RĪGAS TEHNISKĀ UNIVERSITĀTE

Datorzinātnes un informācijas tehnoloģijas fakultāte
Datorvadības, automātikas un datortehnikas institūts

Jurijs KORŅIJENKO

Automātikas un datortehnikas doktora programmas Lēmumu atbalsta sistēmu virziena doktorants

KLASIFIKĀCIJAS LIKUMU AUTOMĀTISKĀS ĢENERĀCIJAS METOŽU IZSTRĀDE UN IZPĒTE

Promocijas darba kopsavilkums

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

Rīga 2007

UDK 004.85(043.2)
Ko 735 k

J. Korņijenko. Klasifikācijas likumu automātiskās ģenerācijas metožu izstrāde un izpēte. Promocijas darba kopsavilkums. – R:RTU, 2007.– 26 lpp.

Iespiests saskaņā ar DAD institūta 2007. gada 25. maija sēdes lēmumu, protokols Nr. 77.

ISBN 9984-32-xxx-x

PROMOCIJAS DARBS
IZVIRZĪTS RĪGAS TEHNISKAJĀ UNIVERSITĀTĒ
INŽENIERZINĀTŅU DOKTORA GRĀDA IEGŪŠANAI

Promocijas darbs inženierzinātņu doktora grāda iegūšanai tiek publiski aizstāvēts 2007. gada ..... Rīgas Tehniskās universitātes Datorzinātnes un informācijas tehnoloģijas fakultātē, Meža ielā 1, ... auditorijā.

OFICIĀLIE RECENZENTI

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

Asoc.prof., Dr.sc.ing. Irina Jackiva
Transporta un sakaru institūts

Profesors, tehnisko zinātņu doktors Viktors Romaņenko
Ukrainas Nacionāla Tehniskā Universitāte „Kijevas Politehniskais Institūts"

APSTIPRINĀJUMS

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

Jurijs Korņijenko ..... (paraksts)

Datums: .....

Promocijas darbs ir uzrakstīts latviešu valodā, satur ievadu, piecas nodaļas, secinājumus, pielikumus, literatūras sarakstu, 61 attēlu, 35 tabulas, kopā 155 lappuses. Literatūras sarakstā ir 78 nosaukumi.

VISPAPĀRĒJS DARBA RAKSTUROJUMS

Problēmas aktualitāte. Pasaulē un Latvijā viens no visstraujāk progresējošajiem mākslīgā intelekta novirzieniem ir „mašīnapmācība" jeb vienkārši apmācība (*Machine Learning*). Mašīnapmācība ir mākslīgā intelekta joma, kas pēta metodes, ar kuru palīdzību iespējams nodrošināt datora "apmācību" [15]. Tā aptver modeļus, metodes un algoritmus, kas ir orientēti uz automātisku zināšanu veidošanu un uzkrāšanu, pamatojoties uz datu apkopošanu un analīzi. Mašīnapmācība ir orientēta uz tādu intelektuālo sistēmu izveidi, kuras realizē zināšanu automātisko ieguvi (*data mining*).

Pateicoties to pieejamībai un universālām raksturam arvien biežāk zināšanu veidošanas metodes, tiek praktizētas, lai risinātu tādas problēmas kā zināšanu iegūšana no datu bāzēm, robotu kontrole, optimizācija, kā arī jau tradicionālus uzdevumus, piemēram, runas un tēlu atpazīšana, medicīnas datu un spēļu analīzi. Jāatzīmē, ka minētās metodes priekšrocība ir ne tikai iespēja risināt problēmas, kas saistītas ar kāda priekšmetiska apgabala klasifikāciju, bet arī iespēja interpretēt klasifikācijas rezultātus, izmantojot lēmumu kokus vai klasifikācijas likumus. Bez tam mašīnapmācības algoritmi, atšķirībā no adaptētajām un statistikas metodēm, neizvirza nekādus ierobežojumus attiecībā uz ieejas datiem un demonstrē labus rezultātus, strādājot ar vairumu priekšmetisko apgabalu, arī ar tādiem, kuros kādu iemeslu dēļ nav iespējams izmantot citas zināšanu iegūšanas metodes.

Pētījuma mērķis un uzdevumi. Šā darba praktiskie mērķi ir:

- Mašīnapmācības metožu un ģenētisko algoritmu hibrīdu izveide – tas ļauj risināt mašīnapmācības algoritmiem tik raksturīgo lokālā optimuma pārvarēšanas problēmu, pielietojot klasifikatoru ansambļu izveidi. Blakus mērķis ir piedāvāt izmantot jaunradīto algoritmu, lai risinātu uzdevumus, kurus raksturo liels daudzums faktoru vai atribūtu, kas apraksta attiecīgo priekšmetisko apgabalu.
- Izveidot *CART* algoritma un M. Bongarda *CORA* metodes kombinācijas pielietojumu – tas ļauj iegūto algoritmu izmantot tādu principiāli jaunu uzdevumu risināšanai, kas saistīti ar lielu analizējamo datu apjomu. Turklāt, pamatojoties uz šo algoritmu, tiek piedāvāta metodoloģija darbam ar izkropļotiem vai nepilnīgiem datiem.

Uzdevuma nostādne induktīvai apmācībai ir šāda: tiek apskatīts apgabals, kas sastāv no M punktiem $x_i \in \Re^d : x_i = (x_i^1, x_i^2, \ldots, x_i^d)$, $i=1, \ldots, M$. Saskaņā ar datu intelektuālās analīzes terminoloģiju d-dimensiju telpas punkti x_i tiek dēvēti par vektoriem, bet vektoru vērtības gar katru mērījumu tiek sauktas par atribūtiem. Katram punktam x_i tiek nosaukta atbilstošā funkcijas y_i, $i=1, \ldots, M$, vērtība, kur: $y_i \in \Re^d$ regresijas gadījumā un $y_i \in \{-1;+1\}$ klasifikācijas gadījumā. Punktu x_i kopu un ar tiem saistītos y_i nosauksim par „apmācāmo datu kopu". Līdzīgā veidā tiek apskatīta šīs pašas telpas cita punktu kopa un ar tiem saistītās funkcijas „etalonvērtības". Šo kopu nosauksim par apmācošo datu kopu. Tiek pieņemts, ka abām datu kopām piemīt kopējas iezīmes, līdz ar to iespējams runāt par zināmu, tām piemītošu struktūru. Piemēram, abās kopās ir dažādas izlases no veicamo eksperimentu sērijas, kurās tiek pētīta y_i atkarība no x_i parametru komplekta. Izmantojot apmācāmo datu kopu, nepieciešams izveidot tā saucamo klasifikācijas funkciju f_c, kura telpā $\Re^d$ tiktu definēta tādā veidā, lai šīs funkcijas vērtības apmācošās datu kopas punktos būtu iespējami tuvas „etalonvērtībām". Ja klasifikācijas funkcija (klasifikators) ir izveidota sekmīgi, tad ar lielu ticamības pakāpi iespējams apgalvot, ka funkcija atspoguļo datiem piemītošās iekšējās likumsakarības un to var izmantot kā predikatīvu modeli.

Darba gaitā paredzēts atrisināt vairākus uzdevumus:

1. Izpētīt esošos zināšanu iegūšanas algoritmus un noskaidrot to būtiskākos ierobežojumus un trūkumus.
2. Piedāvāt uz *ID3* un ĢA balstītu kombinētu algoritmu, kas risinātu ar lokālā maksimuma sasniegšanu saistītas problēmas. Kā atribūtu selekcijas metode ir izmantots ģenētiskais algoritms ar speciālu procedūru, kas ir paredzēta atribūtu kopas (populācijas) lietderības novērtēšanai.
3. Izstrādāt algoritmu *CART2* kā algoritma *CART* un Bongarda *CORA* metodes kombināciju, kas palīdzētu risināt ar lielām datu bāzēm saistītu klasifikatoru izveides un atjaunināšanas uzdevumu.
4. Izstrādāt programmnodrošinājumu, kas ļautu salīdzināt mašīnapmācības algoritmus, izmantojot datu bāzes no dažādām zināšanu jomām. Piedāvāt tādu programmnodrošinājuma arhitektūru, kas sniegtu iespēju pievienot jaunas metodes, nemainot pamatprogrammas struktūru.
5. Salīdzināt uz *ID3* un ĢA bāzes izstrādāto kombinēto algoritmu ar citām mašīnapmācības metodēm, izmantojot datu bāzes no 24 dažādiem priekšmetiskiem apgabaliem.
6. Aprobēt algoritmu *CART2* klasifikācijas uzdevumu risināšanai nepilnīgu un izkropļotu datu gadījumā. Veikt algoritma pārbaudi un salīdzināšanu ar citām mašīnapmācības metodēm, izmantojot datu bāzes no vairākiem priekšmetiskiem apgabaliem.
7. Piedāvāt koncepciju programmnodrošinājumam, kuru būtu iespējams izmantot spontānu smadzeņu iekšējo asinsizplūdumu prognozēšanai. Paredzēt iespēju izmantot attiecīgo programmnodrošinājumu, lai papildinātu sākotnējo apmācošo izlasi.

Pētījuma objekts un pētījuma priekšmets. Pētījuma objekts ir mašīnapmācības algoritmi. Pētījuma priekšmets ir klasifikatoru jeb klasifikācijas funkciju izveides un pielietošanas process, izmantojot jaunradītos algoritmus.

