RĪGAS TEHNISKĀ UNIVERSITĀTE

Ilze BIRZNIECE

INTERAKTĪVAS UZ INDUKTĪVO APMĀCĪBU BALSTĪTAS KLASIFIKĀCIJAS SISTĒMAS MODEĻA IZSTRĀDE

Promocijas darba kopsavilkums

Rīga 2013

RĪGAS TEHNISKĀ UNIVERSITĀTE
Datorzinātnes un informācijas tehnoloģijas fakultāte
Lietišķo datorsistēmu institūts

Ilze BIRZNIECE

Doktora studiju programmas „Datorsistēmas" doktorante

INTERAKTĪVAS UZ INDUKTĪVO APMĀCĪBU BALSTĪTAS KLASIFIKĀCIJAS SISTĒMAS MODEĻA IZSTRĀDE

Promocijas darba kopsavilkums

Zinātniskā vadītāja
Dr.sc.ing., profesore
M. KIRIKOVA

Rīga 2013

UDK 004.85(043.2)

Bi 818 i

Birzniece Ilze. Interaktīvas uz induktīvo apmācību balstītas klasifikācijas sistēmas modeļa izstrāde. Promocijas darba kopsavilkums. RTU, 2013. – 47 lpp.

Iespiests saskaņā ar DITF LDI padomes 2013. gada 19. marta lēmumu, protokols Nr. 83.

EIROPAS SAVIENĪBA

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

ISBN 978-9934-507-36-6

PROMOCIJAS DARBS IZVIRZĪTS
RĪGAS TEHNISKAJĀ UNIVERSITĀTĒ
INŽENIERZINĀTŅU DOKTORA GRĀDA IEGŪŠANAI

Promocijas darbs inženierzinātņu doktora grāda iegūšanai tiek publiski aizstāvēts 2013.gada 7. oktobrī Rīgas Tehniskās universitātes Datorzinātnes un informācijas tehnoloģijas fakultātē, Meža ielā 1/3, 202. auditorijā.

OFICIĀLIE RECENZENTI:

Profesors, Dr.habil.sc.comp. Arkādijs Borisovs

Rīgas Tehniskā universitāte, Latvija

Asoc. profesors, Dr.sc.comp. Jānis Zuters

Latvijas Universitāte, Latvija

Emeritētais profesors, PhD (lietišķā matemātika), HdR Jean-Hugues Chauchat

Lionas Limjēru universitāte 2 (*Université Lumiere Lyon 2*), Francija

APSTIPRINĀJUMS

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

Ilze Birzniece…….. (paraksts)

Datums: 21.06.2013.

Promocijas darbs ir uzrakstīts latviešu valodā un sastāv no ievada, 6 nodaļām, galveno rezultātu un secinājumu apkopojuma. Promocijas darbā ir 47 attēli, 34 tabulas un pamatteksts ir 160 lappuses. Bibliogrāfiskais saraksts satur 139 nosaukumus. Darbam pievienoti 11 pielikumi.

SATURS

IEVADS

Pieaugošais informācijas daudzums pasaulē rada nepieciešamību pēc datu apstrādes tehnikām, kas spētu samazināt cilvēka veiktās rutīnas aktivitātes. Šādas iespējas piedāvā mākslīgā intelekta nozare mašīnapmācība (ang. v. - *machine learning*). Mašīnapmācība sniedz datorprogrammai spēju apmācīties, balstoties uz pagātnes pieredzi, un uzlabot savu sniegumu [1]. Klasifikācija ir viens no mašīnapmācības uzdevumiem, kur klasifikators apgūst noteikt objekta klases piederību, balstoties uz iepriekš iegūtiem faktiem konkrētā problēmsfērā (ang. v. - *domain*). Ar jēdzienu „problēmsfēra" var apzīmēt jebkuru sistēmu vai darbības jomu. Nepieciešamība pēc dažādu objektu klasificēšanas ir sastopama daudzās sfērās, piemēram, medicīnas diagnostikā, kredītņēmēju vērtēšanā, attēlu apstrādē, mārketingā, dokumentu organizēšanā utt. [2].

Tēmas aktualitāte

Klasifikācijas algoritmi izmanto skaitliskus vai nominālus datus, kuri ir strukturēti. *Strukturēti dati* ir sadalīti nelielās, diskrētās vienībās. Tomēr daudzās jomās praktiski pieejamie dati ir tikai *daļēji strukturēti* vai *nestrukturēti*, un tos ir sarežģīti organizēt noteiktās struktūrās. Tas apgrūtina iespēju izmantot tradicionālās mašīnapmācības metodes automātiski un izslēdz no analīzes lielu daudzumu pieejamo datu. Par *automātisku klasifikāciju* promocijas darbā tiek saukts datorizēts klasifikācijas process, kurā no klasifikatora apmācības brīža (neskaitot apmācības datu sagatavošanu un klasifikācijas algoritma parametru iestatīšanu) līdz lēmuma pieņemšanai par jaunu objektu (jeb no apmācības puses skatoties - piemēru) klasifikāciju netiek iesaistīts sistēmas lietotājs vai eksperts.

Promocijas darbā īpaša uzmanība ir pievērsta vienai no cilvēku darbības sfērām, kam raksturīgi konceptuāli sarežģīti dati, proti, studiju priekšmetu salīdzināšanai augstākās izglītības jomā. Salīdzināt studiju programmas un atsevišķus priekšmetus ir laikietilpīgs darbs, ja to veic tikai manuāli. Studiju programmu un priekšmetu apraksti, kas tiek izmantoti priekšmetu atbilstības noteikšanai, parasti ir daļēji strukturētu tekstu veidā, kas var saturēt dažādi nosauktas un konceptuāli atšķirīgas sadaļas. Formalizēt un strukturēt šo informāciju traucē neskaidrās objektu attiecības un dabīgā valodā veidotie apraksti, kā arī mašīnapmācībai pieejamais relatīvi nelielais piemēru skaits (eksperta veiktie salīdzinājumi, kas sistēmai kalpo par pieredzi, no kuras izdarīt secinājumus). Turklāt jārēķinās ar apstākli, ka viens studiju priekšmets var atbilst vairākiem citiem, tātad nepieciešams nodrošināt iespēju objektus klasificēt vienlaicīgi vairākās klasēs.

Ja kādā problēmsfērā automātiska klasifikācija sniedz neapmierinošus klasifikācijas rezultātus, savukārt ir pieejams eksperts, kurš var veikt klasifikāciju šīs sfēras ietvaros, tad daļēji automātiskas jeb interaktīvas klasifikācijas izmantošana ļauj izmantot gan automātiskas, gan eksperta veiktas klasifikācijas priekšrocības un iegūt kompromisu starp eksperta ieguldītā darba apjomu un klasifikācijas rezultātu kvalitāti. Promocijas darbs ir veltīts *automatizēta* jeb *daļēji automātiska* klasifikācijas risinājuma izveidei, kas izmanto gan mašīnapmācības sniegtās iespējas, gan interaktīvu sadarbību ar jomas ekspertu klasifikatora lietošanas laikā, ja klasifikators sastopas ar objektu, kura klasifikācija tam ir neskaidra. *Klasifikatoram neskaidrs objekts* jeb *neskaidra klasifikācija* (ang. v. – *uncertain classification*) šī darba kontekstā ietver gan *neklasificētus* (ang. v. – *unclassified*), gan *nepārliecinoši klasificētus* (ang. v. – *low confidence of classification*) objektus. Šie termini sīkāk ir apskatīti darba 3. un 4. nodaļā. Būtisks parametrs, kas nosaka, vai klasifikācijas rezultāti kādā sfērā ir uzticami un praktiskai lietošanai pieņemami, ir nepareizi klasificēto objektu skaits, tādēļ tas arī izvēlēts par mēru klasifikācijas rezultātu novērtēšanai.

Promocijas darba mērķis

Daba mērķis ir izstrādāt automatizētas klasifikācijas sistēmas modeli, kas pieļauj interaktivitāti ar ekspertu klasifikatora lietošanas laikā, ja klasifikators sastopas ar objektu, ko tas nespēj klasificēt vai nav pārliecināts par sava lēmuma pareizību.

Darba uzdevumi

Promocijas darba mērķa sasniegšanai ir izvirzīti šādi uzdevumi:

- Veikt izglītības dokumentu datorizētas salīdzināšanas risinājumu analīzi un identificēt risināmos uzdevumus.
- Veikt klasifikācijas uzdevuma mašīnapmācībā izpēti.
- Veikt esošo interaktīvo klasifikācijas risinājumu izpēti.
- Veikt klasifikācijas sistēmu arhitektūru analīzi interaktīvas klasifikācijas sistēmas izstrādei.
- Izstrādāt interaktīvas klasifikācijas sistēmas modeli, kas apvieno interaktīvas klasifikācijas sistēmas radīšanai nepieciešamās komponentes (algoritmus, metodes, pieejas un arhitektūras).
- Izstrādāt interaktīvas klasifikācijas sistēmas modeļa papildinājumu, kas apvieno interaktīvas daudzkategoriju klasifikācijas sistēmas radīšanai nepieciešamās komponentes.

- Realizēt interaktīvas klasifikācijas sistēmas prototipu, kas ievieš izstrādāto modeli.
- Pārbaudīt izstrādātā modeļa lietderību un prototipa lietojamību, veicot praktiskus eksperimentus.

Pētījuma objekts

Promocijas darba pētījumu objekts ir klasifikācijas uzdevums mašīnapmācībā.

Pētījuma priekšmets

Promocijas darba pētījuma priekšmets ir klasifikācijas sistēmas lietotāja - jomas eksperta iesaistīšana klasifikācijas rezultātu uzlabošanā.

Pieņēmumi un ierobežojumi

Izstrādājamā interaktīvā klasifikācijas sistēma ir paredzēta situācijām, kurās ir spēkā šādi *nosacījumi*:

- klasifikācijai izmantojamie dati ir ekspertam saprotami:
  - pēc savas būtības un struktūras;
  - pēc to apjoma (objekta apraksta apjoms nav *pārāk* liels).
- cilvēks-eksperts ir pieejams.

Tātad datiem, ar kuriem strādās klasifikācijas sistēma, ir jābūt tādiem, kurus eksperts spēj interpretēt. Tas nozīmē, ka ekspertam ir jāsaprot sākotnējo vai apstrādājamo datu jēga, lai viņš varētu klasificēt jaunu objektu, ja sistēma to lūgs. Cits aspekts ir viena objekta aprakstīšanai izmantoto datu apjoms. Ja tas ir pārāk liels, tad eksperts nespēs sniegtos datus operatīvi saprast un, attiecīgi, arī klasificēt. Turklāt ekspertam, kas risināmajā jomā ir spējīgs sniegt sistēmai padomu, vispār ir jābūt pieejamam. Ja šāda eksperta nav, vai jau iepriekš zināms, ka eksperta nodarbināšana ir pārāk dārga, tad interaktīvam risinājumam nav jēgas. Tātad tiek pieņemts, ka interaktīva klasifikācijas sistēma tiek lietota sfērā, kur augstāk minētos nosacījumus ir iespējams ievērot.

No klasifikācijas pieejām darbā ir izvēlēta induktīvā apmācība (lēmumu kokus un likumus veidojošie algoritmi), jo tās sniegtie rezultāti ir lietotājam vislabāk saprotami, kas ir būtiski, ja nepieciešams vairot lietotāja uzticību klasifikācijas sistēmai.

Promocijas darbā tiek apskatītas problēmsfēras, kurām ir raksturīga objektu daudzkategoriju piederība, tas ir, objekts dabiski var vienlaicīgi piederēt vairākām klasēm. Šāda situācija ir, piemēram, ziņu organizēšanā, jo viens raksts var atbilst vairākām tēmām, piemēram, sports un tūrisms. Šis nosacījums ir radies tādēļ, ka daudzkategoriju klasifikācija ir raksturīga primārajā izskatāmajā problēmsfērā - studiju priekšmetu salīdzināšanā.

Lietojot terminus „sistēmas lietotājs" un „eksperts" darbā ir saglabāts princips, ka, lai strādātu ar klasifikācijas sistēmu un caurskatītu iegūtos rezultātus, sistēmas lietotājam nav jābūt problēmsfēras ekspertam, taču, lai sniegtu savas zināšanas un pilnveidotu klasifikatoru, ir nepieciešama padziļināta izpratne par attiecīgo jomu. Tādēļ cilvēks vienmēr var būt klasifikācijas sistēmas lietotājs, bet ne vienmēr ir arī eksperts attiecīgajā jomā.

Darbā tiek izvirzītas šādas **aizstāvamās tēzes.**

T1 Klasifikācijas sistēma, kas realizē interaktīvas klasifikācijas sistēmas modeli eksperta iesaistīšanai klasifikatora lietošanas posmā, ļauj samazināt nepareizi klasificēto objektu skaitu, salīdzinot ar automātisku klasifikāciju.

T2 *Piemērotākā pārliecības sliekšņa lieluma noteikšanas metode* palīdz atrast klasifikatora pārliecības slieksni, ar kuru nepareizi klasificēto piemēru skaits *N* ir minimāls pie izvirzītajiem eksperta ieguldāmā darba ierobežojumiem.

T3 Universitāšu studiju priekšmetu salīdzināšanā ir lietderīgi izmantot *uz induktīvo apmācību balstītu, interaktīvu, daudzkategoriju klasifikācijas sistēmu*.

Zinātniskais jaunieguvums

- Izstrādāts interaktīvas klasifikācijas sistēmas *InClaS* (ang. v. - ***In****teractive Inductive Learning based* ***Cla****ssification* ***S****ystem*) modelis, kas apvieno interaktīvas klasifikācijas sistēmas radīšanai nepieciešamās komponentes.
- Izstrādāts *InClaS* modeļa papildinājums, kas apvieno interaktīvas daudzkategoriju klasifikācijas sistēmas radīšanai nepieciešamās komponentes.

Teorētiskā vērtība

Zinātniskais jaunieguvums ietver vairākus pakārtotus **teorētiskos rezultātus:**

- Izstrādāta klasifikācijas sistēmā ieviešamā interaktivitātes shēma.
- Izstrādāta interaktīvas klasifikācijas sistēmas uzbūve – klasifikācijas sistēmas funkcionālie moduļi, to īpašības un sasaistes.
- Izstrādātas divas klasifikatora atjaunošanas (papildināšanas) shēmas pēc eksperta veiktas klasifikācijas.
- Izstrādāts algoritms klasifikatoram neskaidru piemēru noteikšanai daudzkategoriju klasifikācijas gadījumā.
- Izstrādāta metode atbilstošākā pārliecības sliekšņa noteikšanai, pie kura klasifikatora klasificētos piemērus atzīt par nepārliecinoši klasificētiem un nodot eksperta pārziņā.
- Veikts interaktīvas daudzkategoriju klasifikācijas sistēmas projektējums sistēmu veidojošo moduļu, to ieeju un izeju apraksta veidā.

- Veikta literatūras analīze un iegūti sistematizēti apkopojumi par klasifikācijas un izglītības dokumentu salīdzināšanas tēmām.

Praktiskā nozīmība

Ir izstrādāts interaktīvas klasifikācijas sistēmas *InClaS* prototips daudzkategoriju klasifikācijas uzdevumiem. Prototips ir īpaši pielāgots studiju priekšmetu salīdzināšanai lietotājam ērtākas saskarnes nodrošināšanai.

Kā papildu rezultāts darba gaitā ir izveidota programma daudzkategoriju datu pārveidošanai dažādos atspoguļojuma formātos (atbilstoši atšķirīgajām ieejas datu formāta prasībām programmatūrā *Weka* [3] un bibliotēkā *Mulan* [4]).

Darba aprobācija

Par darba rezultātiem ir ziņots 12 konferencēs:

- 2012. gada 18.-23. novembrī. *The Fifth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services* (CENTRIC 2012). Lisabona, Portugāle.
- 2012. gada 16.-20. jūlijā. *International Conference on Machine Learning and Data Mining* (MLDM 2012). Berlīne, Vācija.
- 2012. gada 16.-18. maijā. *Sixth International IEEE Conference on Research Challenges in Information Science* (RCIS 2012). Valensija, Spānija.
- 2011. gada 6.-8. oktobrī. *10th International Conference on Perspectives in Business Informatics Research* (BIR 2011). Rīga, Latvija.
- 2011. gada 13.-16. oktobrī. RTU 52. starptautiskā zinātniskā konference, sekcija „Datorzinātne". Rīga, Latvija.
- 2011. gada 24.-26. jūlijā. *Intelligent Systems and Agents* (ISA 2011). Roma, Itālija.
- 2011. gada 7.-10. martā. *Rethinking Education in the Knowledge Society* (RED 2011). Monte Verita, Šveice.
- 2010. gada 11.-15. oktobrī. RTU 51. starptautiskā zinātniskā konference, sekcija „Datorzinātne". Rīga, Latvija.
- 2010. gada 5.-7. jūlijā. *Ninth International Baltic Conference on Databases and Information Systems* **(**Baltic DB&IS 2010). Rīga, Latvija.
- 2010. gada 27.-28. maijā. *19th Annual Machine Learning Conference of Belgium and The Netherlands* (BeneLearn 2010). Leuvena, Beļģija.

- 2010. gada 22.-23. aprīlī. *16th International Conference on Information and Software Technologies* (IT 2010). Kauņa, Lietuva.
- 2009. gada 12.-16. oktobrī. RTU 50. starptautiskā zinātniskā konference, sekcija „Datorzinātne". Rīga, Latvija.

Promocijas darba ietvaros veikto pētījumu rezultāti ir atspoguļoti 13 publikācijās starptautiskos zinātniskos izdevumos:

- Birzniece I. Architecture of an Interactive Classification System // The Fifth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services, 2012. IARIA, - 91.-100. lpp. Iekļauts ThinkMind Digital Library datu bāzē.
- Birzniece I. Machine Learning Approach for Study Course Comparison // International Conference on Machine Learning and Data Mining, 2012. IBai publishing - 1.- 13. lpp. **Iegūts atzinības raksts par labāko stenda referātu.**
- Birzniece I. Interactive Use of Inductive Approach for Analyzing and Developing Conceptual Structures // Sixth International Conference on Research Challenges in Information Science: Conference Proceedings, 2012. IEEE - 129.-134. lpp. Iekļauts **Scopus**, IEEE Xplore, DBLP datu bāzēs.
- Birzniece I. Interactive Inductive Learning Based Study Course Comparison // Proceedings of the Red-Conference: Rethinking Education in the Knowledge Society, 2011. Università della Svizzera italiana - 339.-347. lpp.
- Birzniece I., Kirikova M. Interactive Inductive Learning: Application in Domain of Education // RTU zinātniskie raksti. 5. sēr., Datorzinātne. - 47. sēj., 2011. RTU izdevniecība - 57.-64. lpp. Iekļauts VERSITA, DBLP, EBSCO datu bāzēs.
- Birzniece I. Artificial Intelligence in Knowledge Management: Overview and Trends // RTU zinātniskie raksti. 5. sēr., Datorzinātne. - 46. sēj., 2011. RTU izdevniecība - 5.-11. lpp. Iekļauts VERSITA, DBLP, EBSCO, io-port.net datu bāzēs.
- Birzniece I. Interactive Inductive Learning Based Classification System // Proceedings of the IADIS International Conference Intelligent Systems and Agents, 2011. IADIS - 112.-116. lpp. Iekļauts **Scopus** datu bāzē.
- Birzniece I., Rudzājs P. Machine Learning Based Study Course Comparison // IADIS Conference on Intelligent Systems and Agents, 2011. IADIS - 107.-111. lpp. Iekļauts **Scopus** datu bāzē.

