippoglossus Hippoglossus) using RAD Sequencing / C. Palaiokostas, M. Bekaert, A. Davie, et al. // *BMC Genomics*. – 2013. – Vol. 14:566. – P. 1–12. – Available from Internet: http://www.biomedcentral.com/1471-2164/14/566.
87. Marghny M.H., El-Semman E. Extracting Fuzzy Classification Rules With Gene Expression Programming // *ICGST Conference on Artificial Intelligence and Machine Learning, AIML 05, Cairo, Egypt, 2005*. Serial Number: P1120535114.
88. Measuring Similarity by Prediction Class between Biomedical Datasets via Fuzzy Unordered Rule Induction / S. Fong, O. Mohammed, J. Fiaidhi, et al. // International Journal of Bio-Science and Bio-Technology. – 2014. – Vol. 6 (2). – P. 159–168. –Available from Internet: http://www.sersc.org/journals/IJBSBT/vol6_no2/16.pdf.
89. Meyer D. Support Vector Machines. The Interface to libsvm in package e1071. Online-Documentation of the package e1071 for R. – Wien: Technische Universität Wien, 2001. – P. 1–8. – Available from Internet: http://citeseerx.ist.psu.edu/viewdoc/download;jsessionid=BCDB6D08469CF19CF416EAADC044C6B3?doi=10.1.1.151.5271&rep=rep1&type=pdf.
90. Missing Value Estimation Methods for DNA Microarrays / O. Troyanskaya, M. Cantor, G. Sherlock, et al. // *Bioinformatics*. – 2001. – Vol.17, Issue 5. – P. 520–525.
91. Missing Value Imputation Improves Clustering and Interpretation of Gene

Expression Microarray Data / J. Tuikkala, L. L. Elo, O. S. Nevalainen, et al. // *BMC Bioinformatics*. – 2008. – Vol. 9:202. – P. 1–14. – Available from internet: http://www.biomedcentral.com/1471-2105/9/202.