Pētījuma hipotēzes:

1. Zināšanu iegūšana un *hill-climbing.* Zināšanu veidošanas algoritmiem ir viens kopīgs trūkums: klasifikatora izveides procesā tiek izvēlēts visaugstākais kvalitātes pieaugums, kas var novest pie lokālā maksimuma izvēles, kuram, iespējams, neatbildīs vislabākais risinājums.
2. Klasifikatora izveide un atjaunošana lielām datu bāzēm. Izstrādājot un testējot gandrīz visus pastāvošos algoritmus, tika pieņemts, ka tie tiks izmantoti ar nelielām vai vidēji lielām datu bāzēm. Tādēļ gandrīz visi algoritmi operē ar sākumdatiem, kas viegli satilpst datora operatīvajā atmiņā. Bet praksē diemžēl bieži jāstrādā ar lieliem datu apjomiem, kas sastāv no miljoniem rindu un vairākiem tūkstošiem atribūtu. Tomēr, praksē nepietiek ar vienkāršu klasifikatora uzbūvēšanu uz kaut kādu datu bāzes pamata, jo pasaule turpina attīstīties, apmācošā izlase aizvien papildinās un tātad ir jāmodificē pirms tam izveidotais klasifikators, lai cik labs tas arī būtu. Ja izlase nepārsniedz vairākus tūkstošus ierakstu, nav problēmu pilnīgi no jauna izveidot klasifikatoru, taču ja apmācošā izlase sastāv no vairākiem miljoniem ierakstu, šāda pieeja ir vienīgi lieka skaitļošanas mašīnu resursu tērēšana un laika zaudēšana. Neviens no esošajiem zināšanu veidošanas algoritmiem nepiedāvā šīs problēmas komplekso risinājumu.

3. Apmācība ar lielu atribūtu skaitu. Daudzām mašīnapmācības algoritmu potenciālā pielietojuma sfērām ir raksturīgs liels daudzums – līdz simtiem tūkstošu – iespējamo atribūtu, kas apraksta katru no objektiem. *ID3* un *C5.0*, kā arī citi populārie algoritmi neļauj sasniegt pietiekami labus rezultātus šādu uzdevumu risināšanā.
4. Trokšņaino un nepilnīgu datu problēma. Piemēri ar lielu skaitu nepilnīgu vai izkropļotu datu sniedz salīdzinoši maz informācijas. Standarta mašīnapmācības algoritmu var viegli maldināt kropļoti vai neesoši atribūti, kā rezultātā tiks uzbūvēti nepareizi klasifikatori. Tāpēc praktiskajos lietojumos ir uzmanīgi jāizvēlas, kādi atribūti tiks piegādāti algoritmam apmācībai.

Pētījuma metodes. Šajā darbā tika pielietoti algebras un matemātiskas analīzes elementi, ģenētiskie algoritmi, lēmumu pieņemšanas, mākslīga intelekta un programmatūras projektēšanas metodes (*UML*).

Darba zinātniskais jaunieguvums. Darbs ietver mašīnapmācības algoritmu uzvedības analīzi, kas balstīta gan uz eksperimentāliem rezultātiem, gan uz teorētiskiem secinājumiem. Tā ir izmantojama jaunu induktīvu algoritmu izstrādāšanai.

Darba ietvaros ir izstrādāts uz *ID3* un ĢA balstītais kombinētais algoritms, kas risinātu ar lokālā maksimuma sasniegšanu saistītas problēmas. Kā atribūtu selekcijas metode tiek izmantots ģenētiskais algoritms ar speciālu procedūru, kas paredzēta atribūtu kopas (populācijas) lietderības novērtēšanai.

Darba ietvaros ir izstrādāts algoritms *CART2* kā algoritma *CART* un M. Bongarda *CORA* metodes kombinācija, kas palīdzētu risināt ar lielām datu bāzēm saistītu klasifikatoru izveides un atjaunināšanas uzdevumu.

Teorētiskā vērtība. Darbā tiek piedāvātas:

- Jaunas metodes un algoritmi, kas ļauj risināt jauna veida uzdevumus, tajā skaitā arī tādus, kuriem raksturīgs liels datu apjoms vai kuriem parasto risinājuma metožu pielietojums nedod atbilstošus rezultātus. Katra no ieteiktajām metodēm ir izstrādāta, izmantojot divu metodoloģiju kombināciju un apvienojot katras šīs metodes labākās īpašības.
- Principiālas nostādnes citu nozīmīgu ar induktīvo apmācību saistītu zinātnisku problēmu risināšanai:
    - Ar lokālā optimuma sasniegšanu saistītu problēmu pārvarēšana, izmantojot mašīnapmācības algoritmus.
    - Iespēja izmantot ĢA kā meklēšanas algoritma heiristiku.
    - Apmācības problēma lielu datu bāzu gadījumos. Izmantojot mācību pabeigšanas principus un balsošanas metodoloģiju uz M. Bongarda metodes bāzes, kļuvusi iespējama klasifikatoru izveide lielām datu bāzēm. Piedāvātās metodes realizācija sniedz iespēju izlasi, uz kuras bāzes tiek veidots klasifikators, neglabāt datora operatīvajā atmiņā kā vienu masīvu.
    - Esošā klasifikatora uzlabošanas problēma, izmantojot jaunus priekšmetiskā apgabala datus. Ar šo metodoloģiju, saņemot jaunus datus, atkrīt nepieciešamība veidot klasifikatoru no jauna, pietiek tikai precizēt (*adjust*) esošo.

Praktiskā vērtība. Darba izstrādāšanas laikā radās šādi praktiski pielietojamie produkti:

1. Tika sastādīta *C++* valodā izmantojot *MFC* klašu bibliotēku speciāla programma *Machine Learning Methods Comparison*. Programmas galvenā priekšrocība ir iespēja to papildināt ar jaunu induktīvo algoritmu.
2. Izmantojot *MLC++* (*Machine Learning C++*) bibliotēku, uz *ID3* un ĢA bāzes izveidota kombinētā algoritma realizācija, kas deva iespēju izpētīt un izvērtēt jaunizveidoto algoritmu.
3. Tika piedāvāta koncepcija programmnodrošinājumam, kuru būtu iespējams izmantot smadzeņu spontāno iekšējo asinsizplūdumu prognozēšanai. Paredzēta iespēja attiecīgo programmnodrošinājumu izmantot, lai papildinātu sākotnējo apmācošo izlasi.

Darba aprobācija. Ar pētījumu rezultātiem tika iepazīstināti sekojošo zinātnisko konferenču dalībnieki:

1. Kornienko Y., Borisov A. (2003). Investigation of a hybrid algorithm for decision tree generation. Proceedings of the Second IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2003, Lviv, Ukraine, September 8-10, P. 63-68.
2. Kornienko J. & Borisov A. (2000). Production rules induction algorithm based on the finish learning principle. Fourth International Conference on Application of Fuzzy Systems and Soft Computing ICAFS'2000, Siegen, Germany, June 27-29, P. 287-292.
3. Kornienko, Y. & Borisov. A. (1999). The CART2 inductive algorithm in comparison with standard machine learning methods. Proceedings of the International Scientific-Technical Workshop"Problems of Transfer Technology", Ufa, Russia, 30. September – 1. October, P. 154-161.
4. Kornienko Y. and Borisov A. (1999). The CART methodology for production rules induction. 5th International Conference on Soft Computing MENDEL'99, Brno, Czech Republic, June 9-12, P. 362-366.
5. Kornienko Y. & Borisov A. (1998). Genetic-based decision trees. MENDEL'98 - 4th International Conference on Genetic Algorithms, Optimization Problems, Fuzzy Logic, Neural Networks and Rough Sets, Brno, Czech Republic, June 24-26, P.42-44.
6. Kornijenko Y. and Borisov A. (1998). Application of genetic algorithms for generating decision trees. International Conference on Parallel Computing in Electrical Engineering, PARELEC'98, Bialystok, Poland, September 2-5, P. 277-279.

Publikācijas. Pētījumu rezultāti tika publicēti vienpadsmitos darbos, gan autora patstāvīgi sarakstītajos, gan sadarbība ar līdzautoriem. Seši no šiem darbiem ir publicēti ārvalstīs.

Personīgais ieguldījums. Visi darba rezultāti, kurus ietver dotais doktora darbs, ir autora patstāvīgu pētījumu ceļā iegūti.

Doktora darba struktūra un apjoms. Doktora darbs sastāv no septiņām nodaļām, noslēguma, literatūras saraksta un pielikuma. Promocijas darba pamatteksts ir izklāstīts 128 lapaspusē un paskaidrots ar 61 attēlu, 35 tabulām un 3 pielikumiem. Literatūras sarakstā ir iekļauti 78 nosaukumi.