- Birzniece I. The Use of Inductive Learning in Information Systems // Proceedings of 16th International Conference on Information and Software Technologies, 2010. Technologija Kaunas - 95.-101. lpp. Iekļauts **Thomson Reuters Web of Science** datu bāzē.
- Birzniece I. From Inductive Learning Towards Interactive Inductive Learning // RTU zinātniskie raksti. 5. sēr., Datorzinātne. - 43. sēj., 2010. RTU izdevniecība - 106.-112. lpp. Iekļauts VERSITA, DBLP, io-port.net, EBSCO datu bāzēs.
- Birzniece I. Interactive Inductive Learning System // Frontiers of AI and Applications. Databases and Information Systems VI, Vol. 224, Selected Papers of Baltic DB&IS, 2011. IOS Press - 380.-393. lpp. Iekļauts ACM, DBLP, io-port.net datu bāzēs.
- Birzniece I., Kirikova M. Interactive Inductive Learning Service for Indirect Analysis of Study Subject Compatibility // Proceedings of the BeneLearn, 2010. Katholieke Universiteit Leuven - 1.-6. lpp.
- Birzniece I. Interactive Inductive Learning System: The Proposal // Proceedings of the Ninth International Baltic Conference on Databases and Information Systems, 2010. University of Latvia Press - 245.-260. lpp.

Darba struktūra

Darbs sastāv no ievada, 6 nodaļām, rezultātu un secinājumu apkopojuma, literatūras avotu saraksta un pielikumiem. Ievadā ir pamatota risināmā problēma, definēts darba mērķis, uzdevumi un aizstāvamās tēzes. Tāpat ievadā ir izklāstīts arī darba izpildes process, galvenie rezultāti un promocijas darba saturs.

Pirmā nodaļa "Pētījuma pamatojums" apraksta līdzšinējos centienus datorizētā studiju programmu un priekšmetu salīdzināšanā, izvēršot aktuālās problēmas un definējot šīs disertācijas ietvaros risināmos uzdevumus un darba mērogu.

Otrajā nodaļā "Saistīto darbu analīze: iestrādnes un pastāvošās problēmas" atspoguļoti un apkopoti līdzšinējie pētījumi par darbam aktuālām tēmām. 2.1. apakšnodaļa apskata klasifikācijas uzdevumu mašīnapmācībā, galveno uzmanību pievēršot induktīvās apmācības problēmu aprakstam, daudzkategoriju klasifikācijai, tās novērtējuma mēriem, klasifikācijā risināmajām problēmām, piemēram, nespējai klasificēt jaunus piemērus. 2.2. apakšnodaļa veltīta interaktīvajām pieejām klasifikācijā, bet 2.3. apakšnodaļa apkopo līdzšinējo veikumu klasifikācijas sistēmu arhitektūru izstrādē.

Trešā nodaļa "Interaktīvas uz induktīva apmācību balstītas klasifikācijas sistēmas *InClaS*) pamatmodelis" sniedz izstrādāto interaktīvās klasifikācijas sistēmas modeļa galveno

komponenšu izklāstu – ekspertam vaicājamo piemēru noteikšanu, eksperta sniegto zināšanu iekļaušanu klasifikatorā un sistēmas uzbūvi.

Ceturtā nodaļa "*InClaS* modelis daudzkategoriju klasifikācijas uzdevumam" paplašina sistēmas pamatmodeli ar daudzkategoriju klasifikācijai nepieciešamajām komponentēm, definējot algoritmu klasifikatoram neskaidro piemēru noteikšanai un metodi piemērotākā pārliecības sliekšņa atrašanai. Šeit ir paskaidroti ar sistēmas projektēšanu un realizācijas detaļām saistītie lēmumi studiju priekšmetu salīdzināšanas uzdevumam.

Piektajā nodaļā "*InClaS* prototips" apkopotas izstrādātās modeļa komponentes un aprakstīta to realizācija sistēmas prototipa veidā. Nodaļā paskaidrotas izstrādātā *InClaS* prototipa atšķirības no citiem klasifikācijā izmantotiem rīkiem un sniegts ieskats prototipa funkcijās un lietotāja saskarnē.

Sestā nodaļa "*InClaS* modeļa novērtējums" apraksta eksperimentu plānu un iegūtos rezultātus izglītības un medicīnas jomās, salīdzinot promocijas darbā piedāvātās interaktīvās un klasiskās automātiskās klasifikācijas sniegtos rezultātus un pārbaudot izstrādātā *InClaS* modeļa lietderību.

Darbu noslēdz "Galvenie rezultāti un secinājumi", kas sniedz darba teorētisko un praktisko rezultātu, iegūto atziņu un turpināmo darbu aprakstu.

Darbam ir 11 pielikumi: 1. - svarīgāko darbā lietoto terminu skaidrojumu saraksts; 2. - priekšapstrādes apraksts studiju priekšmetu pilno aprakstu izmantošanai klasifikācijā; 3. - forma, ar kuras palīdzību tiek iegūti dati no eksperta studiju priekšmetu salīdzināšanai ar Eiropas e-kompetenču ietvara starpniecību; 4. - studiju priekšmetu formālo aprakstu iegūšanas demonstrējums tiešajā un netiešajā salīdzināšanā; 5. - rezultāti, kas iegūstami ar utilītprogrammu datņu pārveidošanai starp dažādiem daudzkategoriju atspoguļošanas formātiem; 6. - induktīvās apmācības algoritmu iedalījums; 7. - klasifikācijas sistēmu projektēšanas pieeju apkopojums; 8. - klasifikācijas sistēmu uzbūves modeļu apkopojums; 9. - promocijas darba eksperimentos izmantoto klasifikācijas metožu un algoritmu nosaukumu saīsinājumu skaidrojums; 10. - klasifikācijas modeļu atspoguļojuma formāti dažādiem algoritmiem; 11. - pilns eksperimentu rezultātu atspoguļojums piemērotākā sliekšņa lieluma noteikšanai studiju priekšmetu salīdzināšanas uzdevumā.

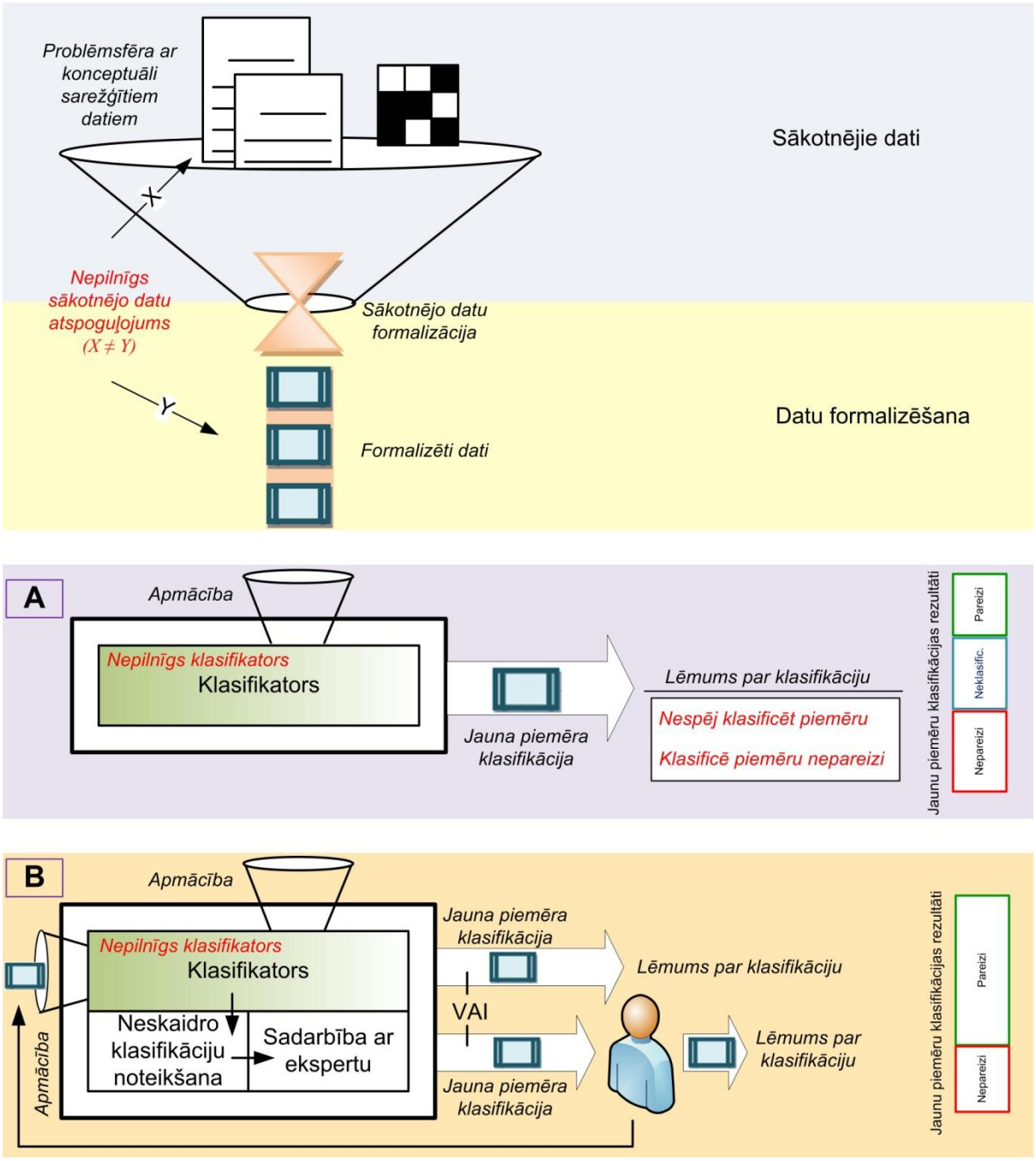
1. PĒTĪJUMA PAMATOJUMS

Darba pirmajā nodaļā ir izklāstīta problemātika, kas pamato promocijas darbā risināmo uzdevumu un nosaka nepieciešamos pētījumus un izstrādnes. Vispirms izskaidroti automātiskās klasifikācijas ierobežojumi lietošanai sfērās, kuras ir grūti formāli definējamas un nav aprakstītas ar pilnīgu piemēru kopu. Kā nozīmīga darbības sfēra, kurā parādās nepieciešamība pēc automatizēta, uz induktīvo apmācību balstīta klasifikācijas risinājuma, analizēta izglītības joma, kur ir aktuāla dažādu studiju priekšmetu savstarpējas atbilstības noteikšana. Pamatojoties uz šo problēmsfēru, izvirzītas prasības pret izstrādājamo risinājumu.

1.1. Automātiskas klasifikācijas ierobežojumi

Mašīnapmācības pieejas saskaras ar izaicinājumu jomās, kurām būtu lietderīgi izmanot mašīnapmācību tādēļ, ka tajās pastāv laikietilpīgas cilvēka veiktas aktivitātes, bet kuras neatbilst tipiskiem parametriem mašīnapmācības piemērošanai datu apjoma un strukturētības ziņā. Ja sākotnējie dati ir daļēji strukturētā vai citādi nepilnīgi izmantojamā formā, tad pārveidojot pieejamos datus no sākotnēji daļēji strukturēta vai nestrukturēta formāta uz klasifikācijas algoritmiem lietojamu – strukturētu formātu, var pazaudēt daļu būtiskas informācijas vai atspoguļot to neprecīzi, tādējādi iegūstot apmācības kopu, kas nepilnīgi atspoguļo pētāmo problēmsfēru. Ja turklāt šo datu ir maz, tad klasifikatoram nepietiek avotu, kur smelties pieredzi. Pārveidotos (formalizētos) datus tālāk izmantojot mašīnapmācībai un klasifikatora iegūšanai, pastāv iespēja, ka arī klasifikators būs nepilnīgs un nespēs noteikt klasi jauniem piemēriem vai noteiks to nepareizi. Šāda situācija ir atspoguļota 1.1. attēlā, ja klasifikācijai izmanto tradicionālo automātisko sistēmu (attēla A daļa). Paskaidrojot atšķirību starp terminiem „klasifikators” un „klasifikācijas sistēma”, jāmin, ka klasifikators ir klasifikācijas modelis (piemēram, likumu kopa), kas kalpo jaunu piemēru klases (klašu) noteikšanai konkrētā problēmsfērā, savukārt klasifikācijas sistēma ir programmatūra, kas ietver klasifikatoru, lietotāja saskarni un citas saistītās komponentes, piemēram, pirms un pēcapstrādi. Klasifikācijas sistēmas izveide ietver arī klasifikatora veidošanu. Principā klasifikators var būt arī indukcijas rezultātā iegūts likumu saraksts uz papīra, pēc kura vadīties jauna piemēra klasifikācijas gaitā. Klasifikācijas sistēma ir klasifikators un tā perifērija, kas kopumā nodrošina datorizētu klasifikācijas procesu. Tātad klasifikators ir daļa no klasifikācijas sistēmas, kura nodrošina klasifikācijas veikšanu. Lai risinātu problēmas jaunu piemēru klasifikācijā nepilnīga klasifikatora gadījumā, novērstu

neklasificētus piemērus un samazinātu nepareizi klasificēto piemēru skaitu, promocijas darbā ir izstrādāta interaktīva klasifikācijas pieeja.



1.1. att. Automātiska (A) vai interaktīva (B) klasifikācija nepilnīgu apmācības datu gadījumā

Ja automātisko klasifikācijas sistēmu (1.1. attēla A daļu) aizvieto ar daļēji automātisku (automatizētu) sistēmu (1.1. attēla B daļu), kurā klasifikators ir papildināts ar elementiem (1) neskaidro klasifikāciju noteikšanai un (2) saziņai ar ārēju ekspertu, tad jauno piemēru klasifikāciju veiktu pati sistēma vai, ja tā nav droša par savu spēju pieņemt lēmumu, jomas eksperts, tādējādi likvidējot neklasificētus piemērus un samazinot nepareizi klasificēto

piemēru skaitu. Klasifikators arī iegūtu papildu apmācības iespējas, jo eksperta klasificētie piemēri var kalpot par jaunu pieredzi.

Lai pārliecinātos par daļēji automātiskas klasifikācijas pieejas nepieciešamību, darba tālākā gaitā ir izpētīti līdzšinējie datorizētie klasifikācijas risinājumi ar augstāko izglītību saistītās sfērās.

1.2. Klasifikācijas uzdevumi izglītības jomā

Nepieciešamība atbalstīt dažādu izglītības dokumentu salīdzināšanas procesu ar datoru palīdzību ir pamatota literatūrā un izteikta no augstākās izglītības jomā strādājošo puses. Ar terminu *izglītības dokumenti* darbā tiek apzīmēti dažāda veida materiāli, kas raksturo izglītības saturu un novērtējumu, tai skaitā studiju priekšmetu apraksti, mācību materiāli, izglītību apliecinošie diplomi u.c. dokumenti. Augstākās izglītības dokumentu salīdzināšanas nepieciešamība parādās vairākās formās un ir nosacīti iedalāma trīs kategorijās [5-12].

- Studiju priekšmetu salīdzināšana apmaiņas programmām, tālākizglītībai, izglītības dokumentu pielīdzināšanai u.c. vajadzībām. Galvenokārt tiek izmantoti dokumenti, kas apraksta studiju programmu un priekšmetu saturu un apliecina iegūto izglītību.
- Studiju programmu izstrāde, kas iesaista salīdzinājuma veikšanu ar citām līdzīgām izglītības programmām. Galvenokārt tiek izmantoti dokumenti, kas apraksta studiju programmu un priekšmetu saturu.
- Mācību materiālu dalīšana kategorijās, piemēram, e-apmācības sistēmās, lai interesentam tiktu piedāvāti atbilstoši mācību materiāli.

Kā apstiprina [8], izglītības dokumentu salīdzināšana ir sarežģīts uzdevums gan cilvēkiem, gan datorsistēmām, tāpēc šī procesa automatizācija prasa specifiskas pieejas un eksperta līdzdalību. Daļēji strukturētais dokumentu formāts prasa dažādu informācijas izguves metožu lietošanu. Elektroniskās akadēmiskās padomdošanas sistēmas autori [8] norāda, ka sistēmas rezultātus noteikti varētu uzlabot, palielinot apmācības kopu, kā arī iesaistot ekspertu. Piemēram, ieviešot vienkāršu mehānismu manuālai izgūto datu pārbaudei un papildināšanai, būtu iespējams uzlabot tālākā darbībā izmantotās informācijas kvalitāti. Klasifikācijai izmantojamās datu kopas palielināšana arī prasa cilvēka darbu, jo sagatavot apmācības piemērus var tikai problēmsfēras eksperts, līdz ar to pieejamās pieredzes daudzums vienmēr būs ierobežots, un procesa uzlabošanas galvenais ierobežojums ir darba apjoms, ko eksperts var ieguldīt. Līdzšinējo darbu analīze apliecina, ka izglītības dokumentu salīdzināšanā, lai iegūtu uzticamu rezultātu, risinājumam ir jābūt automatizētam, bet ne automātiskam. Ir vērts pievērst uzmanību arī faktam, ka, neskatoties uz to, ka starp dažādām

studiju programmām nepastāv viennozīmīga studiju priekšmetu savstarpējā atbilstība, ko apstiprina arī pētītās literatūras autori [9, 10], studiju priekšmetu salīdzināšana ir veikta, meklējot 1 pret 1 atbilstību. Nevienā no realizētajām sistēmām līdz šim nav apskatītas iespējas studiju priekšmetu vienā studiju programmā attiecināt uz vairākiem studiju priekšmetiem citā programmā jeb izmantot daudzkategoriju (ang. v. - *multi-label*) klasifikāciju. Lai arī ir uzsvērta cilvēka iesaistīšanās nepieciešamība, līdzšinējos risinājumos izmantotās mašīnapmācības metodes neveicina iesaistītās personas izpratnes pilnveidošanu un atgriezeniskās saites uzlabošanu.

Līdzšinējo darbu analīze norāda nepieciešamos uzlabojumu virzienus un pastāvošos ierobežojumus, lai izglītības dokumentu salīdzināšanā varētu ieviest zināmu automatizāciju. Izdarot secinājumus par šīs sfēras problemātiku, promocijas darbā ir izskatīts uzdevums par *uz induktīvo apmācību balstītu, interaktīvu un daudzkategoriju klasifikācijas sistēmas izstrādi studiju priekšmetu salīdzināšanas atbalstam*.