92. Mehmet K., Reda A. Utilizing Genetic Algorithms to Optimize Membership Functions for Fuzzy Weighted Association Rules Mining // *Applied Intelligence*. – 2006. – Vol. 24:1. – P. 7–15.
93. Molecular Alterations in Primary Prostate Cancer after Androgen Ablation Therapy / C. J. Best, J. W. Gillespie, Y. Yajun, et al. // *Clinical Cancer Research*. – 2005. – Vol. 11 (19, Pt 1). – P. 6823–6834.
94. Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring / T. R. Golub, D. K. Slonim, P. Tamayo, et. al. // *Science*. – 1999. – Vol. 286. – P. 531–537.
95. Murata T., Ishibuchi H. Adjusting Membership Functions of Fuzzy Classification Rules by Genetic Algorithms // *Proceedings of International Joint Conference of the 4th IEEE International Conference on Fuzzy Systems and The 2nd International Fuzzy Engineering Symposium., Yokohama, Japan, 20–24 March, 1995*. Vol. 4. – IEEE Int., 1995 – P. 1819–1824.
96. Nakanishi H., Turksen I.B., Sugeno M. A Review and Comparison of Six Reasoning Methods // *Fuzzy Sets and Systems*. – 1993. – Vol. 57. – P. 257–294.
97. Nakashima T., Yokota Y., Ishibuchi H. Learning Fuzzy If-Then Rules for Pattern Classification with Weighted Training Patterns // *4th Conference of the European Society for Fuzzy Logic and Technology (EUSFLAT 2005) and the 11th Rencontres Francophones sur la Logique Floue et ses Applications (LFA 2005), September 7–9, 2005, Barcelona, Spain.* – Barcelona: Universitat Politecnica De Catalunya, 2005. – P. 1064–1069.
98. Nigles M., Linge J.P. *Bioinformatics* [Internet]. – France: Institut Pasteur, 2015. – Available from internet: http://www.pasteur.fr/recherche/unites/Binfs/definition/bioinformatics_definition.html.
99. Nojima Y., Kaisho Y., Ishibuchi H. Accuracy Improvement of Genetic Fuzzy Rule Selection with Candidate Rule Addition and Membership Tuning // *IEEE International Conference on Fuzzy Systems (FUZZ-IEEE'2010), 18-23 July, 2010, Barcelona, Spain*. – IEEE, 2010. – P. 1–8.
100. Novel Unsupervised Feature Filtering of Biological Data / R. Varshavsky, A. Gottlieb, M. Linial et al. // *Bioinformatics*. – 2006. – Vol. 22, Issue 14. – P. e507–e513.
101. Ohno–Machado L., Vinterbo S., Weber G. Classiication of Gene Expression Data Using Fuzzy Logic // *Journal of Intelligent & Fuzzy Systems*. – 2002. – Vol. 12. – P. 19–24.
102. Pelleg D., Moore A. X-means: Extending K-means with Efficient Estimation of the Number of Clusters // *Proceedings of the 17th International Conf. on Machine Learning, 2000*. – Morgan Kaufmann, 2000. – P. 727–734.
103. Pomeroy S. L. Gene Expression-Based Classification and Outcome Prediction of Central Nervous System Embryonal Tumors // *Nature*. – 2002. – Vol. 415. – P. 436–442.
104. Prediction of central nervous system embryonal tumour outcome based on gene expression / S. L. Pomeroy, P. Tamayo, M. Gaasenbeek, et al. // *Nature*. – 2002. – Vol. 415. – P. 436–442.
105. Prognostic Gene Expression Signatures can be Measured in Tissues Collected in RNAlater Preservative / D. Chowdary, J. Lathrop, J. Skelton, et al. // *J. Molecular Diagnostics*. 2006. – Vol. 8 (1). – P. 31–39.
106. Profiling and Functional Annotation of mRNA Gene Expression in Pediatric Rhabdomyosarcoma and Ewing's Sarcoma / C. Baer, M. Nees, S. Breit, et al. // *International Journal of Cancer*. – 2004. – Vol. 110 (5). – P. 687–694.
107. Quinlan J. R. *C4.5: Programs for Machine Learning*. – San Mateo, CA: Morgan Kaufmann Publishers, 1993. – 302 p.
108. Ressom H., Reynolds R., Varghese R. S. Increasing the Efficiency of Fuzzy Logic-based Gene Expression Data Analysis // *Physiological Genomics.* – 2003. – Vol. 13, No. 2. – P. 107–117. – Available from Internet: http://physiolgenomics.physiology.org/content/13/2/107.
109. Ross T. J. *Fuzzy Logic with Engineering Applications*. 3rd Edition. – Great Britain: John Wiley & Sons, 2010. – 606 p.