DARBA SATURA IZKLĀSTS

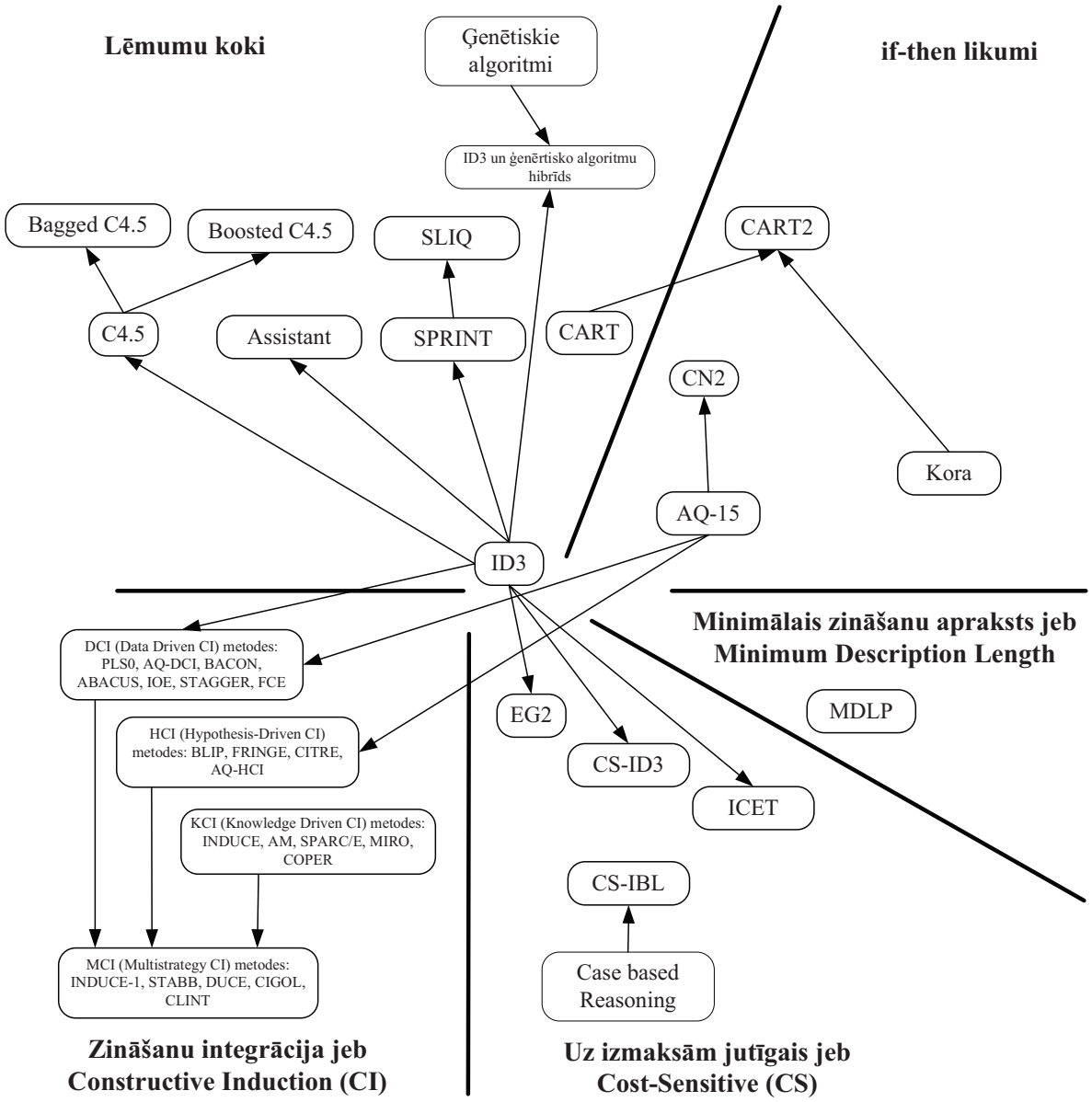
Pirmajā sadaļā sniegts vairāku svarīgāko zināšanu atklāšanas algoritmu apraksts, kuri ir ietekmējuši mākslīgā intelekta mašīnapmācības apakšnozares attīstību, kā arī apraksts metodēm, kas tiek piedāvātas jaunu, šai jomai specifisku uzdevumu risināšanai. Šīs sadaļas mērķis ir parādīt esošo zināšanu veidošanas metožu daudzveidību, kuras tiek izmantotas noteiktu uzdevumu risināšanai, kā arī veikt esošo metožu trūkumu analīzi.

Pastāv liels daudzums metožu, ar kurām var veikt intelektuālo datu analīzi un kuras var iedalīt induktīvajās, adaptīvajās un statistiskajās metodēs. Savukārt induktīvās metodes, pamatojoties uz risināmo uzdevumu specifiku, tiek nosacīti sadalītas šādi (sk. 1. attēlu):

- *inductive learning* jeb zināšanu iegūšanas uzdevums, kas balstās uz esošo datu bāzu informācijas apkopošanas tehniku [10, 12, 14, 26, 35, 50 un 57],
- *constructive induction* — uzdevums, kas paredz ar induktīvās apmācības palīdzību iegūto risinošo likumu integrēšanu ar jau esošajām zināšanām noteiktajā priekšmetiskajā apgabalā [4, 7, 8, 16, 53 un 52] ,
- *cost-sensitive* — zināšanu ieguves uzdevums jomām, kurās pastāv katra atribūta mērīšanas izmaksas [65 un 67],
- *minimum description length* zināšanu iegūšanas uzdevums, kas balstās uz konkrēta priekšmetiska apgabala minimālā apraksta atrašanas principa [54 un 55].

Sadaļā ir sniegts īss induktīvu metožu apraksts un pielietojums. Izanalizējot esošās metodes, var secināt, ka induktīvajai apmācībai ir šādas galvenās problēmas:

1. *Hill-climbing* – t.i., situācija, kad klasifikatora izveides procesā tiek izvēlēts lielākais kvalitātes pieaugums, kā rezultātā var notikt lokālā maksimuma izvēle, kurai ne vienmēr atbilst labākais risinājums.
2. Apmācība pie liela atribūtu skaita.
3. Klasifikatora izveide un atjaunošana lielām datu bāzēm.
4. Kropļotu un nepilnīgu datu problēma.



1. attēls. Mašīnapmācības algoritmu savstarpējās saites un sadalījums, izejot no pielietošanas specifikas

Pirmās divas problēmas ir atrisināmas, izmantojot klasifikatoru ansambļu uzbūves tehniku, kas tiek teorētiski pamatots nākamajā sadaļā.

Lai novērstu 3. un 4. problēmu, tiek piedāvāts *CART2* algoritms – induktīvā *CART* algoritma un Bongarda *CORA* metodes kombinācija.

Otrajā sadaļā tiek piedāvāta autora izstrādātā *ID3* un ĢA kombinētā klasifikatoru ansambļu izveides metode, kas balstās uz šādiem teorētiskajiem principiem:

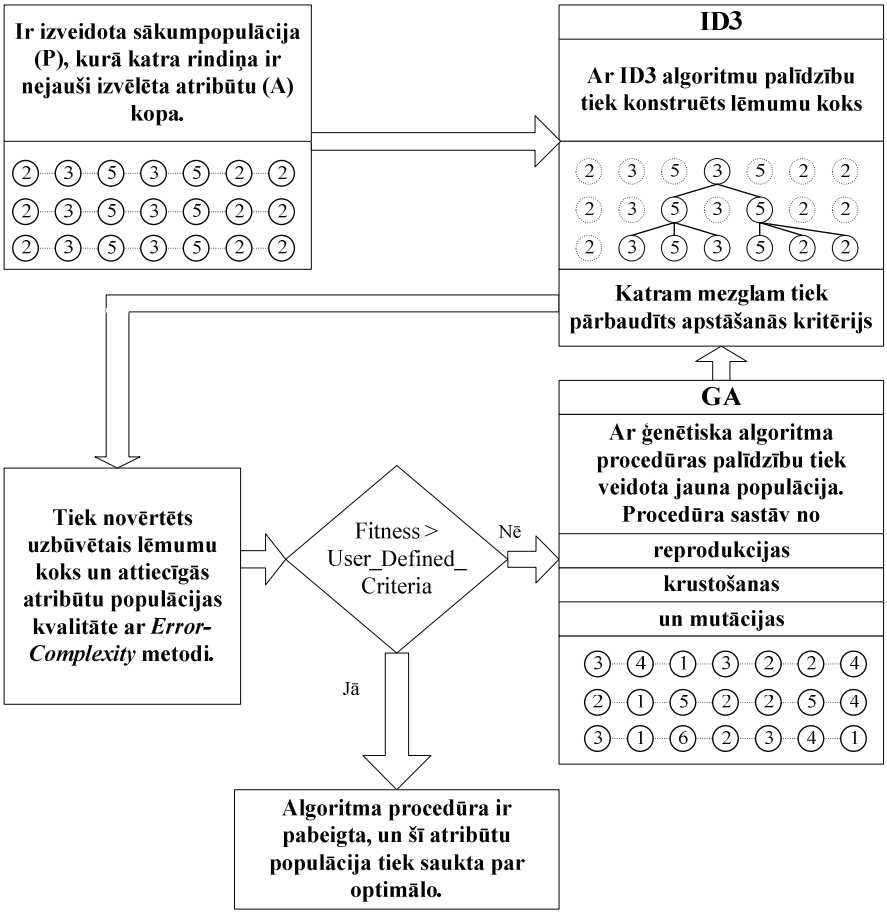
- Ģenētisko algoritmu izmantošana ļauj novērst lineārās programmēšanas un ekspertsistēmu nepilnības dažu veidu uzdevumu risināšanā [1].
- ĢA pieeja ir daudz labāka par parastajām optimizācijas metodēm, kas uz parametru kopas operē ar vienīgo piekļuves punktu.
- Standartveidā izmantojot mašīnapmācības algoritmus, rezultātā tiek būvēts viens klasifikators, taču pēdējā laikā aizvien populārākas kļūst veselu klasifikatoru

ansambļu veidošanas metodes. Šie ansambļi demonstrē apmācošajā izlasē neiekļauto objektu labāku atpazīšanas kvalitāti.