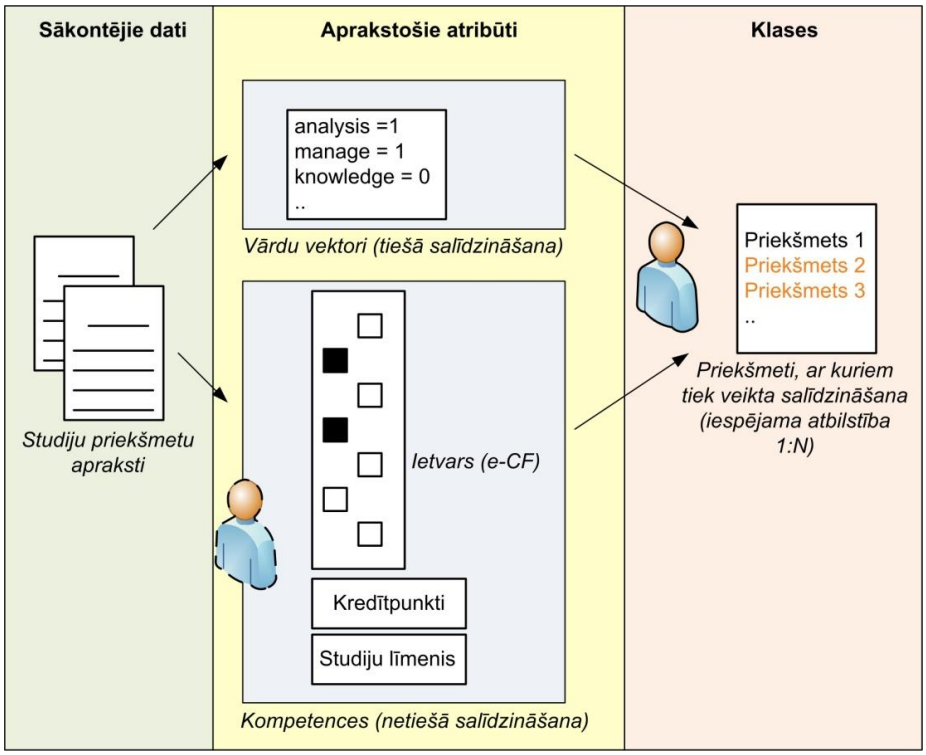
Formalizējot studiju priekšmetu salīdzināšanas uzdevumu kā prasības pret mašīnapmācības risinājumu, ir definētas šādas ***raksturīgās īpašības***:

- iegūtie rezultāti ir jāsaprot klasifikatora lietotājam un ekspertam;
- pieejamā apmācības kopa ir maza;
- sākotnējie dati ir daļēji strukturēti vai nestrukturēti;
- problēmsfērā ir raksturīgas daudzas klases, kuras sastopamas vienlīdz bieži;
- objekts var piederēt vienlaicīgi vairākām klasēm.

1.3. Izglītības jomas uzdevuma interpretācija mašīnapmācības kontekstā

Šī apakšnodaļa sniedz skaidrojumu promocijas darbā risinātajai augstākās izglītības studiju priekšmetu salīdzināšanas problēmai no mašīnapmācības puses. Tajā ir paskaidrotas daļēji strukturēto studiju priekšmetu aprakstu transformēšanas metodes klasifikācijas algoritmiem izmantojama datu formāta iegūšanai. Tas nozīmē strukturētu priekšmetus aprakstošo atribūtu izgūšanu un klašu jeb kategoriju, kurās klasificējamie priekšmeti jāiedala, definēšanu. Apmācības datu iegūšana un klašu definēšana shematiski parādīta 1.2. attēlā. Sākotnējie dati ir studiju priekšmetu apraksti tādā formā, kādā izglītības institūcija tos ir publicējusi. Lai iegūtu uz atribūtu-vērtību pāriem balstītu atspoguļojumu, ko izmantot induktīvās apmācības procesā, šāds apraksta veids ir jāformalizē, nosakot klases, definējot un izgūstot raksturīgos atribūtus un to vērtības. Priekšmetu apraksta formalizācija eksperimentu nolūkos tiek veikta divos neatkarīgos veidos – (1) kā tekstu salīdzināšana (iegūstot vārdu vektorus angļu valodā) un (2) kā kompetenču salīdzināšana (atspoguļojot sasniedzamos

mācību rezultātus pret vienotu standartizētu ietvaru). Priekšmetu klasifikācija, izmantojot vārdu vektoru iegūšanu no priekšmetu aprakstiem, tiek saukta par tiešo salīdzināšanu, jo, lai iegūtu aprakstā izmantotos vārdus, tiek veikta tikai sintaktiska sākotnējo datu apstrāde. Vārdu vektors satur tos vārdus, kas sastopami apmācības datu kopā, un konkrētam studiju priekšmetam tiek norādīts, kuri no vārdiem šī priekšmeta aprakstā ir iekļauti. Savukārt netiešajā salīdzināšanā no studiju priekšmetu aprakstiem eksperts izsecina iegūstamās kompetences, atspoguļojot tās vienotā ietvarā (izvēlēts Eiropas e-kompetenču ietvars (*e-CF*) [13]), tādējādi sākotnējie dati tiek pārveidoti arī semantiski. Kā galvenie klasifikācijas atribūti tiek izmantotas konkrētas kompetences (to esamība vai neesamība), un priekšmeti tiek salīdzināti pastarpināti, par starpslāni izmantojot kompetenču ietvaru. Balstoties uz veikto literatūras analīzi, tiek izteikts pieņēmums, ka priekšmetu tiešā salīdzināšana sniegs sliktākus klasifikācijas rezultātus kā netiešā; to ir paredzēts pārbaudīt eksperimentāli.



1.2. att. Priekšmetu aprakstu formalizēšanas koncepcija

Klases tiek izvēlētas atbilstoši salīdzināšanas mērķim; piemēram, klases var būt studiju priekšmeti no studiju programmas, attiecībā pret kuru ir nepieciešams noskaidrot citu studiju priekšmetu atbilstību. Iepazīstoties ar jauna studiju priekšmeta aprakstu, piešķiramās klases šim priekšmetam nosaka eksperts, turklāt viņš var piešķirt vairākas klases vienam priekšmetam, jo neviennozīmīgās priekšmetu atbilstības dēļ viens priekšmets savā saturā var pārklāties ar vairākiem citiem. Sīkāka analīze studiju priekšmetu salīdzināšanas problēmas risinājumam ir sniegta saistībā ar praktisko eksperimentu plānu darba 6. nodaļā.

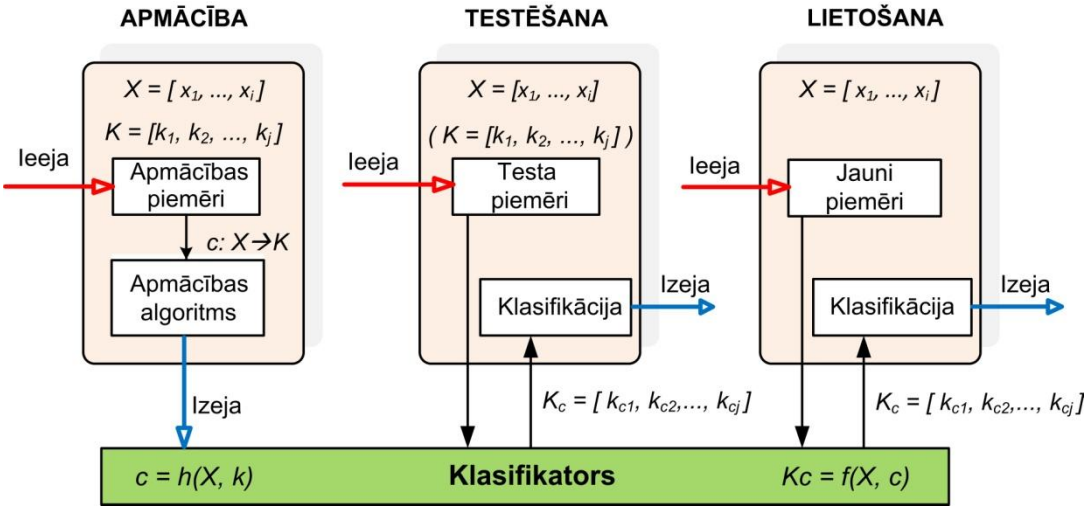
2. SAISTĪTO DARBU ANALĪZE: IESTRĀDNES UN PASTĀVOŠĀS PROBLĒMAS

Otrā nodaļa apkopo līdzšinējos pētījumus un teorētisko bāzi dažādos klasifikācijas aspektos. 2.1. apakšnodaļā ir apskatīts automātiskās klasifikācijas uzdevums, sevišķu uzmanību pievēršot induktīvās apmācības pieejai, daudzkategoriju klasifikācijai un tās novērtēšanas mēriem, kā arī problēmām jaunu piemēru klasificēšanā. 2.2. apakšnodaļa ir veltīta dažāda veida līdzšinējiem interaktīviem risinājumiem induktīvajā apmācībā un citās klasifikācijas pieejās. 2.3. apakšnodaļa sniedz klasifikācijas sistēmu arhitektūru apkopojumu, lai interaktīvas klasifikācijas sistēmas izveidē izmantotu jau esošus un pārbaudītus risinājumus, ja tādi ir, un pārņemtu labo praksi.

2.1. Klasifikācijas uzdevums mašīnapmācībā

Klasifikācija ir būtiska daudzos problēmu risināšanas uzdevumos. Lai veiktu klasifikāciju, ir nepieciešams realizēt kāda veida spriešanu. Šī darba kontekstā galvenā uzmanība tiek pievērsta induktīvās spriešanas izmantošanai datu automātiskā vai daļēji automātiskā apstrādē, tādēļ tiek lietots termins „induktīvā apmācība". Uzsverot induktīvās apmācības priekšrocības, jāmin, ka tās iegūtie klasifikācijas rezultāti ir saprotami ne tikai datorsistēmai, bet arī cilvēkam, kas ir neatsverami sistēmās, kur nepieciešams tālāk apstrādāt iegūtos spriedumus un izprast lēmuma pieņemšanas ceļu [14]. Induktīvā apmācība datorzinātnē nozīmē mācīšanos no piemēriem, kad sistēma cenšas inducēt koncepta aprakstu $c: X \rightarrow K$ no datu kopas $X = \{x_1, ..., x_i\}$, kurai ir zināma klašu kopa $K = \{k_1, ..., k_j\}$. Katrs piemērs x sastāv no atribūtu-vērtību pāriem $x = \{(a_1, v_{a1}), ..., (a_n, v_{an})\}$. Induktīvā apmācība un klasifikācija ir plaši apskatīta autores bakalaura un maģistra darbos [15, 16]. 2.1. att. atspoguļots klasifikācijas process kā klasifikatora apmācības, testēšanas un lietošanas posmi.

Klasifikācijas uzdevumi atšķiras pēc objektiem piešķiramo kategoriju skaita, ko nosaka problēmsfēra. Visbiežāk ir nepieciešams noteikt objekta piederību tikai vienai klasei. Tas ir, katrs piemērs ir saistīts ar vienu klasi k no nepārklājošos klašu kopas K, $|K| > 1$. Tomēr ir arī sfēras, kurās objekti var piederēt vienlaicīgi vairākām klasēm. Tādā gadījumā runa ir par daudzkategoriju klasifikāciju, un katrs piemērs ir saistīts ar apakškopu $Y \subseteq K$. Par sava veida daudzkategoriju klasifikāciju var uzskatīt arī izplūdušo klasifikāciju (ang. v. – *fuzzy classification*), kur objekti var piederēt dažādām klasēm ar noteiktu piederības pakāpi. Tomēr daudzkategoriju un izplūdušās klasifikācijas mērķi un risinātās problēmas ir atšķirīgas [17].



2.1. att. Klasifikācijas process

Uz izplūdušo loģiku balstītā klasifikācija ir līdzeklis neskaidrības kliedēšanai starp klases aprakstošajiem atribūtiem un var tikt uzskatīta kā sagatavošanās bloks pirms klasifikācijas, lai nošķirtu dažādas klases. Izplūdušās piederības vērtības pēc normalizācijas iegūst summāro vērtību "1", kamēr daudzkategoriju klasifikācijā vairākas klases var saņemt piederības vērtību "1" , jo klases nav savstarpēji izslēdzošas [18]. 2.1. tabulā sniegts piemērs vienkategorijas, daudzkategoriju un izplūdušās klasifikācijas rezultātam.

2.1. tabula

Vienas kategorijas, daudzkategoriju un izplūdušās klasifikācijas piemērs

Pieeja	**Objekts (aprakstošie atribūti)**	**Piešķirtās klases**				**Secinājums**
		A	**B**	**C**	**D**	
Vienkategorijas klasifikācija	a1 = 1, a2 = 1, a3 = 0	**1**	**0**	**0**	**0**	Objekts pieder klasei A
Daudzkategoriju klasifikācija		**1**	**0**	**1**	**0**	Objekts pieder klašu kopai {A, C} , nepieder klašu kopai {B, D}
Izplūdusī klasifikācija		**0.5**	**0.1**	**0.3**	**0.1**	Objekts pieder klasei A ar visaugstāko piederības vērtību

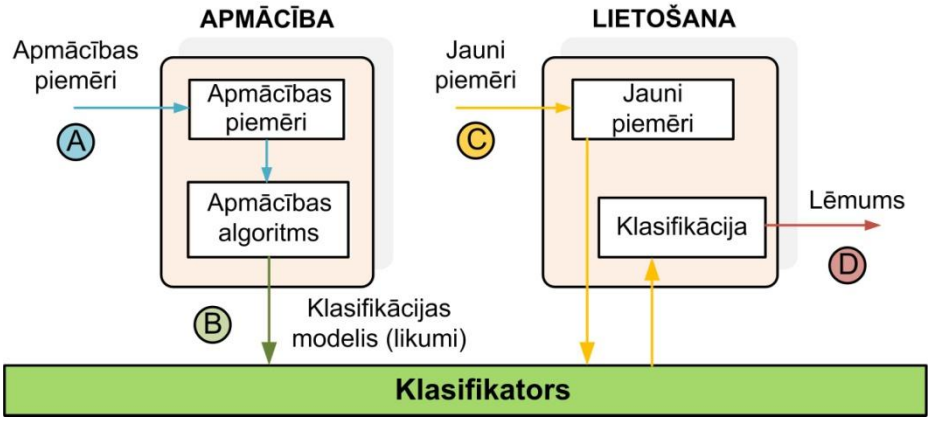
Promocijas darbā uzsvars ir likts uz daudzkategoriju klasifikācijas uzdevumu, jo galveno darbā risināmo problēmsfēru – studiju priekšmetu salīdzināšanu – raksturo īpašība, ka viens mācību priekšmets var būt līdzīgs vai atbilstošs vairākiem citiem. Daudzkategoriju datu apstrādi var veikt divos veidos – (1) caur problēmas transformāciju uz vienu vai vairākiem vienkategorijas uzdevumiem (piemēram, izmantojot binārās saistības (ang. v. - *Binary relevance*) metodi vai klašu kopu (ang. v. - *Label powerset*) veidošanu) vai (2) algoritmu adaptāciju tiešai daudzkategoriju uzdevuma risināšanai [18].

Lai varētu spriest, vai apmācības laikā iegūtais klasifikators ir derīgs jaunu piemēru klases piederības noteikšanai, tas ir jānovērtē. Klasifikatoriem, kas darbojas ar daudzkategoriju datiem, ir savas specifiskas novērtēšanas metodes. Darbā apskatītas biežāk izmantotās metrikas, kas aprakstītas avotos [19-22].

Daudzas problēmas, ar kurām jāsaskaras induktīvās apmācības lietošanas gaitā, tiek sekmīgi risinātas, piemēram, datu pirmapstrāde. Tomēr jaunu piemēru klasificēšanā ir trūkumi, kuri pagaidām nav novērsti. Sīkāk apskatītas problēmas, kas rodas, ja pieejamā apmācības kopa ir neliela un iegūtais klasifikators nespēj noteikt klases piederību jaunam piemēram. Metodes klasifikācijas nodrošināšanai gadījumos, kad klasifikatora likumu bāze nespēj nodrošināt jauna piemēra klasificēšanu, nav pilnīgas un piemērotas visām dzīves situācijām. Šādu neklasificētu piemēru problēmas risināšanai populāra ir algoritmos *AQ* [23] un *CN2* [24] piedāvātā noklusētā likuma lietošana, kurš piešķir apmācības kopā visizplatītāko klasi tiem piemēriem, kurus klasifikators nav varējis klasificēt [24]. Lai arī metode ir vienkārša un labi pamatojama, ne vienmēr tā strādā pieņemami. Piemēram, situācijā, kad klašu ir daudz un tās visas parādās caurmērā vienlīdz bieži, šāda pieeja nedod labu klasifikācijas rezultātu. Tādēļ var secināt, ka promocijas darba uzdevums – eksperta iesaistīšana neklasificēto un nepārliecinoši klasificēto piemēru klases piederības noteikšanai – ir jauns un potenciāli lietderīgs ieviesums klasifikācijas rezultātu uzlabošanā, tādējādi arī paplašinot induktīvās apmācības lietošanas iespējas jaunās jomās.

2.2. Interaktivitāte klasifikācijā un induktīvajā apmācībā

Pirms piedāvāt jaunu interaktīvu risinājumu induktīvajā apmācībā, promocijas darba autore ir apkopojusi [25, 26] dažādu avotu sniegtās līdzšinējās pieejas [27-32] interaktīvai induktīvajai apmācībai, kā arī citām klasifikācijas metodēm. Apkopojuma rezultātā ir izstrādāta shēma (skat. 2.2. att.), kurā izdalītas tās fāzes klasifikatora veidošanā un lietošanā, kurās saskarsme ar lietotāju vai ekspertu, saskaņā ar dažādām pieejām, ir nepieciešama vai vēlama. Klasifikatora veidošanas posmā izdalīta apmācības datu sagatavošana un atlase (A) un izveidoto likumu atlase un apstrāde (B). Klasifikatora lietošanas posmā izdalīta jauno klasificējamo piemēru apstrāde (C) un pieņemtā lēmuma apstrāde (D).



2.2. att. Interaktivitātes brīži starp sistēmu un tās lietotāju

Dažādās līdz šim aprakstītajās sistēmās sadarbība ar lietotāju izpaužas posmā A vai gan posmā A, gan B, vai arī posmā D, bet ne posmā C. Esošās lietotāja iesaistes notiek vai nu par ātru, vai par vēlu, kad lēmums par jauna piemēra klasifikāciju jau ir pieņemts. Savukārt tieši neklasificētu un nepārliecinoši klasificētu piemēru apstrādē ir lietderīgi iesaistīt sfēras ekspertu posmā C. Apskatīti arī speciāli interaktīvas klasifikācijas gadījumi – aktīvā mācīšanās [33] un *RippleDown* likumi [34].