110. Rule Based Classifier for the Analysis of Gene-Gene and Gene-Environment Interactions in Genetic Association Studies / T. Lehr**,** J. Yuan**,** D. Zeumer**,** et al. // *BioData Mining*. – 2011. – Vol. 4:4. – P. 1–14.
111. Saeys Y., Inza I., Larranaga P. A Review of Feature Selection Techniques in Bioinformatics // *Bioinformatics*. – 2007. – Vol. 23, No. 19. – P. 2507–2517.
112. Schaefer G. Fuzzy Rule-Based Classification Systems and Their Application in the Medical Domain // *16th International Conference on Soft Computing MENDEL 2010, June 23–25, 2010, Brno, Czech Republic*. – Brno: Brno University of Technology, 2010. – P. 229–235.
113. Sensitivity and Specificity [Elektroniskais resurss]. – Tiešsaistes pakalpojums. – USA: Michigan State University, Office of Medical Education Research and Development, College of Human Medicine, 2008. – Pieejas veids: tīmeklis WWW. URL: http://omerad.msu.edu/ebm/Diagnosis/Diagnosis4.html. – Resurss aprakstīts 2015. g. 21. aprīlī.
114. Stanislawski J., Kotulska M., Unold O. Machine Learning Methods can Replace 3D Profile Method in Classification of Amyloidogenic Hexapeptides // *BMC Bioinformatics*. – 2013. – Vol. 14:21. – P. 1–9.
115. Tan P.-N., Steinbach M., Kumar V. *Introduction to Data Mining*. – Addison-Wesley, 2005. – 769 p.
116. The Identification of Complex Interactions in Epidemiology and Toxicology: A Simulation Study of Boosted Regression Trees / E. Lampa, L. Lind, M. Lind et al. // *Environmental Health*. – 2014. – Vol. 13:57. – P. 1–17.
117. Theron H., Cloete I. BEXA: A Covering Algorithm for Learning Propositional Concept Descriptions // *Machine Learning*. – 1996. – Vol. 24, Issue 1. – P. 5–40.
118. The WEKA Data Mining Software: An Update / M. Hall, E. Frank, G. Holmes, et al. // *ACM SIGKDD explorations newsletter*. – 2009. – Vol. 11, Issue 1. – P. 10–18.
119. Top 10 algorithms in data mining / W. Xindong, V. Kumar, J.R. Quinlan et al. // *Knowledge and Information Systems*. – 2008. – Vol. 14, No. 1. – P. 1–37.
120. Towards Missing Data Imputation: A Study of Fuzzy K-means Clustering Method / D. Li, J. Deogun, W. Spaulding, et al. // *Rough Sets and Current Trends in Computing, Proceedings of 4th International Conference, RSCTC 2004, Uppsala, Sweden, June 1–5, 2004*. – Berlin Heidelberg: Springer, 2004. – P. 573–579.
121. Treatment-Specific Changes in Gene Expression Discriminate In Vivo Drug Response in Human Leukemia Cells / M. H. Cheok, W. Yang, C. H. Pui, et al. // *Nature Genetics*. – 2003. – Vol. 34 (1). – P. 85–90.
122. van der Ploeg T., Austin P. C., Steyerberg E. W. Modern Modelling Techniques are Data Hungry: A Simulation Study for Predicting Dichotomous Endpoints // *BMC Medical Research Methodology*. – 2014. – Vol. 14:137. – P. 1–13.
123. van Zyl J., Cloete I. FuzzConRi – A Fuzzy Conjunctive Rule Inducer // *Proc.of the Workshop W8 on Advances in Inductive Rule Learning*, ECML/PKDD'2004, Pisa, September 20–24, 2004. – P. 194–203.
124. van Zyl J. *Fuzzy Set Covering as a New Paradigm for the Induction of Fuzzy Classification Rules*: PhD thesis. – Mannheim: University of Mannheim, 2007. – 263 p.
125. Vinterbo S., Kim E.-Y., Ohno–Machado L. Small, Fuzzy and Interpretable Gene Expression Based Classifiers // *Bioinformatics*. – 2005. – Vol. 21, No. 9. – P. 1964–1970.
126. Weitschek E., Fiscon G., Felici G. Supervised DNA Barcodes Species Classification: Analysis, Comparisons and Results // *Biodata Mining*. – 2014. – Vol. 7:4. – P. 1–18. – Available from Internet: http://www.biodatamining.org/content/7/1/4.
127. Witten I. H., Frank E., Hall M. A. *Data Mining: Practical Machine Learning Tools and Techniques*. 3rd Edition. – San Francisco: Morgan Kaufmann Publishers, 2011. – 664 p.
128. Woolf P. J., Wang Y. A Fuzzy Logic Approach to Analyzing Gene Expression Data // *Physiological Genomics*. – 2000. – Vol. 3. – P. 9–13.

129. Yasunobu S., Miyamoto S., Ihara H. A Fuzzy Control for Train Automatic Stop Control // *Trans. of the Society of Instrument and Control Engineers*. – 2002. – Vol. E-2, No. 1. – P. 1–9.

130. Yu L., Liu H. Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution // *Proceedings of the 20th International Conference on Machine Learning (ICML-2003), August 21-24, 2003*. – Washington DC: AAAI Press, Menlo Park, California, 2003. – P. 856–863.

131. Zadeh L.A. Fuzzy Sets // *Information and Control*. – 1965. – Vol. 8. – P. 338–353.

132. Zhang N., Lu W.F. An Efficient Data Preprocessing Method for Mining Customer Survey Data // *Industrial Informatics, 5th IEEE International Conference, 23-27 June, 2007.* – Vienna: IEEE, 2007. – P. 573–578.

133. Zhu T. Q., Xiong P. Optimization of Membership Functions in Anomaly Detection Based on Fuzzy Data Mining // *Proceedings of International Conference Machine Learning and Cybernetics (ICMLC 2005), 18-21 August, 2005, Guangzhou, China*. – Vol. 4. – IEEE, 2005. – P. 1987–1992.