Apvienojot ģenētiskos algoritmus ar citām mākslīgā intelekta pieejām, par mērķi tiek izvirzīta šo mākslīgā intelekta pieeju kvalitātes uzlabošana. Izmantojot kā piemēru *ID3* induktīvā secinājuma algoritmu, šajā sadaļā ir detalizēti izskatīts, kā tiek radīts ĢA un heiristisko meklēšanas stratēģiju hibrīds [31, 32, 33 un 34]. Sadaļā tiek arī paskaidroti ĢA un *ID3* kombinācijas apvienojuma iegūšanas principi. Ir parādīts, ka ĢA un *ID3* kombinācija ļauj novērtēt veselu piekļuves punktu populāciju, kā to dara ĢA. Ar ģenētisko operatoru palīdzību tiek sniegta iespēja manipulēt ar pašreizējo populāciju, lai radītu jaunu. Šīs sadaļas mērķis ir arī parādīt ĢA un *ID3* hibrīda realizācijas iespējas, izmantojot dažādus lēmumu koka saīsināšanas mērus.

Tālāk seko hibrīda ĢA un *ID3* algoritma apraksts un interpretācija: ĢA un *ID3* hibrīda darbības rezultātā tiek būvēts lēmumu koks, kuru var izmantot likumu formulēšanai. Koka mezgli ir atribūti, tā zari ir attiecīgās šo atribūtu vērtības. Uzdevums, ko hibrīdā pilda ĢA, ir *ID3* nodrošināšana ar iespējamo atribūtu virkni, kurus izmantojot tiks veikta turpmāka piemēru kopas sadalīšana apakškopās. Algoritmam uzsākot darbību, ir jāizveido sākumpopulācija, kurā katru virkni veidos ar gadījuma atlases metodi izvēlēta atribūtu kopa. Populācijas lielums nevar pārsniegt atribūtu skaitu, jo katru atribūtu iespējams izmantot katrā zarā tikai vienu reizi. Lai novērtētu katras virknes lietderīgumu, var izmantot *ID3* algoritmam izstrādātos sadalīšanas mērus. Katrai kopai tiks noteikts vidējais tā sastāvā ietilpstošo atribūtu efektivitātes mērs. Tādējādi rodas iespēja pielietot visus ĢA operatorus.

Algoritma kopējā shēma ir parādīta 2. attēlā.



2. attēls. Algoritma darbības uz ĢA un ID3 kombinācijas bāzes kopshēma

Tālāk sadaļā ir sniegts detalizēts metodes apraksts un neliels ilustratīvs piemērs tam, kā šī metode tiek izmantota. Ir ievērots, ka ĢA un lēmumu koku kombinācija ļauj veikt sekojošo:

- veidot lēmumu koku sēriju (ansambli), no kuras tiks izvēlēts vislabākais koks, kas atbilst lietotāja definētajam kritērijam;
- populācijas lielums un hromosomu skaits virknē ļauj regulēt lēmumu koka dziļumu un platumu;
- nav nepieciešama lēmumu koka saīsināšana.

Šajā sadaļā tika nodemonstrēts, ka lēmumu koku un ĢA apvienošanas metode ļauj risināt heiristiskās meklēšanas uzdevumus tajos priekšmetiskajos apgabalos, kuros atribūtu apgabala lielums ievērojami pārsniedz normālo līmeni.

Trešajā sadaļā tiek piedāvāts autora izstrādātais produkcijas likumu ģenerēšanai paredzētā induktīvā algoritma *CART2* apraksts. *CART2* ir jauns induktīvais algoritms, kas ir detalizēti aprakstīts darbos [35, 36 un 37]. *CART2* algoritms tika veidots ar nolūku novērst dažas nepilnības, kas piemīt *CART* algoritmam [10], kā arī radīt metodi, kas būtu izmantojama uz lielo datu bāzu pamata būvēto klasifikatoru apstrādei un uzturēšanai. *CART2* metode ir *CORA* [73] (jēdzienu vispārināšana pēc pazīmēm) un *CART* datorprogrammas kombinācija. *CART* algoritms un M. Bongarda *CORA* metode ir labi zināmi speciālistiem, kas nodarbojas ar induktīvās apmācības problēmām.

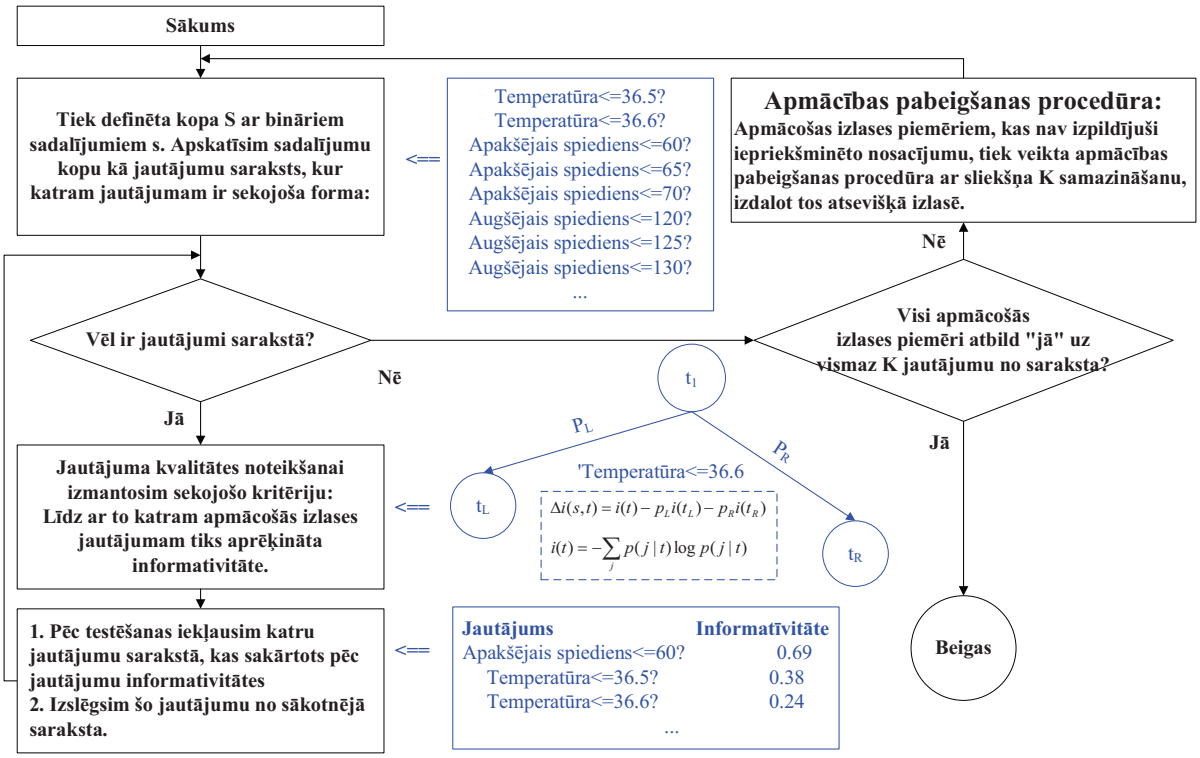
Veidojot *CART2* algoritmu, par mērķi tika izvirzīta dažu standarta *CART* algoritma nepilnību novēršana [30, 35, 36, 37 un 38]. Tā kā *CART* algoritms ir lēmumu koku būvēšanas līdzeklis, *CART2* iegūst šādas priekšrocības, ja tas tiek kombinēts ar *CORA* metodi:

- algoritma darbības rezultātā tiks iegūti loģiskie likumi;
- katrs apmācošās izlases piemērs jāaptver (*cover*) vismaz ar K likumiem, kur K ir vesels skaitlis. Šī priekšrocība ļauj atrast un pielietot dažas slēptas sakarības, saskaņā ar kurām atribūti sadalās vienai klasei piederošajos piemēros. Tas ir, katram piemēram, kuru var attiecināt uz konkrētu klasi, jābūt aptvertam ar K likumiem vienlaicīgi;
- izvēloties jaunu jautājumu, tika izskatīta pilna piemēru kopa, izņemot apmācības pabeigšanas procedūru.

CART2 algoritms ir iteratīva procedūra, un tas balstās uz šādiem teorētiskajiem pamatojumiem:

- bināro jautājumu kopa α, kuri ir izteikti formā $\{Is\ x \in A?\}$, $A \subset X$, kur x – atribūta vērtība, X – visu atribūtu vērtību kopums;
- sadalīšanas kvalitātes kritērijs $\phi(s, t)$, kas var tikt novērtēts jebkuras kopas t jebkurai sadalīšanai s;
- minimālais jautājumu skaits, kuriem ir jāapraksta katrs apmācošās izlases piemērs, vai katru piemēru aprakstošo pazīmju minimālais skaits;
- „apmācības pabeigšanas" procedūra;
- jaunu piemēru identifikācijas procedūra.

Algoritma shēma ir paradīta 3. attēlā.