2.3. Klasifikācijas sistēmu arhitektūra

Lai izstrādātu interaktīvas klasifikācijas sistēmas arhitektūru, ir analizētas dažādas esošās klasifikācijas sistēmu arhitektūras [35-44], ar arhitektūru saprotot gan sistēmas projektēšanas posmus, gan uzbūvi. Pētījuma rezultātus autore pirmo reizi ir publicējusi [45]. Atšķirības starp aprakstītajām arhitektūrām galvenokārt nosaka risinājuma mērogs un lietojuma sfēra. Detalizēta klasifikācijas sistēmu uzbūve ir specifiska katram lietojumam, ar mazu atkārtotu izmantojamību, kas ļauj secināt, ka arī interaktīvas sistēmas uzbūve būs unikāla un veidojama no jauna. Savādāk ir ar pašu sistēmas projektēšanas procesu; lielākā daļa no 9 apskatītajiem projektēšanas modeļiem satur kopīgas iezīmes – (1) uzsverot sākotnējo problēmsfēras izpētes nozīmību un (2) labākā risinājuma meklēšanu kā daļu no projektēšanas aktivitātēm. Tādējādi sistēmas projektēšanas modeļi ir atkārtoti izmantojami, un interaktīvas klasifikācijas sistēmas izstrādei nav nepieciešams radīt jaunu projektēšanas ciklu, bet var izmantot kādu no līdz šim aprakstītajiem un praksē pārbaudītajiem.

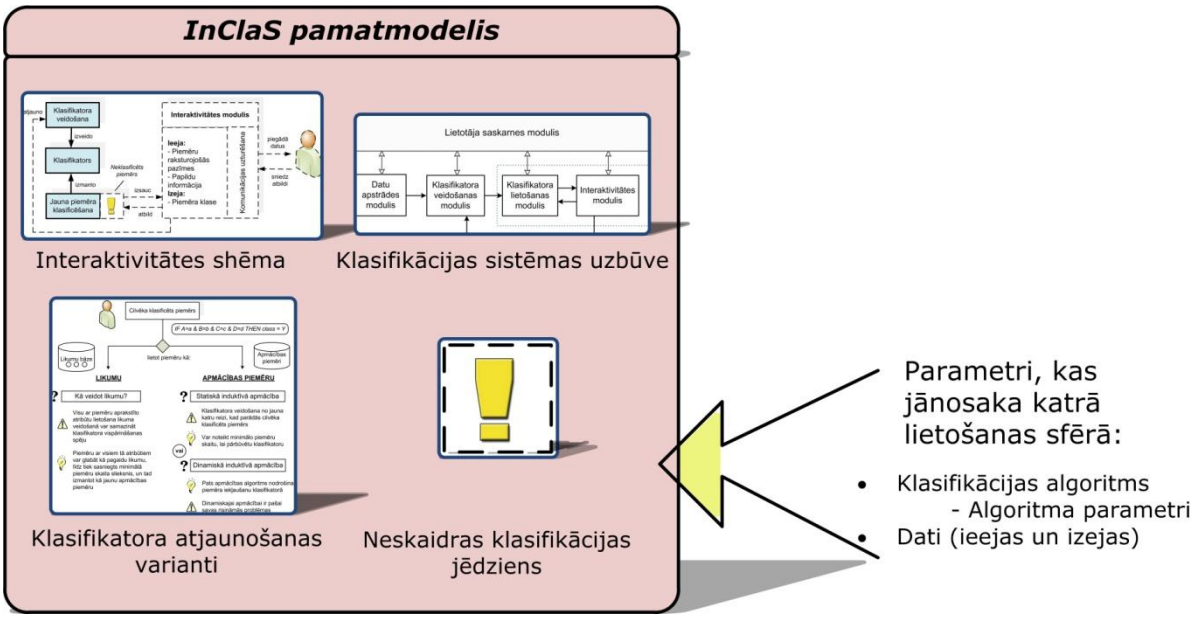
Balstoties uz veikto literatūras analīzi un konstatētajiem pastāvošo pieeju trūkumiem, ir izstrādāts interaktīvas uz induktīvo apmācību balstītas klasifikācijas sistēmas pamatmodelis. Tā komponentes ir aprakstītas nākošajā nodaļā.

3. INTERAKTĪVAS UZ INDUKTĪVO APMĀCĪBU BALSTĪTAS KLASIFIKĀCIJAS SISTĒMAS (*INCLAS*) PAMATMODELIS

Klasifikācijas sistēmas papildināšana ar interaktivitāti ne tikai var uzlabot klasifikācijas rezultātus, bet arī ļaut lietotājiem ieskatīties sistēmas darbībā. Šajā nodaļā izklāstīts promocijas darba autores izstrādātais interaktīvas klasifikācijas sistēmas ***InClaS*** (ang. v. - ***In****teractive Inductive Learning Based* ***Cla****ssification* ***S****ystem*) modelis, kura atsevišķo komponenšu apraksti ir publicēti [25, 45-48].

InClaS modelis definē šādas nepieciešamās komponentes interaktīvas klasifikācijas sistēmas izveidei (skat. 3.1. att.):

- vispārējo ieviešamo interaktivitātes shēmu (skat. 3.1. apakšnod.);
- interaktīvajā sistēmā apstrādājamās *neskaidrās klasifikācijas* jēdzienu (skat. 3.2. apakšnod.);
- ieteicamos klasifikatora atjaunošanas variantus ar skaidrojumu, kā tie nodrošina klasifikatora likumu bāzes saskanīgumu (skat. 3.3. apakšnod.);
- klasifikācijas sistēmas uzbūvi, tās moduļus un saites starp tiem (skat. 3.4. apakšnod.).



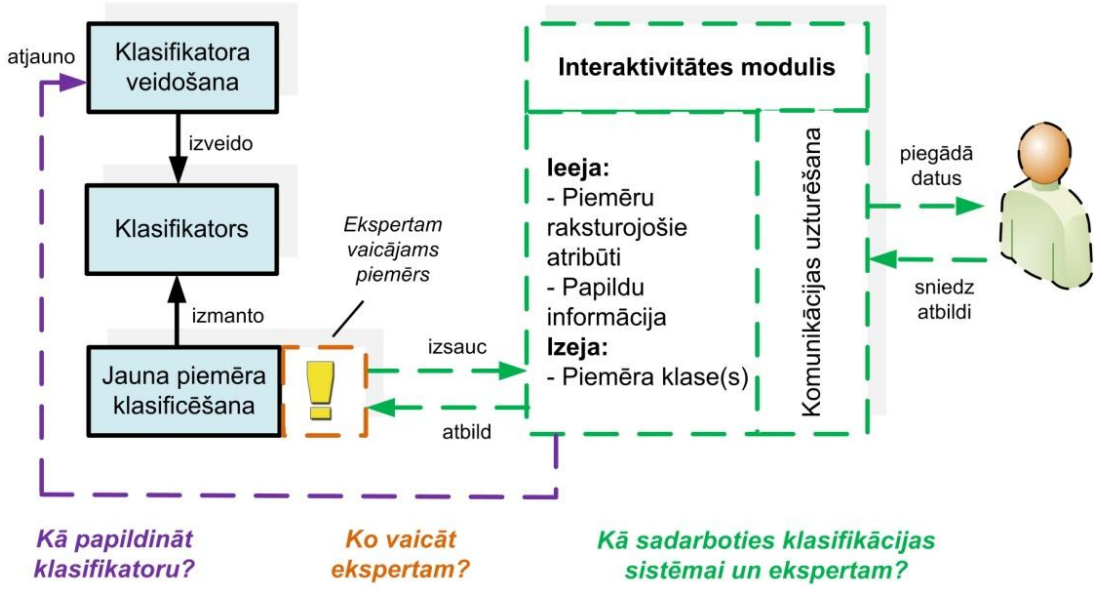
3.1. att. Interaktīvās klasifikācijas sistēmas pamatmodelis

Ir identificēti arī parametri, kas jānosaka katrā *InClaS* lietošanas sfērā (skat. 3.1. att.). Klasifikācijas algoritma un tā iespējamo parametru izvēle ir jāveic ikvienā jomā, kur ir plānots izmantot klasifikāciju, un ir spēkā arī interaktīvas sistēmas kontekstā. Sistēmas

izveide konkrētai problēmsfērai nevar tikt iepriekš pilnībā definēta, bet tās izstrāde ir atbalstīta ar darba autores adaptētu Bielavski un Levanda piecu soļu metodi *intelektuālu sistēmu projektēšanā* [42], kas atvieglo analītisko darbu klasifikācijas sistēmas izveidē.

3.1. Vispārējā ieviešamā interaktivitātes shēma

Interaktīvas klasifikācijas sistēmas mērķis nav kāda konkrēta induktīvās apmācības algoritma uzlabošana, bet gan klasifikācijas iespēju paplašinājuma izstrāde, kas ļautu lietot interaktīvu pieeju visiem tiem induktīvās apmācības algoritmiem, kuros nav definēti citi mehānismi neklasificēto piemēru apstrādei (piemēram, noklusētais likums) vai šie mehānismi var tikt izmainīti, neizdarot lielas izmaiņas pašā algoritmā. Šī pieeja klasifikācijas procesā maina veidu, kā klasifikators tiek lietots jaunu piemēru klases piederības noteikšanai, nevis to, kā klasifikators tiek veidots (skat. 3.2. att.). Attēlā redzama ieviešamā interaktivitātes shēma, iekļaujot klasifikācijas sistēmas pamatelementus un saites (attēla elementi ar nepārtrauktu līniju), kā arī elementus un saites, kas nodrošina interaktivitātes ieviešanu (attēla elementi ar raustītu līniju) neklasificēto piemēru apstrādei. Brīdī, kad klasifikators sastopas ar piemēru, kura klasifikācija tam ir neskaidra, klasifikācijas sistēmai ir iespēja pavaicāt sistēmas lietotājam - ekspertam, kādu klasi viņš piešķirtu šim objektam. Iegūtās zināšanas var izmantot ne tikai lēmumam par konkrētā piemēra klasi, bet arī klasifikatora pilnveidošanai, paplašinot sākotnējo apmācības kopu.



3.2. att. Interaktivitātes iekļaušana vispārīgajā klasifikācijas modelī

Jautājumi, kas rodas, paplašinot vispārēju klasifikācijas pieeju ar interaktivitāti (redzami 3.2. att.), tiks atbildēti nākamajās apakšnodaļās, izklāstot citas *InClaS* komponentes.

3.2. Neskaidras klasifikācijas jēdziens

Lai atbildētu uz jautājumu „Ko vaicāt ekspertam?", ir izskaidrots *neklasificēta, nepārliecinoši klasificēta* un *klasifikatoram neskaidra* piemēra jēdziens, kas lietots šajā darbā.

Neklasificēts piemērs (ang. v. – *unclassified instance*) ir tāds piemērs, kam pēc klasifikatora lietošanas nav izdevies noteikt klases piederību, balstoties uz klasifikācijas modeli (neviens likums vai zars lēmumu kokā neatbilst klasificējamam piemēram). Neklasificētu piemēru apstrādes iespējas ar noklusēto likumu un tā trūkumi tika minēti iepriekš darba 2.1. apakšnodaļā.

Ja tiek ņemta vērā pārliecība, ko klasifikators asociējis ar piemēru klasificējošo likumu (vai lapu lēmumu kokā), tad klasifikatora pieņemto lēmumu var izvērtēt pēc tā pārliecības. ***Nepārliecinoši klasificēts piemērs*** jeb piemērs ar zemu klasifikācijas pārliecību (ang. v. – *low confidence of classification*) ir tāds piemērs, kam klasifikatora iegūtā klases pārliecība ir pārāk zema, lai piešķirtu klasi. Pārliecība ir balstīta uz piemēru sadalījumu apmācības kopā, kas tika izmantota klasifikatora izveidošanai. Dažādas klasifikācijas metodes un algoritmi lieto atšķirīgus pārliecības noteikšanas veidus. Piemēram, ja likums, kurš nosaka klases *A* piešķiršanu, pārklāj 3 piemērus ar klasi *A* un 2 piemērus ar klasi *B*, tad klašu sadalījums ar šo likumu klasificētam jaunam piemēram ir 0.6 klasei A un 0.4 klasei *B,* iegūstot klasifikatora pārliecību - 0.6 (par piešķiramo klasi - *A*). Tradicionāli pārliecības līmenis 0.5 ir slieksnis, lai likumu sāktu izmantot klasifikācijā. Tomēr var tikt lietoti arī citi sliekšņa lielumi, atlasot klasifikatora lēmumus ar lielāku pārliecību, tādējādi cenšoties iegūt vairāk pareizi klasificētu piemēru.

Klasifikatoram neskaidrs piemērs jeb neskaidra klasifikācija (ang. v. – *uncertain classification*) šī darba kontekstā ietver gan neklasificētus piemērus, gan nepārliecinoši klasificētus piemērus.

Daudzkategoriju klasifikācijas gadījumā, kur piemērs var piederēt vairākām klasēm, pārliecības noteikšana par klasifikatora pieņemto lēmumu kļūst sarežģītāka. Šī temata tālāks izklāsts, tāpat kā piemērotākā pārliecības sliekšņa lieluma noteikšanas metode, tiks sniegta darba nākamajā nodaļā, kas veltīta *InClaS* paplašināšanai daudzkategoriju klasifikācijas uzdevumiem.

3.3. Ieteicamie klasifikatora atjaunošanas varianti

Kad klasifikatoram neskaidrais piemērs ir nodots ekspertam un ir saņemts eksperta lēmums par piemēra klasifikāciju, sistēmai nepieciešams apstrādāt iegūto informāciju -

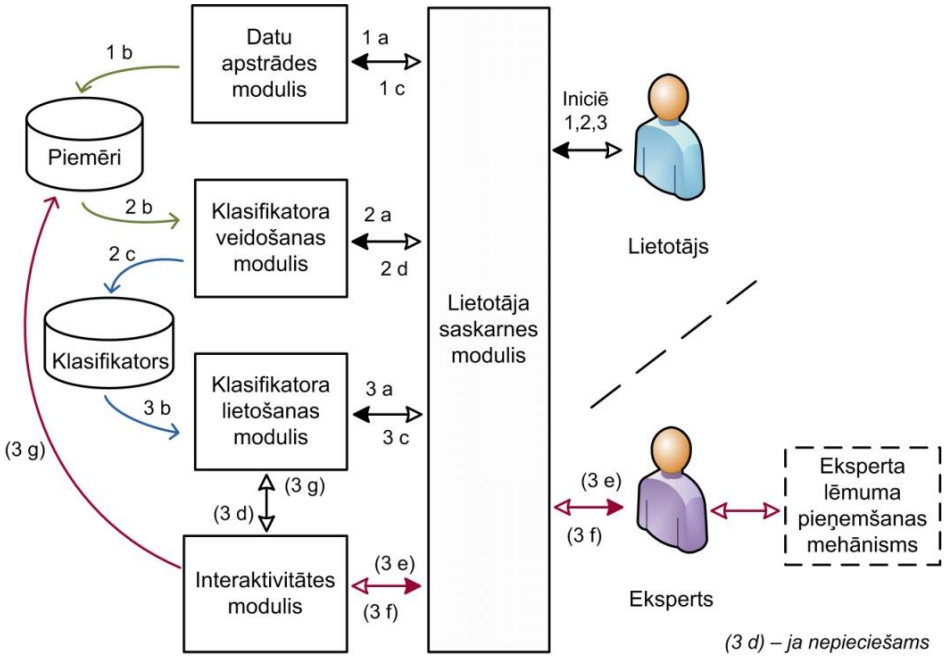
atbildēt uz jautājumu „Kā papildināt klasifikatoru?”. Klasifikatora atjaunošanai darba autore piedāvā divas alternatīvas (tās sīkāk ir aprakstītas [47]) :

Uz slieksni balstīta statiskās apmācības pieeja, kas, kā nosaukums liecina, izmanto statisku apmācības algoritmu (šādi algoritmi paredz klasifikatora papildināšanu ar jauniem apmācības piemēriem, tikai veidojot klasifikatoru pilnībā no jauna) un eksperta klasificētos piemērus pievieno kā pagaidu likumus, līdz ir sasniegts iepriekš uzstādītais piemēru skaita slieksnis. Kad slieksnis ir sasniegts, piemēri tiek pievienoti sākotnējai apmācības piemēru kopai, un tiek veikta klasifikatora atkārtota apmācība, bet pagaidu likumi - dzēsti.

Dinamiskās apmācības pieeja jau sākotnējā klasifikatora veidošanai izmanto dinamiskās apmācības algoritmu, kas ļauj papildināt apmācības kopu klasifikatora lietošanas laikā, neveidojot klasifikatoru no jauna, eksperta klasificēto piemēru izmantojot kā jaunu apmācības piemēru un apstrādājot to atbilstoši savam algoritmam.

3.4. Klasifikācijas sistēmas uzbūve

Daļu no atbildes „Kā sadarboties klasifikācijas sistēmai un ekspertam?” sniedz klasifikācijas sistēmas uzbūve. Balstoties uz izstrādāto interaktīvās klasifikācijas sistēmas vispārīgo shēmu (3.2. attēlā) un 2.3. nodaļā iegūto klasifikācijas sistēmu arhitektūru apkopojuma analīzi, projektējamajā sistēmā ir izdalītas komponentes, kas atbild par dažādām funkcijām. Interaktīvai klasifikācijas sistēmai ir izvēlēta modulāra arhitektūra. Sistēmas uzbūve aprakstīta 3.1. tabulā, bet 3.3. attēls atspoguļo tipisku sistēmas lietošanas scenāriju, neaprakstot iekšējos procesus moduļos.



3.3. att. Interaktīvas klasifikācijas sistēmas funkcionēšana

Caur lietotāja saskarni sistēmas lietotājs sniedz apmācības piemērus (3.3. attēlā solis *1a*), iniciē klasifikatora izveidi (*2a*) un izsauc *klasifikatora lietošanas moduli* (*3a*), kurš, izmantojot no *apmācības piemēriem* iegūto *klasifikatoru*, veic piemēru klases piederības noteikšanu. Ja klasifikācija ir veiksmīga, t.i., klasifikators spēj atrast atbilstošus likumus, rezultāti tiek demonstrēti lietotājam (*3c*). Ja piemērs netiek klasificēts vai tiek klasificēts nepārliecinoši, *klasifikatora lietošanas modulis* sūta pieprasījumu *interaktivitātes modulim* risināt situāciju (*3d*). *Interaktivitātes modulis* lūdz neklasificētā piemēra klasifikāciju veikt sistēmas lietotājam – ekspertam (*3e*); šajā gadījumā darbība tiek iniciēta no sistēmas puses, nevis no lietotāja puses, kā iepriekšējos gadījumos. Saņemot lietotāja atbildi (*3f*), *Interaktivitātes modulis* sniedz atbildi *Klasifikatora lietošanas modulim* un papildina *Piemēru bāzi* (*3g*), kā rezultātā arī iespējama *Klasifikatora* atjaunošana, sazinoties ar *Klasifikatora veidošanas moduli (2b, 2c)*. Eksperta lēmumu pieņemšanas mehānismi ir ārpus šī darba apskatāmo jautājumu loka. Klasifikācijas sistēmā tiek sagaidīts viens eksperta lēmums par piemēra klases piederību.