3. attēls. *CART2* algoritma shēma

Tālāk sadaļā ir sniegts detalizēts metodes apraksts un neliels ilustratīvs piemērs, kas tika aprēķināts pašrocīgi un uzbūvēts, lietojot apmācošo izlasi no 10 piemēriem.

Ceturtajā sadaļā ir aprakstīti eksperimenti ar mašīnapmācības algoritmiem (*ID3*, *C4.5*, *Bagged-C4.5*, *Boosted-C4.5*, *Naive Bayes*) un algoritmu, kas ir izveidots, pamatojoties uz ģenētisko algoritmu un *ID3* savienojumu.

Standartveidā izmantojot mašīnapmācības algoritmus, tiek būvēts viens klasifikators, bet pēdējā laikā aizvien populārākas kļūst veselu klasifikatoru ansambļu veidošanas metodes. Šie ansambļi demonstrē apmācošajā izlasē neiekļauto objektu labāku atpazīšanas kvalitāti. Šobrīd pastāv šādas metodes, kuras var izmantot klasifikatoru ansambļu veidošanā [57]:

- *Stacking*: apmācībai izmantoto piemēru apraksts tiek paplašināts, iekļaujot piemēru klasifikācijas rezultātus ar sākotnēji izvēlētajiem klasifikatoriem. Šāda informācija tiek atkārtoti analizēta, un uz tās pamata tiek būvēti citi klasifikatori utt.
- *Windowing*: tiek būvēts klasifikators uz sākotnējas apmācošās izlases bāzes. Tad apmācošā izlase tiek palielināta uz to piemēru rēķina, ar kuriem klasifikators ieguvis sliktu analīzi, tad tiek ģenerēts jauns klasifikators uz palielinātās izlases bāzes, un process turpinās tālāk.
- *Bagging*: no oriģinālās datubāzes tiek veidota sākotnēja apmācoša izlase. Daži piemēri var netikt iekļauti šajā izlasē, daži var tur nokļūt vairakkārt. Balstoties uz šo izlasi, tiek būvēts klasifikators, un tad izlase tiek veidota no jauna. Pēc vairākiem šādiem cikliem gatavie klasifikatori tiek kombinēti, lai izveidotu gala klasifikatoru.
- *Boosting*: katram apmācošās izlases piemēram ir noteikts svars. Katrā iterācijā tiek būvēts klasifikators, izmantojot šādus piemērus; tad katra piemēra svars tiek labots atkarībā no tā, vai šis piemērs ticis klasificēts pareizi vai nepareizi.

- ĢA un heiristisko meklēšanas stratēģiju kombinācija, kas tiek piedāvāta apskatīšanai šajā sadaļā. Tā ļauj izmantot ĢA priekšrocības kopā ar heiristiskās meklēšanas iespējām un iekļauj lēmumu koku būvēšanu vai likumu kopas atrašanu, kas spēj izskaidrot slēptas likumsakarības šajā priekšmetiskajā apgabalā.

Lai veiktu eksperimentus, algoritms, kas bāzējas uz *ID3* un ģenētisko algoritmu savienojuma, tika realizēts kā Stenfordas universitātes *MLC++* bibliotēkas paplašinājums. Jaunā algoritma uzvedība tika izpētīta 24 datubāzēs, ieskaitot datubāzes ar lielu atribūtu skaitu. Šajā sadaļā ir parādīts, ka, pateicoties lokālā maksimuma problēmas atrisināšanai, ar jaunā algoritma palīdzību izveidotā klasifikatora raksturojumi kļuvuši ievērojami labāki. Tika izpētīta algoritma uzvedība, būvējot ierobežotus klasifikatorus, piedāvāti standarta mašīnapmācības algoritmu uzlabošanas veidi.

Sadaļas mērķis ir uz lēmumu koku un ĢA kombinācijas bāzes būvētā algoritma uzvedības izpēte:

- Uz lēmumu koku un ĢA kombinācijas bāzes būvētā algoritma salīdzināšana ar citiem algoritmiem – *ID3*, *C4.5*, *Bagged-C4.5*, *Boosted-C4.5*, *Naive Bayes* [50].
- Pētīt jaunā algoritma uzvedību, strādājot ar lietojumapgabaliem, kas tiks aprakstīti ar lielu atribūtu skaitu. Lielāko daļu zināšanu sfēru iespējams aprakstīt ar ne vairāk kā 10-30 pazīmēm jeb atribūtiem, tomēr ir sastopami gadījumi, kad ir 100 un vairāk šādu pazīmju.
- Izpētīt jaunā algoritma uzvedību, strādājot ar datu bāzēm, kas ir aprakstītas ar normālu atribūtu skaitu. Pierādīt, ka lokālā maksimuma problēmas atrisināšana ir uzlabojusi šādi konstruētā klasifikatora raksturīpašības.
- Pētīt jaunā algoritma uzvedību ierobežotu klasifikatoru konstruēšanas laikā. Ierobežotajiem klasifikatoriem pēc būtības ir gan priekšrocības, gan trūkumi. No vienas puses, šādos klasifikatoros ir iesaistīts neliels elementu skaits (likumu vai lēmumu koka zaru), kas sniedz labas iespējas cēloņu un slēptu sakarību izprašanai konkrētajā priekšmetiskajā apgabalā. No otras puses, var pasliktināties atpazīšanas precizitāte. Tādējādi eksperimentā paredzēts izpētīt atpazīšanas precizitāti atkarībā no apmācāmā klasifikatora ierobežojumu pakāpes.

Par sadaļas galveno rezultātu var uzskatīt pierādījumu tam, ka, pateicoties lokālā maksimuma problēmas atrisināšanai, ievērojami uzlabojušies klasifikatoru raksturojumi (sk. 1. tabulā). Šie rezultāti ir īpaši svarīgi tāpēc, ka šī pieeja var kalpot par pamatu citu "*gridy-search*" algoritmu, tostarp *C4.5, C5.0*, *CART*, uzlabošanai.

Tālāk sadaļā tika izpētīta algoritma uzvedība, veidojot ierobežotus klasifikatorus. Šajā nolūkā tika veikti eksperimenti ar 5, 10 līmeņiem. Rezultātā tika konstatēts, ka dažām datu bāzēm atpazīšanas kvalitāte uzlabojās (*German-org*, *Hypotiroid*). No otras puses, eksperimenti ar vairumu datu bāzu ir parādījuši, ka labākie rezultāti tika iegūti tajos eksperimentos, kuros uzstādītais līmeņu skaits bija no 10 līdz 20. Tas nozīmē, ka klasifikators tiek būvēts bez jebkādiem ierobežojumiem.

1. tabula

ID un ĢA kombinācijas un citu mašīnapmācības metožu salīdzinājums

Datu bāzes nosaukums	Kļūda testēšanas laikā, %						
	ID3-ĢA	ID3	C4.5	Bagged-C4.5	Boosted-C4.5	Naive Bayes	Neirontīkli
Anneal	0.67	2.17	7.33	6.25	4.73	8.20	
Auto	15.50	15.34	37.68	19.66	15.22	41.80	
Br. cancer	28.84	30.85	25.26			35.08	
Breast-w.	3.21	5.14	4.29	4.23	4.09	4.93	5.15
Chess	2.11	0.09	0.47	8.33	4.59	13.00	
Cleve	21.26	28.23	22.77			17.84	
Crx	16.12	20.20	17.00			22.04	
Diabetes	26.36	31.83	30.86	23.63	28.18	23.64	30.47
German-org	28.97	34.53	25.15			27.17	
Glass	31.00	34.43	37.50	27.01	23.55	49.90	
Heart	20.00	27.78	16.67	21.52	21.39	18.33	
Hepatitis	15.73	16.58	19.23	18.52	17.68	13.67	
Horse-colic	18.33	23.33	14.71			20.00	
Hypothyroid	1.33	1.51	0.76			1.90	
Ionosphere	6.40	9.00	11.97			17.04	17.95
Iris	2.00	6.00	8.00	5.13	6.53	3.00	
labor-neg	7.50	30.00	17.65	14.39	13.86	17.50	
Pima	28.52	30.67	23.44			25.59	23.44
Solar	30.54	34.41	26.85			37.21	
Sonar	19.01	25.42	25.71	23.80	19.62	32.68	20.00
Soybean	7.01	10.99	10.53	7.58	7.16	16.70	
Vehicle	25.19	27.85	32.27	25.54	22.72	53.20	
Vote	5.00	7.33	2.96	4.37	5.29	11.00	
Zoo	4.29	7.47	14.71			10.44	

Tālāk sadaļā tika izanalizēti lēmumu koki, kas izveidoti ar jaunā algoritma un *ID3* algoritma palīdzību. Analīzes rezultātā varam izdarīt šādus secinājumus:

- Lēmumu koks, kas izveidots ar *ID3* palīdzību uz ģenētisko algoritmu bāzes, ir sazarotāks un tikai daļēji atgādina parastā *ID3* konstruēto lēmumu koku. No vienas puses, šāds koks precīzāk atspoguļo visas apmācošās izlases īpatnības, no otras puses tas rada zināmas grūtības, ekspertiem analizējot šādu koku.
- Klasifikatora būvēšanai uz ģenētisko algoritmu bāzes ir raksturīgas augstākas apmācības izmaksas nekā standarta mašīnapmācības algoritmiem. Piemēram, *ID3* un ĢA kombinācijai būs nepieciešamas 20–50 reizes vairāk pūliņu ne kā parastajam *ID3*.