3.1. tabula

Interaktīvas klasifikācijas sistēmas moduļi

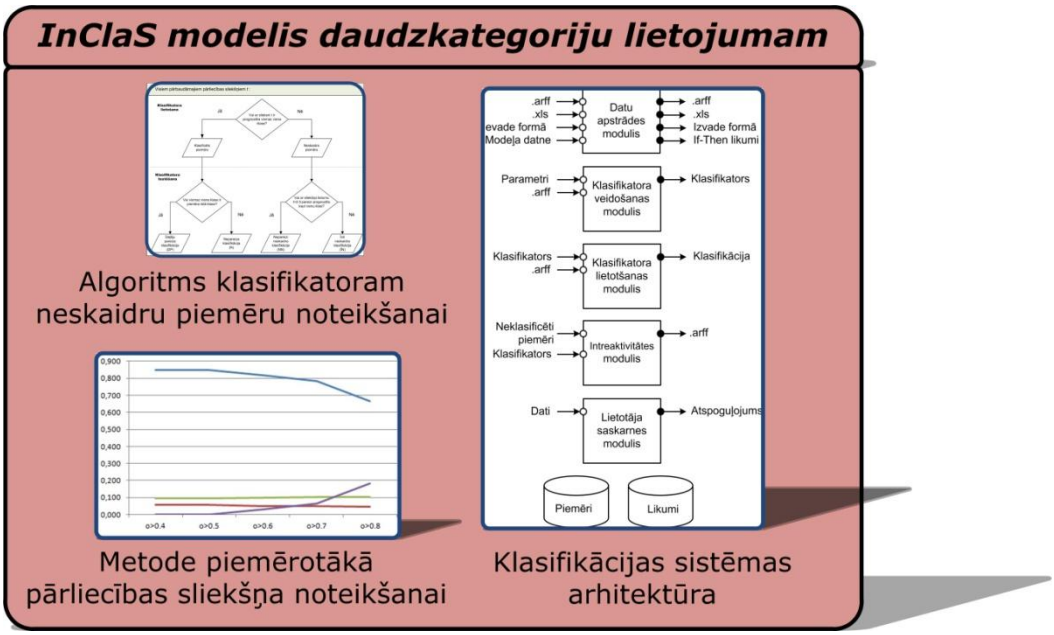
Lietotāja saskarnes modulis
Nodrošina lietotājam draudzīgu saskarni starp sistēmu un tās lietotāju: - Atspoguļo datus. - Nodrošina sistēmas lietotāja pieprasījumu izpildi, izsaucot funkcijas citos sistēmas moduļos.
Datu apstrādes modulis
Nodrošina datu attēlošanu dažādos formātos: - Nodrošina sistēmas lietotājam iespēju pievienot datus dažādos formātos, palīdzot strukturētu datu iegūšanā. - Nodrošina sistēmas lietotājam iespēju apskatīt apmācības datus un klasifikācijas likumus dažādos formātos. - Nodrošina datu transformāciju moduļu iekšējiem un savstarpējiem procesiem.
Klasifikatora veidošanas modulis
Ģenerē klasifikācijas modeli dotajai datu kopai. Klasifikators sistēmas iekšējā struktūrā tiek glabāts sistēmai specifiskā formātā. Ja apmācības algoritma attēlojuma forma ir likumi, tad no šī formāta var izgūt IF-THEN likumus. Šis modulis ir balstīts uz jau implementētu induktīvās apmācības algoritmu bibliotēku izmantošanu.
Klasifikatora lietošanas modulis
Izmanto iepriekš izveidotu klasifikatoru, lai noteiktu klašu piederību jauniem piemēriem, iegūst klasifikatora pārliecības par pieņemto lēmumu. Šis modulis izmanto gatavas implementētas apmācības metodes, kas ir papildinātas ar spēju pārtvert klasifikatoram neskaidros piemērus. Šajā gadījumā tiek izsaukts interaktivitātes modulis.
Interaktivitātes modulis
Nodrošina saziņu ar sistēmas lietotāju, sagatavojot un apstrādājot nepieciešamo informāciju. Cieši saistīts ar klasifikatora lietošanas moduli. - Nodrošina neskaidro klasifikāciju un papildu informācijas sagatavošanu atspoguļošanai ekspertam, saņem atbildi. - Iniciē piemēru bāzes atjaunošanu pēc eksperta atbildes saņemšanas. - Nodrošina likuma atspoguļošanu pēc lietotāja pieprasījuma.

Uz *InClaS* pamatmodeļa pamata tālāk tiks attīstīts nākamais modeļa līmenis interaktīvai klasifikācijas sistēmai daudzkategoriju klasifikācijas uzdevumiem.

4. *INCLAS* MODELIS DAUDZKATEGORIJU KLASIFIKĀCIJAS UZDEVUMAM

Šajā nodaļā ir papildināts 3. nodaļā sniegtais *InClaS* modelis lietošanai daudzkategoriju klasifikācijas gadījumā, tas ir, klasifikācijas uzdevumos, kur katrs objekts jeb piemērs var piederēt vienlaicīgi vairākām klasēm. *InClaS* modeļa papildinājums daudzkategoriju klasifikācijas uzdevumiem paredz šādas komponentes (skat. 4.1. att.):

- algoritms klasifikatoram neskaidru piemēru noteikšanai (skat. 4.1. apakšnod.);
- metode piemērotākā pārliecības sliekšņa noteikšanai (skat. 4.2. apakšnod.);
- klasifikācijas sistēmas arhitektūra – projektēšanas procesa detalizācija un sistēmas uzbūve (skat. 4.3. apakšnod.).



4.1. att. *InClaS* modeļa papildinājums daudzkategoriju klasifikācijai

4.1. Algoritms klasifikatoram neskaidru piemēru noteikšanai

Pirms klasifikatoram neskaidro piemēru noteikšanas algoritma definēšanas, dziļāk ir analizēts neskaidras klasifikācijas jēdziens un ieviesti svarīgi tālākā darbā izmantoti termini un mēri. Vispārējie principi piemēra atzīšanai par neskaidru tika apspriesti jau 3. nodaļā, definējot, ka darba ietvaros par klasifikatoram neskaidriem tiek saukti neklasificēti un nepārliecinoši klasificēti piemēri. Daudzkategoriju klasifikācijas uzdevumi ļauj plašāk palūkoties uz klasifikatora nespēju noteikt piemēra klases piederību. Viens no plaši izmantotiem daudzkategoriju klasifikācijas uzdevuma risināšanas variantiem ir binārās

saistības pieeja [49] – daudzkategoriju problēmu sadalot vairākos vienas kategorijas uzdevumos. Tādējādi piemēra klasifikāciju nosaka kopējie rezultāti no *n* vienas kategorijas klasifikatoriem, kur katrs atsevišķais klasifikators lemj par piemēra piederību tikai vienai klasei. Ja netiek konstatēta piederība nevienai no atsevišķajām klasēm (piemēram, pēc noklusējuma klasifikācijā tradicionāli lietoto pārliecības slieksni – 0.5 – nesasniedz neviena klase), tad piemērs var tikt uzskatīts par neklasificētu.

Papildus pazīstamajām promocijas darba 2.1. apakšnodaļā aprakstītajām daudzkategoriju klasifikācijas novērtēšanas metrikām, autore ir ieviesusi arī citas metrikas, kas konkrētāk saistās tieši ar nepareizi klasificēto piemēru skaita samazināšanu un lietotāja iesaistīšanu klasifikācija procesā. Tās balstās uz vienkāršu parametru novērtēšanu:

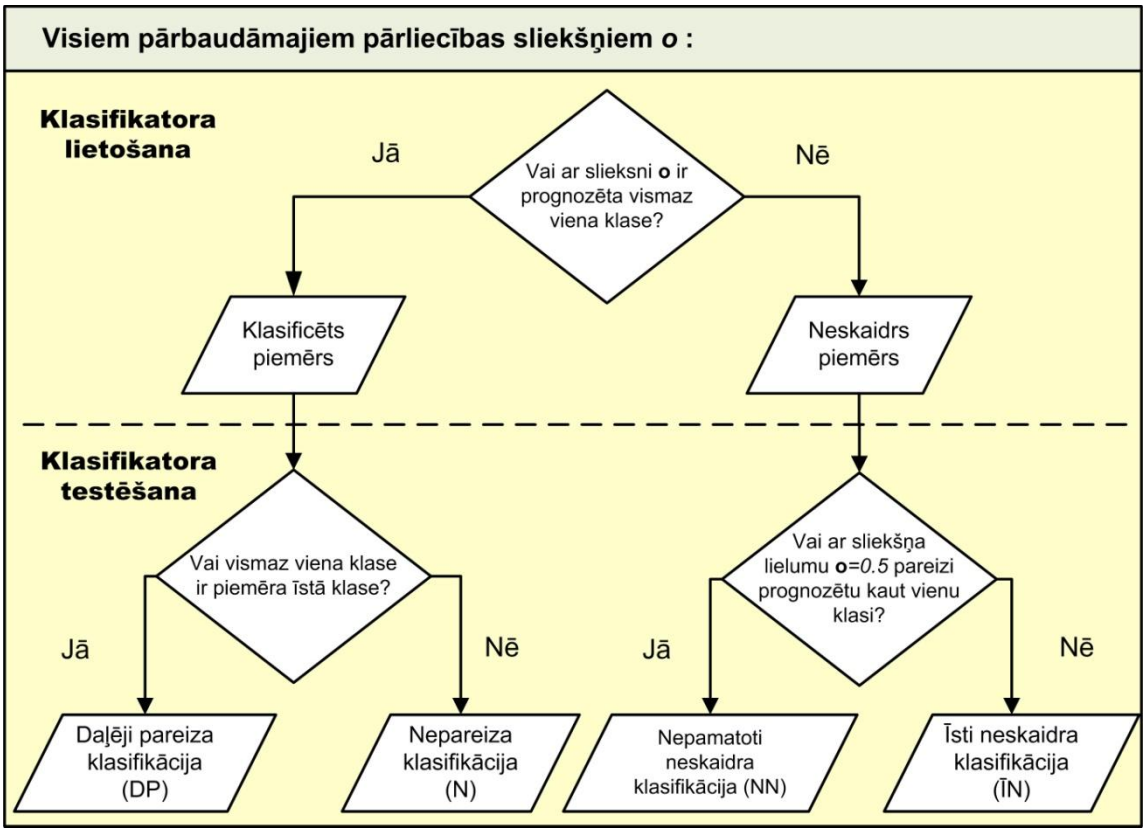
- *daļēji pareizi vai pilnīgi pareizi klasificēts piemērs (DP)* – vismaz viena no prognozētajām klasēm ir piemēra īstā klase, $Y_i \cap Z_i \neq \emptyset$, kur Y_i – īstā klašu kopa *i*-tajam piemēram, Z_i – klasifikatora prognozētā klašu kopa *i*-tajam piemēram;
- *nepareizi klasificēts piemērs (N)* – neviena no klasifikatora noteiktajām klasēm nav piemēra īstā klase, $Y_i \cap Z_i = \emptyset$;
- *īsti neskaidra klasifikācija (ĪN)* – piemērs būtu nepareizi klasificēts (*N*), ja tiktu uzticēts tikai klasifikatoram (tas ir, ja ar noklusēto pārliecības līmeni 0.5 netiktu pareizi prognozēta neviena no īstajām klasēm);
- *nepatiesi neskaidra klasifikācija (NN)* – piemērs būtu klasificēts kā vismaz daļēji pareizs (*DP*), ja būtu uzticēts klasifikatoram (tas ir, ja ar noklusēto pārliecības līmeni 0.5 vismaz viena prognozētā klase būtu piemēra īstā klase).

Protams, klasifikācijai ir jātiecas uz rezultātu, kurā pēc iespējas vairāk piemēru ir klasificēti pareizi vai vismaz daļēji pareizi, bet, ja tas nav iespējams klasifikatora nepilnību dēļ, tad klasifikācijas sistēmai būtu jāatpazīst piemēri, kuru klasifikācija klasifikatoram ir neskaidra.

Mēri cilvēka ieguldītā darba novērtēšanai ir šādi:

- $D_{nelietderīgais}$ – cik pareizi klasificētu piemēru cilvēkam jācaurskata, lai klasificētu vienu nepareizi klasificētu piemēru ($D_{nelietderīgais} = \frac{NN}{ĪN}$);
- $D_{kopējais}$ – cik piemēru cilvēkam pavisam jācaurskata ($D_{kopējais} = ĪN + NN$).

Izstrādātais algoritms klasifikatoram neskaidru piemēru noteikšanai daudzkategoriju klasifikācijas gadījumā, kurš nosaka, ka piemērs tiek atzīts par neskaidru tad, ja ar noteiktu (noklusēto vai izvēlēto) pārliecības sliekšņa lielumu ***o*** piemēram nav prognozēta neviena no piemēra īstajām klasēm, ir atspoguļots 4.2. attēlā.



4.2. att. Algoritms klasifikatoram neskaidru piemēru noteikšanai daudzkategoriju klasifikācijā

4.2. Metode piemērotākā pārliecības sliekšņa noteikšanai

Interaktīvā klasifikācijas sistēmā ir nepieciešams noteikt pārliecības sliekšņa lielumu, pie kura klasifikatora iegūtais rezultāts vairs netiek uzskatīts par uzticamu. Dažādās jomās ir atšķirīga specifika attiecībā uz pārliecības līmeni. Kamēr vienā datu kopā līmenis 0.5 labi nošķir piederību vai nepiederību klasei, tikmēr citā var būt nepieciešams arī augstāks slieksnis, lai lēmums būtu pārliecinošs. Autores izstrādātā piemērotākā pārliecības sliekšņa noteikšanas metode palīdz atrast šo lielumu katrai datu kopai. Kā pārmeklējamā apgabala indikatori ieviesti mēri, kas raksturo **v**idējo klasifikatora **p**ārliecību par klasēm, kurām piemēri ir **p**iederīgi (*VPP*) vai **n**av piederīgi (*VPN*). Metodes uzdevums: atrast atbilstošāko pārliecības slieksni, kur nepareizi klasificēto piemēru skaits N ir minimāls pie lietotāja izvirzītajiem ieguldāmā darba ierobežojumiem (d_1 un d_2): $N \rightarrow min$, $D_{kopējais} \leq d_1$; $D_{nelietderīgais} \leq d_2$.

Metode manuālai piemērotākā pārliecības sliekšņa lieluma izvēlei

1. Izvēlēties klasifikācijas algoritmu, veikt klasifikatora apmācību un testēšanu, noteikt *N, ĪN, NN,* $D_{kopējo}$ un $D_{nelietderīgo}$ pie pārliecības sliekšņa 0.5. Ja neklasificēto piemēru skaits $D_{kopējais}$ prasa pārāk lielu eksperta darbu, izmēģināt citu klasifikācijas algoritmu.

2. Ja neklasificēto piemēru skaits $D_{kopējais}$ ir pieņemams praktiskam lietojumam, mainīt sliekšņa lielumus un noteikt *N, ĪN, NN,* $D_{kopējo}$ un $D_{nelietderīgo}$Standarta solis ir 0.1, bet iespējams lietot sīkāku soli.
3. Atspoguļot un izvērtēt rezultātus ar mēriem $D_{kopējais}$, $D_{nelietderīgais}$, *N*.

Novērtējumam par pamatu tiek ņemts nepareizi klasificēto piemēru skaits. Izvērtēt sliekšņa lielumus, ņemot vērā šādus faktorus:

- ja pie konkrēta sliekšņa *N* nav lielāks kā iepriekšējā solī, ir vērts izskatīt sliekšņa palielināšanu;
- ja ir vienāds nepareizi klasificēto piemēru skaits vairākos stāvokļos, tad dot priekšroku stāvoklim ar mazāku $D_{kopējo}$ (un *ĪN* skaitu, kas ir tieši saistīti lielumi);
- ir iespējami vairāki savstarpēji ekvivalenti stāvokļi, kuros parametri nemainās.

Ir izstrādāta arī automātiska piemērotākā pārliecības sliekšņa lieluma izvēles metode, kura kā ieejas parametrus saņem $D_{kopējo}$ un/vai $D_{nelietderīgo}$, bet izdod pārliecības līmeni, pie kura *N* ir minimāls, ņemot vērā ieejas parametrus.

Jāņem vērā, ka šādā veidā iegūtie aprēķini balstās uz sadalījumu apmācības kopā un var nebūt pilnībā patiesi datiem, ar kuriem klasifikācijas sistēma saskarsies nākotnē, klasificējot jaunus objektus.

4.3. Interaktīvas daudzkategoriju klasifikācijas sistēmas arhitektūra

Sistēmas projektējums iekļauj pieņemto lēmumu aprakstu atbilstoši izvēlētajai projektēšanas procedūrai – Bielavski and Levanda *Intelektuālu sistēmu projektēšanai* piecos soļos [42] –, analizējot (1) problēmas identificēšanu, (2) zināšanu iegūšanu un attēlošanu, (3) rīku izvēli, (4) prototipēšanu un izstrādi, kā arī (5) testēšanu un uzturēšanu. Tā ietvaros ir detalizēti studiju priekšmetu salīdzināšanas problemātikai raksturīgie parametri un klasifikācijas sistēmu veidojošo moduļu realizācijas detaļas daudzkategoriju klasifikācijas veikšanai. Šī *InClaS* modeļa komponente aprakstīta arī publikācijās [7, 50-52].

Iegūtais *InClaS* pamatmodelis un tā papildinājums ar daudzkategoriju klasifikācijas nodrošināšanai nepieciešamajām komponentēm sniedz pietiekamu teorētisko un metodisko bāzi interaktīvas klasifikācijas sistēmas realizācijai programmatūras veidā.

5. *INCLA*S PROTOTIPS

Lai izstrādāto *InClaS* modeli pārbaudītu un interaktīvo klasifikācijas sistēmu ieviestu praktiskā lietošanā, ir realizēts *InClaS* prototips valodā *Java*. Nodaļā sniegts prototipa funkcionalitātes apraksts, pievēršot uzmanību tam, kā atsevišķās *InClaS* komponentes no pamatmodeļa un tā paplašinājuma daudzkategoriju klasifikācijai ir realizētas programmatūrā.

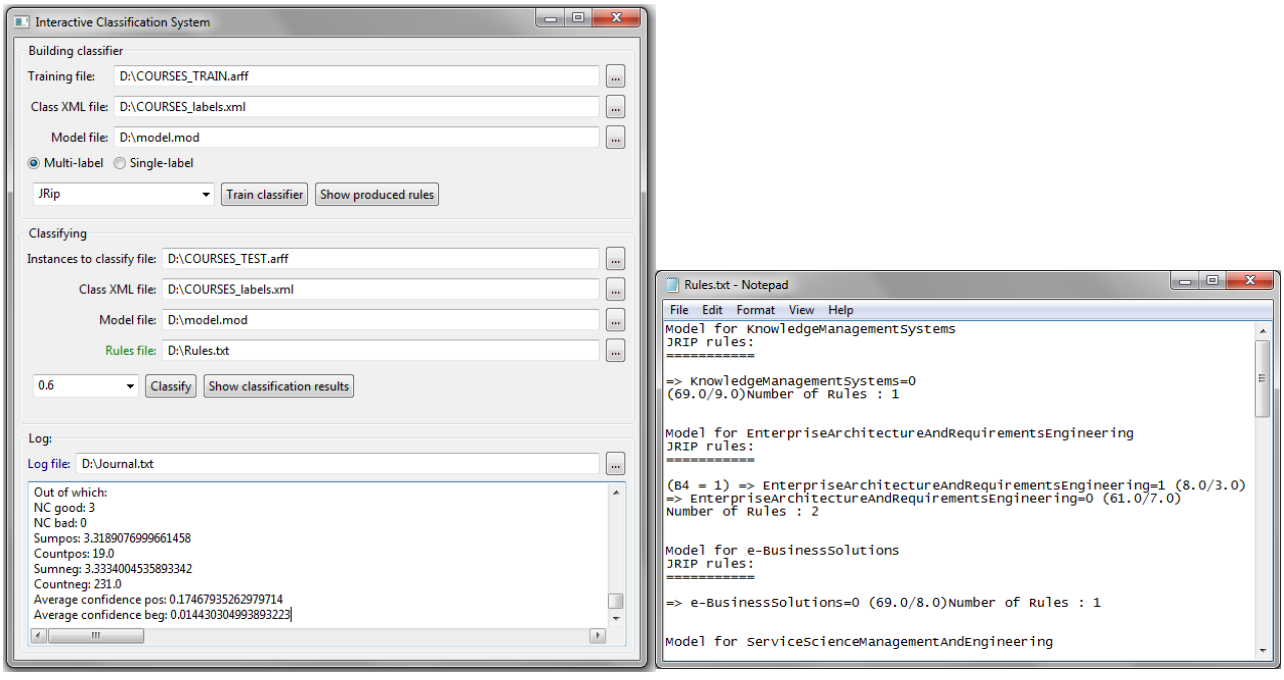
- Prototipa ietvaros tiek izmantoti iepriekš realizēti klasifikācijas algoritmi; bāzes algoritmi tiek izsaukti no programmatūras *Weka* [3], bet tos izmantojošās daudzkategoriju klasifikācijas metodes aprakstītas bibliotēkā *Mulan* [4]. Sistēmā šobrīd ieviesti 11 statiskie klasifikācijas algoritmi no *Weka* un *Mulan* klāsta, izmantojot to noklusētos parametrus.
- Interaktivitātes shēma realizēta, jaunu piemēru klasifikācijas procedūru papildinot ar iespēju izsekot nepārliecinoši klasificētus piemērus – klasificēšanas brīdī pārbaudot, ar kādu pārliecību veikta klasifikācija, un demonstrējot rezultātus lietotājam. Atbilstoši, lietotājs var izvērtēt ar dažādām pārliecības pakāpēm piešķirtās klases un sniegt savu klasifikāciju gadījumos, kad neviena klase nav piešķirta ar pārliecību 0.5 vai vairāk.
- No klasifikatora atjaunošanas variantiem tiek izmantota *Uz slieksni balstītā statiskā pieeja* ar slieksni 1. Tas nozīmē, ka klasifikators tiek atjaunots pēc katra jauna lietotāja klasificēta piemēra parādīšanās. Praktiski, ja vienā jaunu piemēru klasifikācijas reizē tiek klasifikācijai nodoti vairāki piemēri, un arī lietotājs klasificē vairāk par vienu piemēru, tad klasifikators tiks atjaunots, reizē izmantojot visus jaunos piemērus.
- Nodrošinātas datu ievades iespējas un sistēmas iegūto rezultātu izvade ekrāna formā.
- Algoritms klasifikatoram neskaidru piemēru noteikšanai ir pilnībā realizēts sistēmas prototipā klasifikatora lietošanas modulī. Jaunu piemēru klasifikācijas laikā tiek iegūtas klasifikatora pārliecības par visām piešķiramajām klasēm, kuras tālāk tiek atbilstoši izmantotas lēmumu pieņemšanai.
- Sistēma izgūst un saglabā teksta datnē lietotājam lasāmā formā klasifikatoru veidojošos likumus.
- Strādājot eksperimentu veikšanas režīmā, sistēmai nav pilnvērtīga lietotāja saskarne, bet ir pieejamas plašākas iespējas, tajā skaitā 20 algoritmu pārbaude (ko var vēl paplašināt ar citām metodēm un algoritmiem) un iespēja iegūt nepieciešamos sistēmas darbības rezultātus dažādiem novērtējumiem. Iegūstami rezultāti gan par

populārākajiem daudzkategoriju klasifikācijas novērtējumu mēriem, gan aprēķini darba autores izvirzītajiem mēriem – *DP, N, ĪN, NN*, *VPP* un *VPN*.

- Piemērotākā pārliecības sliekšņa noteikšana ir veicama ar manuālo metodi.

Uzsverot izstrādnes jauninājumus, tālāk paskaidrots, kā šis sistēmas prototips atšķiras un papildina klasifikācijas uzdevumiem plaši izmantoto rīku *Weka* un daudzkategoriju bibliotēku *Mulan*. Par galvenajām papildinošajām atšķirībām jāmin (1) izstrādātā lietotāja saskarne daudzkategoriju klasifikācijas bibliotēkas *Mulan* izmantošanai (*Mulan* izstrādātāji nepiedāvā savu grafisko lietotāja saskarni), (2) iespēja lietotājam ērti apskatīt klasifikatora iegūto likumu kopu (ja izvēlētais algoritms producē interpretējamu klasifikatoru) un (3) lietotāja saskarne interaktivitātes nodrošināšanai. Šie jauninājumi kopumā veido unikālu vidi daudzkategoriju klasifikācijas nodrošināšanai lietotājam ērtākā formā nekā bija iespējams ar pastāvošajiem rīkiem, kā arī līdz šim nebijušas interaktivitātes iespējas starp klasifikācijas sistēmu un lietotāju.

5.1. attēla A daļa sniedz ieskatu sistēmas prototipa lietotāja saskarnē, bet attēla B daļā redzams klasifikatora iegūtās likumu kopas fragments uz kompetencēm balstītajā (netiešajā) priekšmetu salīdzināšanā.



A **B**

5.1. att. Klasifikatora testēšanas rezultātu izvade (A) un klasifikatora iegūtie likumi (B)

Izstrādātais prototips ir galvenā programmatūras bāze praktisko eksperimentu veikšanai, ar kuru palīdzību ir pārbaudīts *InClaS* modelis un piedāvātā interaktīvā pieeja.

6. *INCLAS* MODEĻA NOVĒRTĒJUMS

Šī nodaļa sniedz eksperimentu plānu un galvenos iegūtos rezultātus darbā ar interaktīvas klasifikācijas sistēmas prototipu daudzkategoriju klasifikācijas uzdevumos. Eksperimenti ir veikti ar mērķi pārbaudīt interaktīvās pieejas un *InClaS* modeļa lietderību un sistēmas prototipa lietojamību, kā arī apstiprināt izvirzītās tēzes (T1, T2 un T3). Atgādinot problēmapgabalu, kas tiek risināts promocijas darbā - no mašīnapmācības viedokļa tiek risināts uzdevums, kuram raksturīgas šādas īpašības: (1) iegūtie rezultāti ir jāsaprot klasifikatora lietotajam un ekspertam, (2) pieejamā apmācības kopa ir maza, (3) sākotnējie dati ir daļēji strukturēti vai nestrukturēti, (4) problēmsfērā ir raksturīgas daudzas klases, kuras sastopamas vienlīdz bieži, (5) objekts var piederēt vienlaicīgi vairākām klasēm.

Aprobācijai izmantotās problēmsfēras ir studiju priekšmetu salīdzināšana un diagnostika medicīnā. Daļa no eksperimentu rezultātiem atspoguļoti autores publikācijā [53].

Ar eksperimentu palīdzību tiek pārbaudīti šādi aspekti *InClaS* lietderības novērtēšanai:

- *nepareizi klasificēto piemēru skaita salīdzināšana*, izmantojot klasisko neinteraktīvo klasifikācijas pieeju un piedāvāto interaktīvo pieeju (attiecas uz T1);
- *piemērotākā pārliecības sliekšņa lieluma noteikšanas metodes* pārbaude, atrodot atbilstošāko pārliecības slieksni, kur nepareizi klasificēto piemēru skaits *N* ir minimāls pie izvirzītajiem eksperta ieguldāmā darba ierobežojumiem (attiecas uz T2).

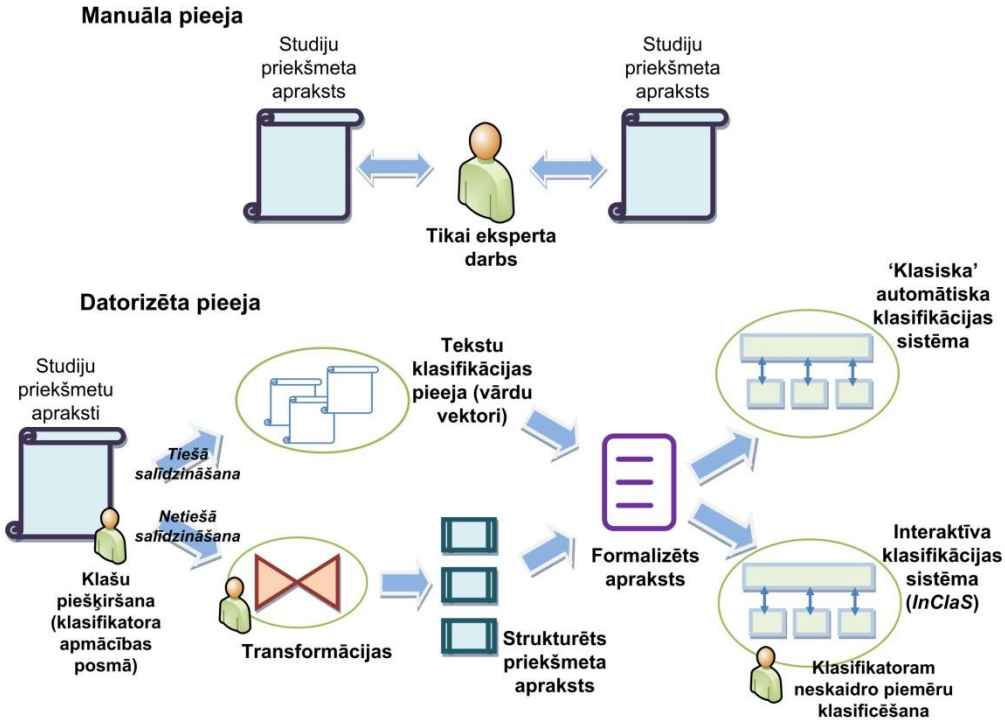
Attiecībā uz risinājuma lietderību izglītības sfērā:

- pārbaude darbā izteiktajam apgalvojumam, ka šī problēmsfēra nav piemērota tradicionāliem mašīnapmācības risinājumiem, bet uz *induktīvo apmācību balstīta, interaktīva, daudzkategoriju klasifikācijas sistēma studiju priekšmetu salīdzināšanas atbalstam* var dot pieņemamu risinājumu (attiecas uz T3);
- studiju priekšmetu *tiešās un netiešās salīdzināšanas novērtēšana* – izmantojot priekšmetu aprakstus pilnībā vai veicot pastarpinātu salīdzināšanu caur *e-CF* ietvara kompetencēm.

6.1. Eksperimenti izglītības jomā

6.1. attēlā ir shematiski parādīti vairāki veidi, kā studiju priekšmeti var tikt salīdzināti. Manuālā pieejā vienīgi cilvēks – eksperts lieto savas zināšanas problēmsfērā un spriež par priekšmetu atbilstību. Datorizētajās pieejās cilvēka sākotnēji ieguldītais darbs tiek izmantots automātisku vai interaktīvu klasifikācijas sistēmu izveidei. Attēlā atspoguļoti divi veidi

formalizēta un klasifikācijas algoritmiem piemērota ieejas datu formāta iegūšanai – (1) caur daļēji strukturētu priekšmetu aprakstu tekstu tiešu izmantošanu un (2) strukturētu semantiski nozīmīgu daļu izgūšanu no pieejamajiem priekšmetu aprakstiem, izmantojot pastarpinātu unifikācijas ietvaru. Vienā vai otrā veidā iegūto formalizēto aprakstu iespējams apstrādāt ar 'klasisku' neinteraktīvu klasifikācijas sistēmu vai ar šajā darbā piedāvāto interaktīvo klasifikācijas sistēmu. Tādējādi datorizētajai pieejai realizējamas 4 kombinācijas: (1) tekstu klasifikācija ar 'klasisko' klasifikāciju, (2) strukturētā aprakstu pieeja ar 'klasisko' klasifikāciju, (3) tekstu klasifikācija ar *InClaS* un (4) strukturētā aprakstu pieeja ar *InClaS*.



6.1. att. Studiju priekšmetu manuālas un automatizētas klasifikācijas iespējas

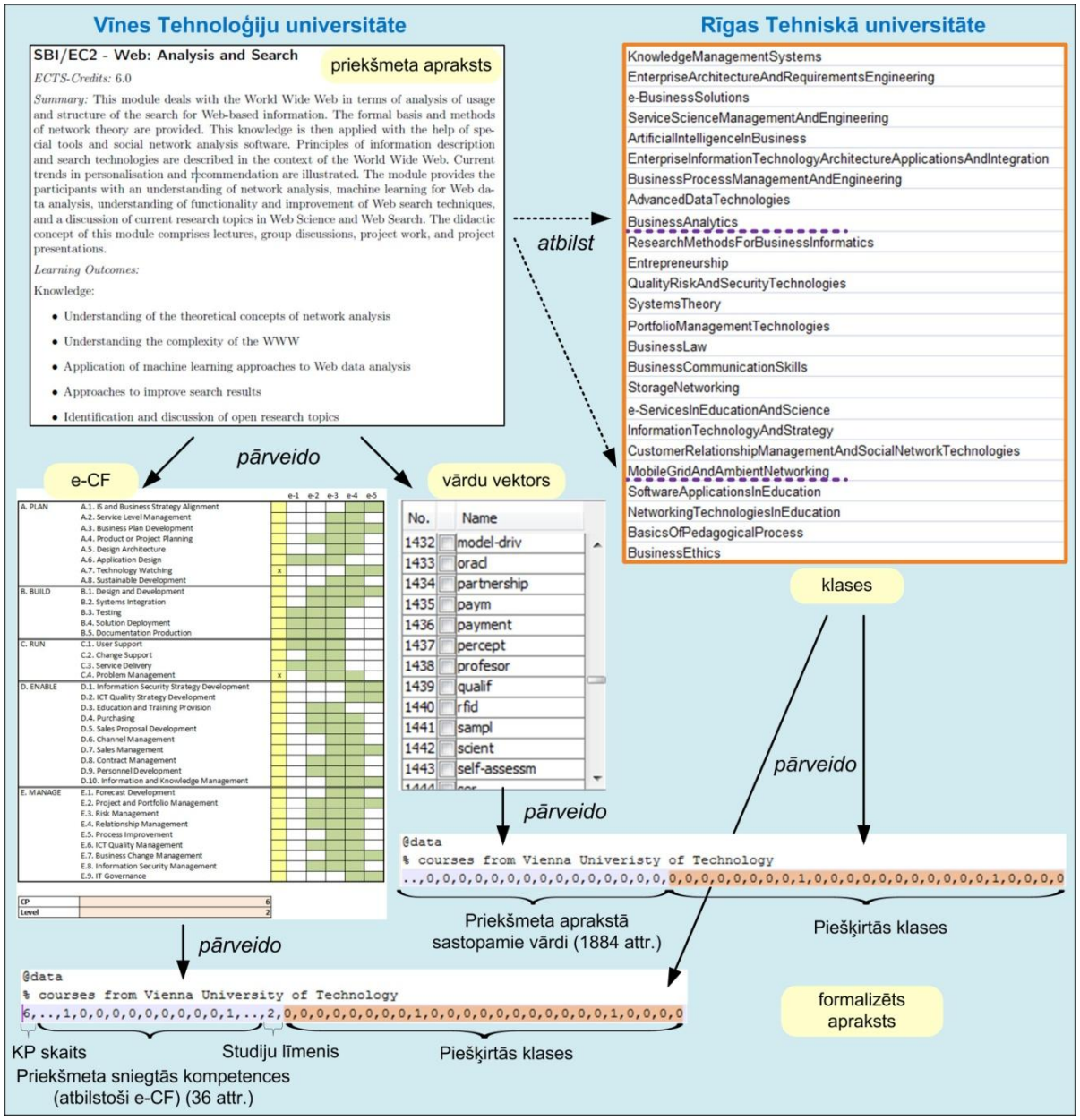
Formalizētais studiju priekšmetu apraksts sniegts 6.1. tabulā, savukārt studiju priekšmeta formalizēšanas piemērs, iegūstot Vīnes Tehnoloģiju universitātes priekšmeta atspoguļojumu aprakstu abos formalizācijas veidos un norādot tam 2 klases jeb nosakot tā atbilstību 2 studiju priekšmetiem no pavisam 25 Biznesa informātikas studiju programmas priekšmetiem Rīgas Tehniskajā universitātē, demonstrēts 6.2. attēlā.

6.1. tabula

Atribūti un klases tiešās un netiešās studiju priekšmetu salīdzināšanas gadījumā

Atribūti a	**Iespējamās vērtības** v_a	**Datu tips**	
Atribūti netiešās studiju priekšmetu salīdzināšanas gadījumā *(n = 38)*			
Kredītpunktu skaits (ECTS)	[3; 6; 9; 15]	Nominmāls	
Studiju līmenis	[bakalaura; maģistra]	Nominmāls	
Kompetence A1	[0; 1]	Nominmāls	36 atribūti
Kompetence A2	[0; 1]	Nominmāls	
..	..	..	
Kompetence E9	[0; 1]	Nominmāls	

Atribūti tiešās studiju priekšmetu salīdzināšanas gadījumā - vārdi priekšmetu apraksts *(n = 1884)*			
attrib	[0; 1]	Nominmāls	1884 atribūti
compliance	[0; 1]	Nominmāls	
..			
model-diven	[0; 1]	Nominmāls	
Klases *K (j = 25)*			
Biznesa analītika	[0; 1]	Nominmāls	25 klases
Zināšanu vadības sistēmas	[0; 1]	Nominmāls	
..	..	..	