Darba 6. sadaļā jaunā metode tiek veiksmīgi pielietota smadzeņu spontāno iekšējo asinsizplūdumu prognozēšanai un analīzei ar lēmumu koku palīdzību.

Piektajā sadaļā tika analizēta jaunradīta induktīvā algoritma CART2 efektivitāte. Tika sastādīta speciāla programma *Machine Learning Methods Comparison C++* valodā, izmantojot *MFC* klašu bibliotēku. Mašīnapmācības algoritmu analīze notiek, lietojot šādas metodes:

- gadījumatlases metode (*random subsumpling*);
- *n*-kārtējā šķērsvalidācija (*n-fold cross validation*).

Gadījumatlases metodes būtība ir tajā, ka [69] eksperimentu kopums jāveic ar gadījumrakstura izlasi. Katram jaunam eksperimentam tiek radīts klasifikators. Kļūda tiek

novērtēta kā visu, uz piemēru kopuma bāzes neatkarīgi un nejauši izveidotu, klasifikatoru vidējais aritmētiskais. Gadījumatlases metode sniedz labāku kļūdas novērtējumu par to, ko iespējams iegūt ar vienīgo apmācošo un eksaminācijas atlasi.

N-kārtēja šķērsvalidācijas metodes būtība ir tajā, ka visi piemēri tiek sadalīti *n* apmēram vienlīdz lielās neatkarīgajās eksaminācijas apakškopās. Tad iteratīvi tiek būvēti *n-1* klasifikatori, turklāt katra klasifikatora apmācības atlasē tiek iekļauti visi piemēri, kas netika iekļauti konkrētā eksaminācijas kopā. Visu *n* sadalījumu kļūdas vidējā pakāpe arī ir šķērsvalidēta kļūda.

2. tabula

Gadījumatlases un šķērsvalidācijas metožu salīdzinājums

	Gadījumatlase	**5-kārtēja šķērsvalidācija**
Apmācošā izlase	J	80% visu datu
Eksaminācijas izlase	n-j	20% atlikums
Iterāciju skaits	B<<n	5

Gadījumatlases metodes un *n*-kārtējas šķērsvalidācijas salīdzinājums ir sniegts 2. tabulā. Tālāk seko datu bāzu apraksts, ar kurām tiek salīdzināts jaunais induktīvais algoritms *CART2* un mašīnapmācības standarta algoritmi *ID3* un *CART*.

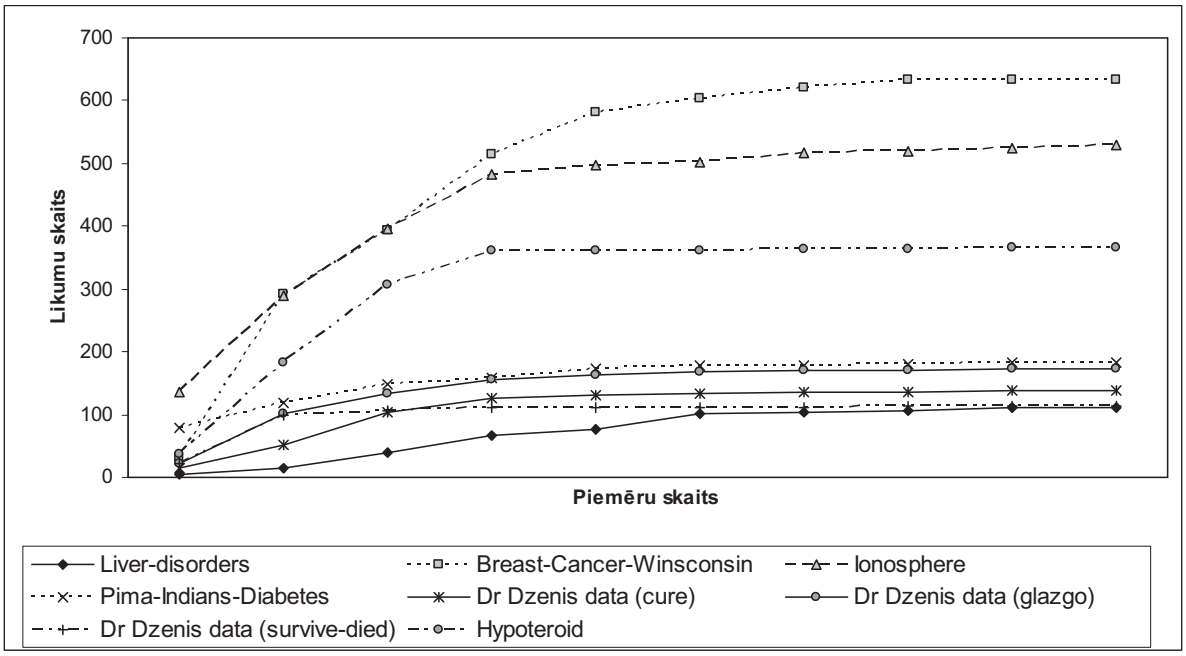
Kā redzams no salīdzinošiem eksperimentu *CART2* algoritms ir demonstrējis labus rezultātus ar datu bāzēm *Hypoteroid, BUPA liver disorders* un *Pima Indians Diabetes*, un to raksturojumi ir salīdzināmi ar *ID3* un *CART* algoritmu raksturojumiem, ieskaitot arī tā uzvedību apstākļos ar ieviesto traucējumu. Viduvējus rezultātus *CART2* algoritms ir parādījis ar *Ionosphere* datu bāzi. Daudzos gadījumos prognozēšanas precizitāte bija tieši proporcionāla apmācošās izlases lielumam. Kā bija sagaidāms, ieviešot traucējumu, algoritma uzbūvēto klasifikatoru darbības precizitāte pasliktinājās. Tajā pašā laikā jaunais *CART2* algoritms ir parādījis labus rezultātus, arī strādājot ar kropļotiem datiem. Nevar kategoriski apgalvot, ka *CART2* algoritms ir labāks par citām metodēm, tomēr rezultāti parāda tā stabili labo uzvedību ar visiem datiem.

Tālāk tika veikta *CART2* algoritma efektivitātes analīze, veicot klasifikatora izveides un inkrementālās atjaunināšanas uzdevumu. Ar katru datu bāzi tika veikta sērija eksperimentu, kuru rezultātā iegūti šādi līkņu grafiki:

- kļūda eksaminācijas izlasei;
- kļūda apmācošajai izlasei;
- kļūda kopējai izlasei.

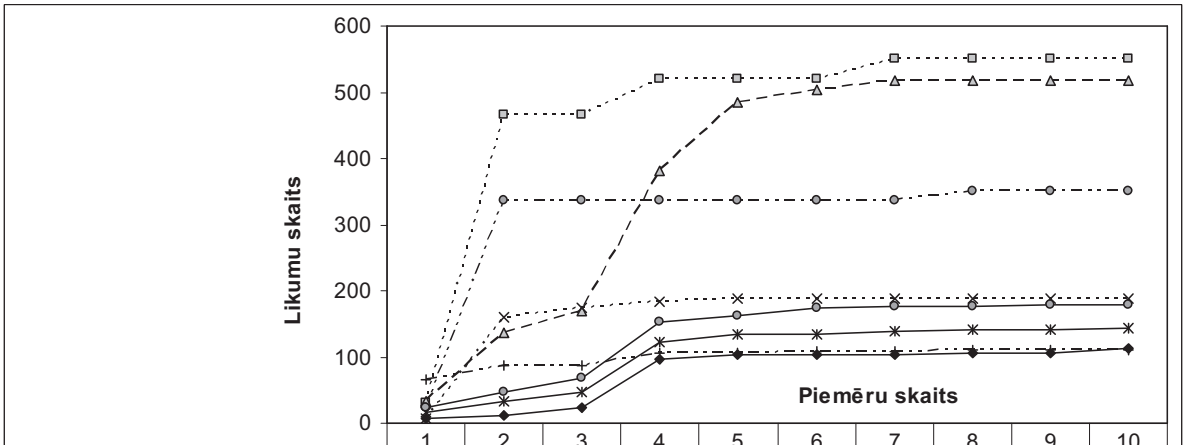
Visi veiktie eksperimenti demonstrē labus rezultātus salīdzinājumā ar līdzīgiem eksperimentiem, veicot apmācību uz pilnas apmācošas izlases pamata. Piemēram, krātuves lielākajai datu bāzei *Hypoteroid* atpazīšanas kļūda bija 1,95% (uz 2870 piemēriem), salīdzinot ar 1,8% kļūdu 5-kārtējai šķērsvalidācijai ar 3000 piemēriem. Jāpiezīmē, ka šo rezultātu salīdzināšana tomēr nebūs tehniski korekta, jo kļūdas aprēķināšana notiek atšķirīgi.

Tālāk sadaļā tika analizētas klasifikācijas likumu daudzuma atkarības no katra eksperimenta apmācošās izlases piemēru skaita vidējotu vērtību grafiku (10 eksperimentu vidējais aritmētiskais), sk. 4. un 5. attēlu. Ir vērojama tendence, ka, sasniedzot noteiktu likumu daudzumu, algoritms pārstāj atjaunināt klasifikatoru vai dara to pavisam nemanāmi.



4. attēls. Klasifikācijas likumu atkarība no piemēru skaita (vidējota)

5. attēlā grafiks ir analogs iepriekšējam, tikai līknes punkts atbilst viena eksperimenta vērtībai. Tāpēc šis grafiks neizskatās tik nogludināts un ir skaidri redzams, ka dažkārt likumu skaits palielinās lēcienveidīgi. Tas ir izskaidrojams ar uzdevuma raksturu: pēc klasifikatora atjaunināšanas tas kādu laiku vēl bijis konsistents attiecībā uz jaunajiem piemēriem, t.i., katrs piemērs tika aprakstīts ar vismaz *K* likumiem. Taču jo vairāk kļuva jaunu piemēru, jo lielāka bija varbūtība, ka kāds nebūs aprakstīts ar vajadzīgo likumu skaitu, kas arī noved pie jaunu likumu radīšanas, t.i., klasifikatora atjaunināšanas.



	1	2	3	4	5	6	7	8	9	10
Liver-disorders	6	12	24	97	103	103	103	107	107	113
Breast-Cancer-Winsconsin	31	465	465	520	520	520	551	551	551	551
Ionosphere	34	136	170	382	484	504	518	518	518	518
Pima-Indians-Diabetes	8	159	175	183	189	189	189	189	189	189
Dr Dzenis data (cure)	16	32	48	122	135	135	139	141	141	143
Dr Dzenis data (glazgo)	23	46	69	152	162	173	176	176	180	180
Dr Dzenis data (survive-died)	67	88	88	105	107	108	109	110	111	111
Hypoteroid	24	336	336	336	336	336	336	350	350	350

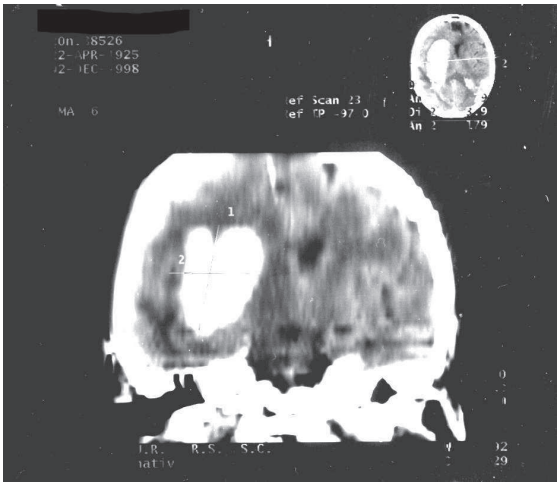
5. attēls. Klasifikācijas likumu atkarības no piemēru skaita vienā eksperimentā

Tika secināts, ka *CART2* algoritms ir daudzsološs un īpašas uzmanības vērts jauns induktīvais algoritms, kas atšķirībā no citām metodēm ir spējīgs risināt izveides un inkrementālās atjaunināšanas uzdevumu lielām datu bāzēm.

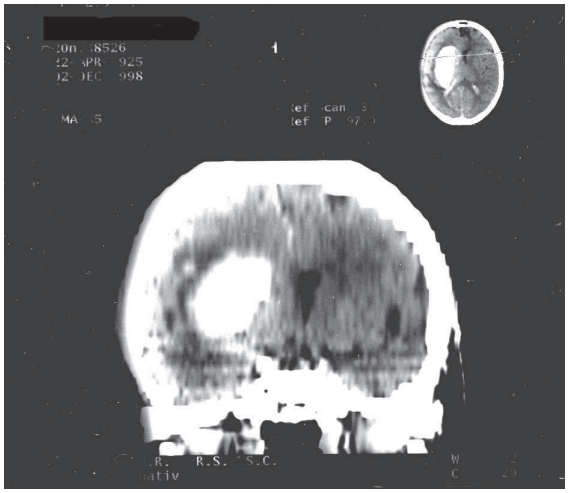
Sestajā sadaļā izklāstīti dažu stāvokļu atribūtu diagnosticēšanas rezultāti spontāniem smadzeņu iekšējiem asinsizplūdumiem (sk. att. 6. – 9.), izmantojot *ID3*, *CART*, *CART2* algoritmus un *ID3* algoritmu kombinācijā ar ģenētiskajām metodēm.

Smadzeņu spontāno iekšējo asinsizplūdumu datu bāze netika izvēlēta nejauši. Savulaik jau tika veikti šādu datu apstrādes mēģinājumi, bet to galvenais uzdevums bija atrast stāvokļu atribūtu savstarpējās sakarības. Turklāt datu bāze jau ir sakārtota, un ir pieejams pieņemams masīvs prognozēšanai ar matemātiskām metodēm.

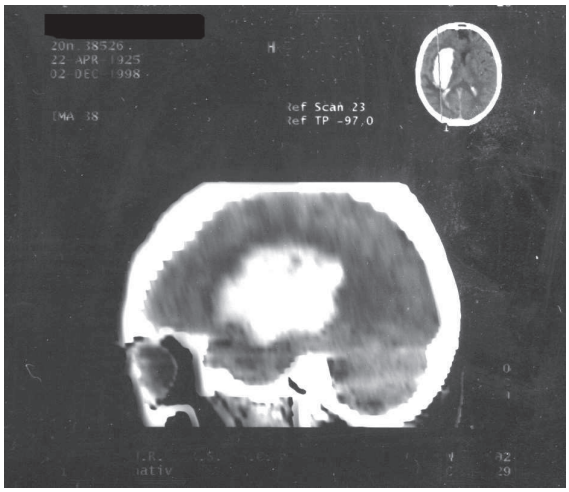
Sadaļā ir formulēts darba uzdevums un īss priekšmetu apgabala apraksts. Tika apstrādāti 200 pacientu anketu dati no Rīgas 7. slimnīcas "Gaiļezers", Maskavas Neiroloģijas institūta un Rīgas P. Stradiņa universitātes klīniskās slimnīcas no 1993. gada līdz 2002. gadam. Datu pilno aprakstu (t. sk. vārdu, uzvārdu, vecumu un novērošanas vietu) var atrast darbā [75].



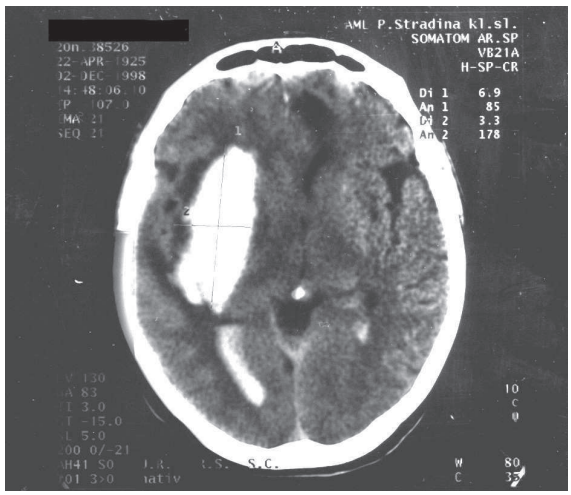
6. attēls. Smadzeņu iekšējā hematoma (SIH) no priekšpuses



7. attēls. Smadzeņu iekšējā hematoma (SIH) no aizmugures



8. attēls. Smadzeņu iekšējā hematoma (SIH) no kreisās puses



9. attēls. Smadzeņu iekšējā hematoma (SIH) no augšas

Katra anketa analizēta pēc 33 atribūtiem. Visu atribūtu iespējamās vērtības ir skaitļi no 1 līdz 9. Pazīmju apraksts sniegts 3. tabulā. Anketas pilns apraksts atrodams 2. pielikumā.

3. tabula

Atribūtu stāvokļu apraksts

Nr.	Atribūta nosaukums
1	Novērojuma Nr.
2	Dzimums
3	Pacienta vecums
4	Blakus slimības
5	Arteriālais spiediens insulta brīdī
6	Pirmie saslimšanas simptomi
7	Samaņas līmenis pirms ārstēšanas (punktos pēc Glazgo komas skalas)
8	Kustību traucējumi pirms ārstēšanas
9	Runātspējas traucējumi pirms ārstēšanas
10	Asinsizplūdumu iemesli
11	Asinsizplūdumu lokalizācija pēc DT (datortomogrāfija)
12	Asinsizplūduma tilpums pēc DT (cm^3)
13	Perifokālās tūskas pakāpe pēc DT datiem (mm)
14	Vēderiņu starpsienas nobīde pēc DT datiem (mm)
15	Okluzīvās hidrocefālijas sindroma pakāpe pēc DT datiem
16	Galvas smadzeņu stumbra deformācijas pakāpe pēc DT datiem
17	Asins klātbūtne vēderiņu sistēmā pēc DT datiem (pakāpe)
18	Ārstēšana
19	Ārstēšanas rezultāts attiecībā uz pacienta dzīvību
20	Komplikācijas
21	Hemiparēzes pakāpe pēc 3 mēnešiem no slimības sakuma
22	Gaitas traucējumu pakāpe pēc 3 mēnešiem no slimības sakuma
23	Hemiparēzes pakāpe pēc 12 mēnešiem no slimības sakuma
24	Gaitas traucējumu pakāpe pēc 12 mēnešiem no slimības sakuma
25	Darba un sadzīves traucējumu pakāpe pēc 12 mēnešiem no slimības sakuma
26	Funkcionālā iznākuma pakāpe pēc Glazgo skalas pēc 12 mēnešiem no slimības sakuma
27	Pacienta novērošanas ilgums (gados) anketēšanas brīdi
28	Hemiparēzes pakāpe anketēšanas brīdi
29	Gaitas traucējumu pakāpe anketēšanas brīdi
30	Darba un sadzīves traucējumi anketēšanas brīdi
31	Funkcionālā iznākuma pakāpe pēc Glazgo skalas anketēšanas brīdi
32	Veselības atjaunošanās procesa dinamika anketēšanas brīdi
33	Invaliditāte anketēšanas brīdi