6.2. att. Priekšmetu apraksta formalizēšanas piemērs

Studiju priekšmetu salīdzināšanas gadījumā izpildāmo eksperimentu vispārīgs plāns ir atspoguļots 6.2. tabulā. Pilna izmantotā datu kopa, kas balstīta uz kompetencēm, sastāv no 79 piemēriem, kas raksturo dažādu Eiropas universitāšu studiju priekšmetu atbilstību RTU *Biznesa informātikas* priekšmetiem. Samazinātajā datu kopā saglabāti piemēri tikai tām

klasēm, kas aprakstītas ar vismaz 4 piemēriem. Datu kopu parametri ir apkopoti 6.3. tabulā. Klašu blīvums (ang. v. - *density*) norāda vidējo viena piemēra klašu skaitu pret kopējo klašu skaitu. Klašu kardinalitāte (ang. v. - *cardinality*) norāda vidējo viena piemēra klašu skaitu. Savukārt klašu kopu (ang. v. - *distinct labelsets*) skaits norāda, cik daudz atšķirīgu klašu kombināciju ir datu kopā. Šie mēri ļauj spriest par datu kopas daudzkategoritātes iezīmēm.

6.2. tabula

Eksperimentu plāns studiju priekšmetu salīdzināšanai

	1. variants	**2. variants**	**3. variants**	**4. variants**
Ieejas datu kopa	Studiju priekšmetu apraksti pilnā apjomā (vārdu vektori)	Studiju priekšmetu kompetences, kredītpunktu skaits, studiju līmenis	Studiju priekšmetu apraksti pilnā apjomā (vārdu vektori)	Studiju priekšmetu kompetences, kredītpunktu skaits, studiju līmenis
Klasifikācijas pieeja	Automātiska klasifikācija	Automātiska klasifikācija	Interaktīva klasifikācija (*InClaS* modelis)	Interaktīva klasifikācija (*InClaS* modelis)
Izmantotās metodes	20 klasifikācijas metodes (no programmatūras *Weka* un bibliotēkas *Mulan* klāsta)	20 klasifikācijas metodes	4 labākās metodes no 1. v. eksperimentiem	4 labākās metodes no 2. v. eksperimentiem
Novērtēšanas mēri	Haminga zaudējums (*Hamming loss*), Mikro-vidējā precizitāte (*Micro-average precision*), Mikro-vidējais atsaukums (*Micro-average recall*), Viena kļūda (*One-error*), Pārklāšana (*Coverage*)	Haminga zaudējums, Mikro-vidējā precizitāte, Mikro-vidējais atsaukums, Viena kļūda, Pārklāšana	*DP, N, ĪN, NN* skaits	*DP, N, ĪN, NN* skaits

6.3. tabula

Studiju priekšmetu datu kopa

	Atribūtu skaits	**Piemēru skaits**	**Klašu skaits**	**Klašu blīvums**	**Klašu kardinalitāte**	**Klašu kopu skaits**
Pilna datu kopa (vārdu vektori)	1884	131 (79)	25	0.0620	1.6203	52
Pilna datu kopa (kompetences)	38	79	25	0.0620	1.6203	52
Samazināta datu kopa (kompetences)	38	64	12	0.1341	1.6094	36

Galvenie eksperimentu rezultāti

6.4. tabulā redzami rezultāti 4 algoritmiem vārdu vektorus izmantojošajai datu kopai (3. eksperimentu variants). Izvēlēti 3 labākos rezultātus uzrādījušie algoritmi no 1. eksperimenta varianta (*RAkEL*, *AdaBoost* un *Bagging)*, kā arī *JRip*, jo tas sniedz skaidrus un lietotājam saprotamus likumus. *BR* ir saīsinājums no *Binary Relevance* jeb binārās saistības metodes. Starp mēriem pastāv šādas sakarības:

DP + Nepareizs (bez interaktivitātes)=1 (visi automātiskas klasifikācijas rezultāti).
DP+ĪN+NN+ Nepareizs (ar interaktivitāti)=1 (visi interaktīvas klasifikācijas rezultāti).
Nepareizs (bez interaktivitātes)= ĪN+NN+ Nepareizs (ar interaktivitāti).

6.4. tabula

Interaktīvas pieejas lietošana pilnai priekšmetu datu kopai (vārdu vektori)

	Daļēji pareizs (DP)	Īsti neskaidrs (ĪN)	Nepamatoti neskaidrs (NN)	Nepareizs (ar interaktivitāti)	Nepareizs (bez interaktivitātes)
RAkEL(J48)	0.267	0.333	0.000	0.400	0.733
BR(AdaBoost)	0.100	0.400	0.000	0.500	0.900
BR(Bagging)	0.067	0.600	0.000	0.333	0.933
BR(JRip)	0.267	0.367	0.000	0.367	0.733

Rezultāti 6.4. tabulā ir interpretējami šādi. Izmantojot klasisko klasifikāciju, kur klasifikators pilnībā pieņem lēmumu par klašu piešķiršanu, pilnīgi vai daļēji pareizi tiktu klasificēti 27% piemēru (*RAkEL* gadījumā), bet nepareizi – 73%. Ja tiek izmantota interaktivitāte, tad daļēji pareizo piemēru īpatsvars saglabājas tāds pats, bet 33% no iepriekš nepareizi klasificētajiem piemēriem tiek atzīti par klasifikatoram neskaidriem un nodoti ekspertam, atstājot nepareizi klasificētus 40%. Lai arī iegūtie rezultāti un sadalījums pa pozīcijām ļoti variējas katram algoritmam, tabulas rezultāti liecina, ka, neizmantojot interaktivitāti, nepareizi klasificēto piemēru skaits jebkuram algoritmam ir daudz lielāks nekā gadījumā, ja sistēmai tiek dota iespēja identificēt neskaidrās klasifikācijas un atsijāt tās no nepareizi klasificēto piemēru klāsta. Jāpiemin, ka eksperimentu rezultātu interpretācijā tiek izmantots pieņēmums, ka eksperta sniegtās klasifikācijas ir pareizas. Kopumā var spriest, ka datu kopa nesniedz pētāmā koncepta pilnīgu aprakstu, kas arī tika pieņemts, uzsākot darbu šajā problēmsfērā.

6.5. tabulā sniegti apkopoti rezultāti 4. varianta eksperimentiem.

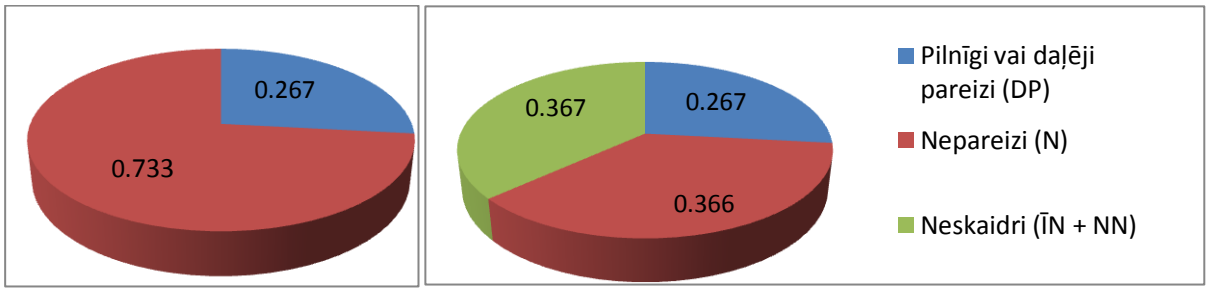
6.5. tabula

Interaktīvas pieejas lietošana pilnai priekšmetu datu kopai (kompetences)

	Daļēji pareizs (DP)	Īsti neskaidrs (ĪN)	Nepamatoti neskaidrs (NN)	Nepareizs (ar interaktivitāti)	Nepareizs (bez interaktivitātes)
BR(NB)	0.234	0.633	0.000	0.133	0.766
BR(Bagging)	0.167	0.733	0.000	0.100	0.833
BR(AdaBoost)	0.267	0.433	0.000	0.300	0.733
BR(JRip)	0.267	0.367	0.000	0.366	0.733

Līdzīgi kā 3. eksperimenta rezultātos, interaktivitāte visiem atlasītajiem algoritmiem ļauj samazināt nepareizi klasificēto piemēru skaitu, atsijājot neskaidros piemērus un nododot tos vērtēšanai ekspertam. Vēl uzskatāmāk tas redzams 6.3. attēlā, kur parādīts *JRip* algoritma rezultāts. Neizmantojot interaktivitāti (6.3. attēla A daļa), nepareizi klasificēti būtu visi klasifikatoram neskaidrie piemēri, iegūstot tikai 27% daļēji pareizu klasifikācijas rezultātu. Šādai klasifikācijas sistēmai tiešām nav vērts uzticēties. Savukārt, izmantojot interaktīvu pieeju (6.3. attēla B daļa) un nosakot neskaidri klasificētos piemērus, trešdaļu piemēru iespējams atzīt par klasifikatoram neskaidriem un pāradresēt izskatīšanai ekspertam. Tādā

veidā nepareizi klasificēti tiek 37% piemēru, kas, protams, arī nav precizitātes ziņā izcils rezultāts, tomēr ir ievērojami labāks par 73% nepareizi klasificētu piemēru.



A **B**

6.3. att. JRip algoritma klasifikācijas rezultātu atspoguļojums priekšmetu salīdzināšanas uzdevumā ar automātisku (A) un interaktīvu (B) klasifikāciju

Ir veikti papildu eksperimenti un salīdzināti pilnās un samazinātās datu kopas rezultāti, imitējot situāciju, kad klasifikatora pieredze jomā ir pieaugusi. Tie ļauj secināt, ka interaktīva klasifikācijas sistēma, kas šajā problēmsfērā sākotnēji sniedz daļēji pilnvērtīgus klasifikācijas rezultātus, darbojas labāk un vēršas pie eksperta arvien retāk, pieaugot apmācības piemēru skaitam. Tātad ir lietderīgi ieguldīt lielāku eksperta darbu sistēmas izmantošanas sākumposmā, lai klasificētu klasifikatoram neskaidros piemērus un ar savām zināšanām papildinātu klasifikatoru, tādējādi uzlabojot klasifikācijas rezultātus nākotnē.

No eksperimentu rezultātiem var secināt arī, ka nav apstiprinājies pieņēmums par sasniedzamo mācību rezultātu izmantošanas lietderības pārākumu salīdzinājumā ar pilnu aprakstu lietošanu. Kompetences ir izmantojamas kā priekšmetu raksturojoši atribūti, bet šobrīd tās nav viegli tiešā veidā iegūstamas un jāiegulda liels eksperta darbs, lai studiju priekšmetus atspoguļotu vienotā formātā, piemēram, *e-CF*. Studiju priekšmetu nestrukturēti vai daļēji strukturēti apraksti, kas pārvērsti vārdu vektoros, ir pietiekami labi izmantojami klasifikatoru iegūšanai, tomēr kā negatīvs aspekts pilnu tekstu izmantošanā jāmin iegūto likumu kopas mazā semantiskā jēga problēmsfēras ekspertam. Tādējādi vairāk lietderīgu zināšanu par sakarībām datu kopā sniedz jēgpilni ieejas dati, šajā gadījumā – kompetences.

6.2. Eksperimenti piemērotākā pārliecības sliekšņa noteikšanai

Pārliecības sliekšņa ietekme uz klasifikācijas rezultātiem ir pārbaudīta studiju priekšmetu salīdzināšanas datu kopai, kas balstīta uz kompetencēm, izmantojot autores piedāvāto piemērotākā pārliecības sliekšņa noteikšanas metodi. Viens no analizētajiem piemēriem ir šāds. Tiek pieņemts, ka ir uzstādīti vairāki ierobežojumi, kuri jāievēro piemērotākā sliekšņa lieluma izvēlē. *A* gadījumā noteikts kopējais piemēru skaits, ko eksperts ir ar mieru klasificēt, *B* gadījumā noteikts lietderīgais darbs, bet *C* gadījumā ierobežojumu

pret ieguldīto darbu nav – jāizvēlas labākais slieksnis (apzīmēts ar **o**) nepareizi klasificēto piemēru samazināšanai.

A) $D_{kopējais} \leq 5$ (uz 10 klasificējamiem piemēriem)

B) $D_{nelietderīgais} \leq 0.5$

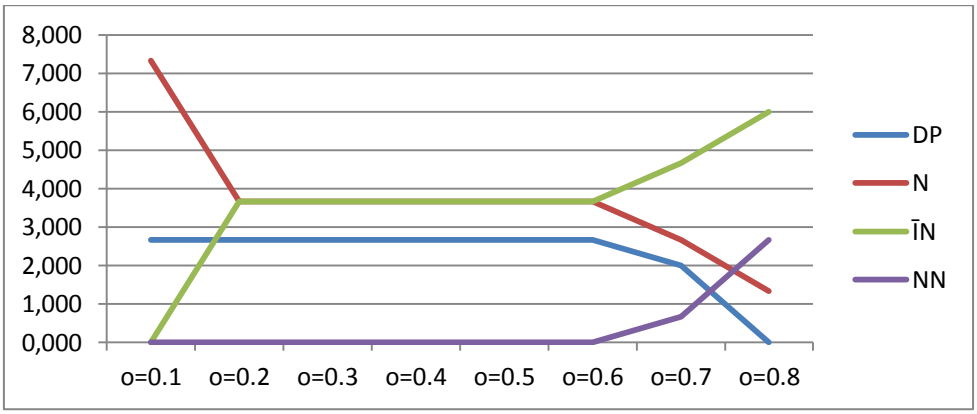
C) Labākais iespējamais (minimāls nepareizi klasificēto (*N*) piemēru skaits).

6.6. tabula un 6.4. attēls sniedz vidējos rezultātus pēc 3-kāršas datu kopas sadalīšanas pilnai priekšmetu datu kopai. *DP, N, ĪN*, un *NN* ir norādīti relatīvi pret testa kopas apjomu.

6.6. tabula

Novērtējamie parametri sliekšņa izvēlei pilnā datu kopā

	o=0.1	**o=0.2**	**o=0.3**	**o=0.4**	**o=0.5**	**o=0.6**	**o=0.7**	**o=0.8**
DP	0.267	0.267	0.267	0.267	0.267	0.267	0.200	0.000
N	0.733	0.366	0.366	0.366	0.366	0.366	0.266	0.133
ĪN	0.000	0.367	0.367	0.367	0.367	0.367	0.467	0.600
NN	0.000	0.000	0.000	0.000	0.000	0.000	0.067	0.267
$D_{nelietderīgais}$	-	0.000	0.000	0.000	0.000	0.000	0.250	0.526
$D_{kopējais}$	0.000	3.667	3.667	3.667	3.667	3.667	5.333	8.667



6.4. att. Grafisks parametru atspoguļojums (X ass – klasifikatora pārliecība, Y ass – piemēru skaits)

No rezultātiem var secināt, ka labākie sliekšņa lielumi, atbilstoši kritērijiem ir šādi:

A) $D_{kopējais} \leq 5$ (uz 10 klasificējamiem piemēriem): o=0.6. Var paplašināt meklēšanu, izvēršot sīkākus soļus starp sliekšņiem 0.6 un 0.7.

B) $D_{nelietderīgais} \leq 0.5$: o=0.7. Var paplašināt meklēšanu, izvēršot sīkākus soļus starp sliekšņiem 0.7 un 0.8. No grafika redzams, ka visi stāvokļi no o=0.2 līdz o=0.6 ir ekvivalenti.

C) Labākais iespējamais (minimāls nepareizi klasificēto piemēru (*N*) skaits): o=0.8

6.3. Eksperimenti medicīnas jomā

Papildus izglītības sfērai, interaktīvās pieejas pārbaudei ir izmantota arī medicīnas datu kopa, kas apraksta pacientu stāvokli, piemēroto terapiju un diagnozi (vai vairākas diagnozes) *ICD-9-CM* kodu veidā. Dati ir publiskoti *Computational Medicine Center's 2007 Medical Natural Language Processing Challenge* [54] ietvaros. Atšķirībā no uzdevuma

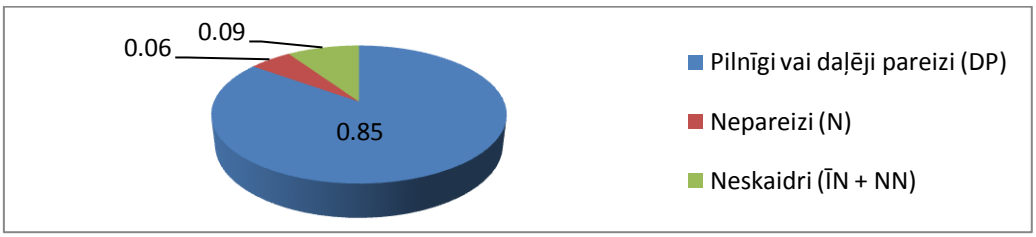
izglītības jomā, šai datu kopai nav raksturīgs tik mazs apmācības piemēru skaits. Tomēr tas nemazina interaktīvās metodes lietošanas *iespējamību* (jo izpildās eksperta pieejamības un problēmas apraksta saprotamības nosacījums), tikai samazina tās *nepieciešamību*, jo arī automātiskas klasifikācijas metodes šeit darbojas pieņemami. Datu kopu raksturojošie parametri atspoguļoti 6.7. tabulā.

6.7. tabula

Medicīnas datu kopa

	Atribūtu skaits	**Piemēru skaits**	**Klašu skaits**	**Klašu blīvums**	**Klašu kardinalitāte**	**Klašu kopu skaits**
Datu kopa	1449	978	45	0.028	1.245	94

Datu kopas sadalījums apmācības un testa kopā saglabāts atbilstoši oriģinālajiem datiem [54]. Izmantojot binārās saistības metodi ar *JRip* algoritmu, kurš uzrādīja labākus rezultātus par citiem izmantotajiem, medicīnas datu kopai iegūti 6.5. attēlā redzamie rezultāti. Lietots noklusētais pārliecības sliekšņa lielums - 0.5.



6.5. att. JRip algoritma klasifikācijas rezultātu atspoguļojums medicīnas uzdevumā

Kā liecina 6.5. att., šajā medicīnas datu gadījumā DP klasifikāciju skaits ir ievērojami lielāks kā studiju priekšmetu uzdevumā – 85%. Tas tā varētu būt, pateicoties daudz apjomīgākai apmācības kopai. Tomēr arī šeit interaktīva pieeja var palīdzēt uzlabot apmācības rezultātu, atklājot vēl 9% piemēru, kas ir klasifikatoram neskaidri un tiktu klasificēti nepareizi, ja lietotu automātisku klasifikāciju. Jāpiemin, ka eksperimentu mērķis nav bijis atrast labāko klasifikatoru šai datu kopai, bet gan demonstrēt, ka automātiskā klasifikācijā iegūtos rezultātus ir iespējams uzlabot, ja tiek izmantota interaktīvā pieeja.

Veiktie eksperimenti pierāda piedāvātā interaktīvā klasifikācijas sistēmas modeļa lietderību un apstiprina eksperimentāli pārbaudāmos aspektus. *InClaS* prototipa lietojamība apstiprinājās eksperimentu veikšanas gaitā. Gan universitāšu priekšmetu salīdzināšanā, gan medicīnas diagnostikas uzdevumā nepareizi klasificēto piemēru skaitu iespējams samazināt, ja tiek izmantota piedāvātā interaktīvā pieeja. Studiju priekšmetu salīdzināšanas gadījumā nepareizi klasificēto piemēru skaita starpība (starp automātisku un interaktīvu pieeju) ir tik būtiska, ka interaktīvas pieejas lietošana paver iespēju izmantot mašīnapmācības metodes, kas bez interaktivitātes ieviešanas šajā jomā nesniedz apmierinošus rezultātus.

GALVENIE REZULTĀTI UN SECINĀJUMI

Promocijas darbā ir izstrādāts *InClaS* modelis, kas definē algoritmus, metodes un citas komponentes, kas ļauj izstrādāt interaktīvu klasifikācijas sistēmu nepareizi klasificēto objektu skaita samazināšanai jomās, kur pieejams cilvēks – eksperts. Modeļa pārbaudei ir izplānoti un veikti eksperimenti divās problēmsfēras – izglītībā un medicīnā –, kas pierāda, ka nepareizi klasificēto piemēru skaitu iespējams samazināt, ja klasifikatoram neskaidrie (t.i., neklasificētie un nepārliecinoši klasificētie) piemēri tiek atlasīti un nodoti eksperta izvērtēšanai. Sevišķi nozīmīgi ir ieguvumi, ja klasifikācijas rezultāti, izmantojot 'klasisku' neinteraktīvu klasifikatoru, ir nepieņemami vāji, kā tas ir studiju priekšmetu salīdzināšanas uzdevumā, kur bez interaktīvās pieejas klasifikators sniedz vairāk nepareizu klasifikācijas lēmumu nekā pareizu. Līdz ar to var secināt, ka ir sasniegts darba mērķis - *izstrādāt automatizētas klasifikācijas sistēmas modeli, kas pieļauj interaktivitāti ar ekspertu klasifikatora lietošanas laikā, ja klasifikators sastopas ar objektu, ko tas nespēj klasificēt vai nav pārliecināts par sava lēmuma pareizību* - un ir iespējams izteikt **rekomendācijas par *InClaS* lietošanu**.

Interaktīvas klasifikācijas sistēmas lietošana *ir iespējama* jomās, kur:

- cilvēks – eksperts ir pieejams un var sniegt klasifikāciju atsevišķiem piemēriem;
- problēmsfēras definēšanai tiek izmantoti ekspertam saprotami atribūti, kuru skaits nav pārāk liels, vai objekta aprakstu iespējams iegūt interpretējamā formā.

Interaktīvas klasifikācijas pieeja *ir piemērotāka* par klasisku automātisku klasifikāciju jomās, kur izpildās vismaz viens no apstākļiem:

- ir būtiski iegūt pareizu klasifikāciju pēc iespējas vairāk piemēriem, un tā sasniegšanai ir pieņemami ieguldīt eksperta darbu un laiku;
- ir grūti izgūt vai definēt raksturīgās iezīmes, kā rezultātā atribūti neapraksta pētāmo konceptu pilnīgi;
- ir pieejama tikai neliela sākotnējā apmācības kopa, un pastāv aizdomas, ka tā nav pietiekami reprezentabla.

Darba teorētiskie rezultāti

Darba izstrāde devusi teorētiskos rezultātus, kurus iespējams grupēt sekojoši:

- Izstrādāts interaktīvas klasifikācijas sistēmas *InClaS* modelis, kas apvieno interaktīvas klasifikācijas sistēmas radīšanai nepieciešamās komponentes:

- Izstrādāta realizējamā interaktivitātes shēma, kas parāda atšķirības attiecībā pret 'klasisku' neinteraktīvu klasifikācijas pieeju.
- Izstrādāta interaktīvas klasifikācijas sistēmas vispārīga struktūra - klasifikācijas sistēmas funkcionālie moduļi un to sasaistes, kā galvenos moduļus izdalot *Datu apstrādes, Lietotāja saskarnes, Klasifikatora veidošanas, Klasifikatora lietošanas* un *Interaktivitātes moduli*.
- Izstrādātas divas klasifikatora atjaunošanas (papildināšanas) shēmas pēc eksperta veiktas klasifikācijas – *Uz slieksni balstītā statiskās apmācības pieeja* un *Dinamiskās apmācības pieeja*.

- Izstrādāts *InClaS* modeļa papildinājums, kas apvieno interaktīvas daudzkategoriju klasifikācijas sistēmas radīšanai nepieciešamās komponentes:
  - Izstrādāts algoritms klasifikatoram neskaidru piemēru noteikšanai daudzkategoriju klasifikācijas gadījumā.
  - Izstrādāta metode atbilstošākā pārliecības sliekšņa noteikšanai, pie kura algoritma klasificētos piemērus atzīt par klasifikatoram neskaidriem un nodot eksperta pārziņā.
  - Ieviesti un pamatoti vairāki mēri daudzkategoriju klasifikācijas novērtēšanai – *vidējā klasifikatora pārliecība par klasēm, kurām piemēri ir piederīgi* (*VPP*) un *nav piederīgi* (*VPN*), eksperta ieguldītā darba mēri: $D_{nelietderīgais}$ - cik pareizi klasificētu piemēru ekspertam jācaurskata, lai klasificētu vienu nepareizi klasificētu piemēru, $D_{kopējais}$ - cik piemēru ekspertam tiek lūgts klasificēt un jēdzieni *Daļēji pareizi vai pilnīgi pareizi klasificēts piemērs* (*DP*), *Nepareizi klasificēts piemērs* (*N*), *Īsti neskaidra klasifikācija* (*ĪN*), *Nepatiesi neskaidra klasifikācija* (*NN*).
  - Ir adaptēta piecu soļu metode intelektuālu sistēmu projektēšanā [42], kas atvieglo analītisko darbu, ieviešot interaktīvo klasifikācijas sistēmu konkrētā problēmsfērā.
  - Veikts interaktīvas daudzkategoriju klasifikācijas sistēmas projektējums sistēmu veidojošo moduļu, to ieeju un izeju apraksta veidā.
- Oriģinālapkopojumi līdzšinējo darbu analīzes rezultātā:
  - Izglītības dokumentu datorizētas salīdzināšanas risinājumu apskats.
  - Induktīvās apmācības algoritmu klasifikācija pēc dažādiem parametriem.
  - Esošo interaktīvo klasifikācijas pieeju sistematizācija un salīdzinājums.
  - Klasifikācijas sistēmu arhitektūru apkopojums un salīdzinājums.

Darba praktiskie rezultāti

Darba izstrāde ļāvusi sasniegt šādus praktiskos rezultātus:

- Izstrādāts interaktīvas klasifikācijas sistēmas prototips daudzkategoriju klasifikācijas uzdevumam, kurš pielāgots studiju priekšmetu salīdzināšanai.
- Radīta utilītprogramma datņu sintaktiskai pārveidošanai no vienkategorijas apraksta formas uz daudzkategoriju (*.arff* formāta datnēm), kas ir praktiski izmantojama arī citos uzdevumos.
- Noteikts priekšmetu atbilstības atspoguļojums starp Rīgas Tehniskās universitātes maģistra studiju programmas *Biznesa informātika* priekšmetiem un vairāku Eiropas universitāšu atbilstošās nozares priekšmetiem.

Turpmākajos pētījumos risināmās problēmas

Promocijas darbā aplūkotajai tēmai ir plašs tālāko pētījumu potenciāls gan no izstrādātā klasifikācijas modeļa pilnveidošanas, gan galvenās risinātās problēmsfēras puses. Šeit minētas tikai dažas no attīstības iespējām.

- Atbilstības definēšanai studiju priekšmetu salīdzināšanas uzdevumā būtu ieteicams izmantot detalizētāku un precīzāku aprēķinu par šobrīd izmantoto kategorisko iedalījumu *atbilst* vai *neatbilst*, piemēram, mainīt priekšmetu svaru atbilstības noteikšanas laikā.
- Turpmākā studiju priekšmetu salīdzināšanas risinājuma pilnveidošanā būtu lietderīgi izskatīt kopīgo apmācību (ang.v. - *co-training*) attiecībā uz studiju priekšmetu tekstuālo aprakstu un strukturēto kompetenču izmantošanu, kā arī transduktīvo vai daļēji pārraudzīto apmācību, kas izmanto gan klasificētus, gan neklasificētus piemērus apmācības uzlabošanai. Papildus apskatāms ir gadījumos sakņotas spriešanas (ang. v. - *case-based reasoning*) izmantošanas potenciāls.
- Ir ieteicams izstrādāt lietotājam pieņemamu risinājumu liela atribūtu skaita gadījumā. Pieeja atribūtu loģiskai grupēšanai un sakārtošanai varētu atvieglot informācijas uztveramību ekspertam, ja klasificējamiem piemēriem ir liels skaits atribūtu.
- Papildinājumu var sniegt sīkāk definēts daļēji pareizas klasifikācijas mērs daudzkategoriju uzdevumu novērtējumā.

BIBLIOGRĀFISKAIS SARAKSTS

1. Cios K.J., Kurgan L.A., Hybrid Inductive Machine Learning: An Overview of CLIP Algorithms, In New Learning Paradigms in Soft Computing. 2002, Physica-Verlag GmbH: Heidelberg, Germany. pp. 276-321.
2. Witten I.H., Frank E. Data Mining: Practical Machine Learning Tools and Techniques. 3rd ed. 2011: Morgan Kaufmann. 629 p.
3. Hall M., Frank E., Holmes G., Pfahringer B., Reutemann P., and Witten I.H. The WEKA Data Mining Software: An Update. SIGKDD Explorations, 2009. Vol. 11, pp. 10-18.
4. Tsoumakas G., Spyromitros-Xioufis E., Vilcek J., Vlahavas I. Mulan: A Java Library for Multi-Label Learning. Journal of Machine Learning Research, 2011. Vol. 12, pp. 2411-2414.
5. Alves H., Figueira Á. A Educational Library based on Clusters of Semantic Proximity. In IADIS European Conference on Data Mining. 2011. pp. 226-228.
6. Anohina-Naumeca A., Graudiņa V., Grundspeņķis J. Curricula Comparison Using Concept Maps and Ontologies. In Proceedings of the 5th International Scientific Conference on Applied Information and Communication Technology. 2012: Latvia, Jelgava. pp. 177-183.
7. Birzniece I., Kirikova M. Interactive Inductive Learning Service for Indirect Analysis of Study Subject Compatibility. In BeneLearn 2010. Belgium, Leuven: Katholieke Universiteit Leuven pp. 1-6.
8. Biletskiy Y., Brown J.A., Ranganathan G. Information extraction from syllabi for academic e-Advising. Expert Systems with Applications, 2009. Vol. 36(3, Part 1), pp. 4508-4516.
9. Biletska O., Biletskiy Y., Li H., Vovk R. A semantic approach to expert system for e-Assessment of credentials and competencies. Expert Systems with Applications, 2010. Vol. 37(10), pp. 7003-7014.
10. Ranganthan G.R., Biletskiy Y., MacIsaac D. Machine Learning for Classifying Learning Objects. IEEE CCECE/CCGEI, 2006, pp. 280-283.
11. Rudzājs P., Kirikova M. Towards Monitoring Correspondence Between Education Demand and Offer. In 21st International Conference on Information Systems Development (ISD2011). 2012: Italy, Prato. pp. 1-12.
12. Rudzājs P., Kirikova M. IT Knowledge Requirements Identification in Organizational Networks: Cooperation between Industrial Organizations and Universities. In Proceedings of the 18th International Conference on Information Systems Development (ISD 2009). 2009: China, Nanchang. pp. 187-199.
13. European e-Competence Framework 2012. Available from: http://www.ecompetences.eu/).
14. Aksoy M.S. Applications of RULES-3 Induction System. In Proceedings of the Innovative Production Machines and Systems. 2008.
15. Valeskalne I. Induktīvās spriešanas metodes un to pielietojums. 2007, Bakalaura darbs, RTU.

16. Birzniece I. Induktīvo apmācības metožu pielietojums tēlu pazīšanā. 2009, Maģistra darbs, RTU.
17. Boutell M.R., J. Luo X.S., Brown C.M. Learning multi-label scene classification. Pattern Recognition, 2004. Vol. 37(9), pp. 1757-1771.
18. Sorower M.S. A Literature Survey on Algorithms for Multi-label Learning, Oregon State University, Corvallis, 2010.
19. Tsoumakas G., Katakis I., Vlahavas I., Mining Multi-label Data, in Data Mining and Knowledge Discovery Handbook, O. Maimon and L. Rokach, (Eds.). 2010, Springer.
20. Fan R.-E., Lin C.-J. A Study on Threshold Selection for Multi-label Classification, National Taiwan University, 2007.
21. Li T., Zhu C.Z.S. Empirical Studies on Multi-label Classificatio. In Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence ICTAI '06. 2006. pp. 86-92.
22. Gao W., Zhou Z.-H. On the consistency of multi-label learning. In Proceedings of the 24th Annual Conference on Learning Theory (COLT'11). 2011: Budapest, Hungary. pp. 341-358.
23. Michalski R.S., Mozetic I., Hong J., Lavrac N. The multipurpose incremental learning system AQ15 and its testing application to three medical domains. In 5th National Conference on Artificial Intelligence. 1986, Morgan-Kaufmann: San Francisco. pp. 1041-1045.
24. Clark P., Niblett T. The CN2 Induction Algorithm. Machine Learning Journal, 1989(3), pp. 261-283.
25. Birzniece I. From Inductive Learning towards Interactive Inductive Learning. Scientific Journal of Riga Technical University. Computer Sciences. - Applied Computer Systems 2010. Vol. 41, pp. 106-112.
26. Birzniece I. Interactive Use of Inductive Approach for Analyzing and Developing Conceptual Structures. In Sixth International Conference on Research Challenges in Information Science (RCIS 2012). 2012, IEEE.
27. Okabe M., Yamada S., Interactive Web Page Retrieval, in Active Mining: New Directions of Data Mining. 2002, OS Press: Amsterdam pp. 31-40.
28. Tanumara R.C., Xie M., Au C.K. Learning Human-Like Color Categorization through Interaction. International Journal of Computational Intelligence, 2007. Vol. 4, pp. 338-345.
29. Buntine W., Stirling D., Interactive Induction, in Machine Intelligence: Towards An Automated Logic of Human Thought, J.E. Hayes, D. Michie, and Ė. Tyugu, (Eds.). 1991, Clarendon Press: New York pp. 121-137.
30. Hadjimichael M., Wasilevska A. Interactive Inductive Learning. International Journal of Man-Machine Studies, 1993(2), pp. 147-167.
31. Wong M.L., Laung K.S. Data Mining Using Grammar-Based Genetic Programming and Applications. 2000, USA: Kluwer Academic Publishers. 228 p.
32. Li X., Feng L., Zhou L., Shi Y. Learning in an Ambient Intelligent World: Enabling Technologies and Practices. IEEE Transactions on Knowledge and Data Engineering, 2009. Vol. 21(6), pp. 910-924.

33. Settles B. Active Learning Literature Survey, University of Wisconsin–Madison, 2010.
34. Brian R.G., Compton P. Induction of ripple-down rules applied to modelling large databases. Journal of Intelligent Information Systems, 1995. Vol. 5(4), pp. 211-228.
35. Mitchell T. Machine Learning. 1997: McGraw Hill. 414 p.
36. Cherkassky V., Mulier F. Learning from Data: Conceptrs, Theory, and Methods. 2nd ed. 2007: John Wiley & Sons. 538 p.
37. Dowdy S.M., Wearden S. Statistics for research. 2nd ed. ed. 1991, New York: Wiley 629 p.
38. Duda R.O., Hart P.E., Stork D.G. Pattern Classification. 2nd ed. 2001: Wiley - Interscience. 654 p.
39. Theodoris S., Koutrumbas K. Pattern Recognition. 3rd ed. 2006: Elsevier. 837 p.
40. Verdenius F., Someren M.W.v. Applications of inductive learning techniques: a survey in the Netherlands. AI Communications. 1997, IOS Press. pp. 3-20.
41. Bradley C.E., Smyth P. The process of applying machine learning algorithms. In Applying Machine Learning in Practice IMLC-95. 1998. Tahoe city, CA.
42. Bielawski L., Lewand R. Intellignet Systems Design: Integrating Expert Systems, Hypermedia, and Database Technologies. 1991: John Wiley & Sons. 302 p.
43. DTI. Neural Computing. "Department of Trade and Industry" ed. 1994: Learning Solutions p.
44. Han J., Kamber M. Data Mining: Concepts and Techniques. 2nd ed. The Morgan Kaufmann Series in Data Management Systems. 2005: Elsevier. 743 p.
45. Birzniece I. Architecture of an Interactive Classification System. In The Fifth International Conference on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services (CENTRIC 2012). 2012, IARIA: Lisbon, Portugal. pp. 91-100.
46. Birzniece I. Interactive Inductive Learning System: The Proposal. In Proceedings of the Ninth International Baltic Conference Baltic DB&IS 2010. 2010. Latvia, Riga: University of Latvia Press, pp. 245-260.
47. Birzniece I. Interactive Inductive Learning System. In Selected papers from the DB&IS 2010. 2010. Latvia, Riga: IOS Press, pp. 380-393.
48. Birzniece I. Interactive Inductive Learning Based Study Course Comparison. In Proceedings of the Rethinking Education in the Knowledge Society. 2011. Switzerland, Ascona, pp. 339-347.
49. Tsoumakas G., Katakis I. Multi-Label Classification: An Overview. International Journal of Data Warehousing and Mining, 2007. Vol. 3, pp. 1-13.
50. Birzniece I. Interactive Inductive Learning Based Classification System. In Proceedings of the IADIS International Conference Intelligent Systems and Agents (ISA 2011). 2011, IADIS. pp. 112-116.
51. Birzniece I., Kirikova M. Interactive Inductive Learning: Application in Domain of Education. Scientific Journal of Riga Technical University. Computer Sciences. - Applied Computer Systems, 2011. Vol. 47, pp. 57-64.

52. Birzniece I., Rudzajs P. Machine Learning based Study Course Comparison. In Proceedings of the IADIS International Conference on Intelligent Systems and Agents 2011 (ISA 2011). 2011. IADIS, pp. 107-111.
53. Birzniece I. Machine Learning Approach for Study Course Comparison. In International Conference on Machine Learning and Data Mining (MLDM 2012). 2012. Berlin, Germany: IBai, pp. 1-13.
54. Computational Medicine Center's 2007 Medical Natural Language Processing Challenge. 2007. Available from: http://computationalmedicine.org/challenge/previous (accessed 2012).