Tad seko eksperimentu plāns, datu apstrādes apraksts un smadzeņu spontāno iekšējo asinsizplūdumu dažu svarīgu faktoru prognozēšanas rezultāti, kas iegūti ar algoritmu palīdzību un apkopoti salīdzinošajā tabulā 4. Saskaņā ar ekspertu vērtējumu, interesantākie izpētes faktori, pēc kuriem vadoties iespējams veidot prognozi, ir:

1. ārstēšanas metodes izvēle (18);
2. ārstēšanas iznākums attiecībā uz pacienta dzīvību (19);
3. komplikācijas pirms- un pēcoperācijas periodā (20);
4. gaitas traucējumu pakāpe pēc 3 mēnešiem no saslimšanas (22) un anketēšanas brīdī (29);

5. darba un sadzīves traucējumu pakāpe pēc 12 mēnešiem (25) un anketēšanas brīdī (30);
6. funkcionālā iznākuma pakāpe pēc Glazgo skalas pēc 12 mēnešiem no saslimšanas (26) un anketēšanas brīdī (31).

4. tabula

Algoritmu darbības rezultātu salīdzinājums

Izmantojamā programma:			***Machine Learning Methods Comparison***			***MLC++***	
Prognozēšanas faktors	**Atribūti prognozes izveidošanai**	**Piemēru skaits**	***ID3, CV-5* kļūda, %**	***CART, CV-5* kļūda, %**	***CART2, CV-5* kļūda, %**	***ID3-GA, CV-5* kļūda,%**	***ID3, CV-5* kļūda, %**
18	1-17	200	23	61	21	-	-
19	1-18	200	6	20	4	8.10	14.36
20	1-18	200	13	13	7.5	12.00	17.76
22	1-18	148	-	-	-	56.58	63.37
25	1-18	148	-	-	-	55.50	61.37
26	1-18	148	37	62	34	35.50	39.29
29	1-18, 27	148	-	-	-	43.19	49.31
30	1-18, 27	148	-	-	-	50.61	58.85
31	1-18, 27	148	-	-	-	29.81	35.91

Tālāk tika veikta svarīgāko stāvokļu atribūtu novērtēšana un diagnosticēšanas likumu analīze, izmantojot ekspertu konsultācijas.

Pateicoties pārliecinošiem rezultātiem, kas tika iegūti, veicot prognozēšanu smadzeņu spontāno iekšējo asisnsizplūdumu jomā, tiek piedāvāts izveidot programmatūras kompleksu, lai veiktu statistiski pamatotas prognozes un ar tām palīdzētu šīs jomas ekspertiem.

Par programmatūras kompleksa bāzi izvēlēts *CART2* algoritms, kas sniedza labāko rezultātu eksperimentos. Tādējādi paredzēts realizēt divus mehānismus, kas varētu prognozēt:

- ārstēšanas iznākumu attiecībā uz pacienta dzīvību;
- komplikācijas pirms- un pēcoperācijas periodā.

Septītajā sadaļa ir doti secinājumi un ir aprakstītas piedāvāto algoritmu galvenās priekšrocības un trūkumi. Tāpat ir izklāstīta tālāko pētījumu perspektīva.

Pielikumos ir piedāvāts *ID3* un ĢA kombinētā klasifikatoru ansambļu izveides metodes testēšanas rezultāts 24 datu bāzēm, smadzeņu spontāno iekšējo asisnsizplūdumu datu anketēs paraugs un programmnodrošinājumu koncepcija.

DARBA PAMATREZULTĀTI

Lai risinātu problēmas, kas saistītas ar lokālā maksimuma un apmācību liela atribūtu skaita gadījumā šajā darbā, ir izmantota klasifikatoru ansambļu izveides tehnoloģija, kas balstās uz ĢA un meklēšanas heiristikas stratēģiju kombinēto izmantošanu. Lielu datu bāzu klasifikatoru izveidei un atjaunināšanai, kā arī ar nepilnīgiem un izkropļotiem datiem saistīto problēmu risināšanai tiek piedāvāts izmantot algoritmu *CART2*, kas ir induktīvā algoritma *CART* un M. Bongarda *CORA* metodes apvienojums.

Galvenie šī darba ieguvumi ir algoritma *CART2* izstrāde, kas piemērots klasifikatoru izveidei un atjaunināšanai lielu datu bāzu gadījumā, un jaunas klasifikatoru ansambļu izveides pieejas izstrāde. Darba procesā ir atrisināti vairāki uzdevumi:

1. Apskatīti un izanalizēti esošie zināšanu iegūšanas algoritmi un noskaidroti to galvenie ierobežojumi un trūkumi.
2. Piedāvāts kombinēts, uz *ID3* un ĢA bāzes veidots algoritms, kas paredzēts lokālā maksimuma problēmas pārvarēšanai.
3. Algoritma *CART* un Bongarda *CORA* metodes kombinācijas rezultātā izstrādāts algoritms *CART2*, kas paredzēts klasifikatoru izveidei un atjaunināšanai lielām datu bāzēm.
4. Izstrādāts programmnodrošinājums, kas ļauj salīdzināt mašīnapmācības algoritmus, izmantojot datu bāzes no dažādām zināšanu sfērām. Piedāvāta programmnodrošinājuma arhitektūra, kas ļauj pievienot jaunas metodes, nemainot pamatprogrammas struktūru.
5. Realizēta kombinētā, uz *ID3* un ĢA bāzes veidotā algoritma salīdzināšana ar citām mašīnapmācības metodēm, izmantojot 24 atšķirīgas datu bāzes no dažādiem priekšmetiskajiem apgabaliem.
6. Aprobēts algoritms *CART2*, lai risinātu ar nepilnīgiem un izkropļotiem datiem saistītus klasifikācijas uzdevumus.
7. Piedāvāta smadzeņu spontāno asinsizplūdumu prognozēšanai paredzēta programmnodrošinājuma koncepcija.

Autors ir gandarīts par pētījumā iegūtajiem rezultātiem un iespēju piedāvāt risinājumu aktuālajām ar mašīnapmācības jomu saistītajām problēmām.

TĀLĀKO PĒTĪJUMU VIRZĪBA

Eksperimentu pozitīvie rezultāti, kas gūti, veidojot lēmumu kokus uz ģenētisko algoritmu procedūru bāzes un izmantojot algoritmu *CART2* klasifikatoru inkrementālai atjaunināšanai un lielu datu bāzu klasifikatoru izveidei, ļauj spert nākamo soli to izpētē. Autors izvirzījis virkni eksperimentu turpmākiem pētījumiem, risinot reālā laika uzdevumus.

Veiktais pētījums ir pamudinājis autoru izvirzīt ideju par iespēju ģenētisko algoritmu lēmumu koku metodoloģiju analizēt arī saistībā ar citām mašīnapmācības metodēm, tajā skaitā *C4.5, C5.0* un *CART*, kuras šī darba kontekstā netika apskatītas, bet ir uzskatāmas par turpmāko pētījumu priekšmetiem.

Tāpat arī būtu interesanti realizēt un izpētīt abas pārējās darbā minētās pieejas, veidojot uz ģenētiskiem algoritmiem balstītus lēmumu kokus, respektīvi:

- Balsošanas procedūrā, kur par katru testējamo piemēru balso ansambļa klasifikatori, uzvar klase, kura savākusi vislielāko balsu skaitu.
- Uz klasifikatoru ansambļa bāzes tiek veidots kopējs lēmumu koks.

IZMANTOTĀS LITERATŪRAS SARAKSTS

1. Al-Attar A. A hybrid ĢA-heuristic search strategy / Al-Attar A. // AI Expert. – 1994. - September. - P. 34-37.
2. Aleksander I. An Introduction to Neural Computing / Aleksander I. and Morton H., London: Chapter and Hall. – 1991. – 240 p.
3. Alexopoulus E. Medical Diagnosis of Stroke Using Inductive Machine Learning / Alexopoulus E., Dounias G. D., Vemmos K. // Proceedings of the CHANIA'99 Conference on Machine Learning & Applications. – 1999.
4. Arciszewski T. Constructive Induction: The Key to Design Creativity / Arciszewski T., Michalski R., Wnek J. // Proceedings of the Third International Round-Table Conference on Computational Models of Creative Design, Heron Island, Queensland, Australia. – 1995. – December 3-7.
5. Arciszewski T. STAR Methodology-Based Learning about Construction Accidents and Their Prevention / Arciszewski T., Michalski R., Dybala T. // Automation in Construction. – 1995. – Vol. 4. – P. 75-85.
6. Belsley D. Regression Diagnostics: Identifying Influential Data and Sources of Collinearity / Belsley D., Kuh E. and Welsch R. – Canada: John Wiley & Sons, Inc., 1980. – 292 p.
7. Bloedorn E. Constructive Induction from Data in AQ17-DCI: Further Experiments / Bloedorn E. and Michalski R. // Artificial Intelligence Center, George Mason University, Fairfax, VA. – 1991.
8. Bloedorn E. Data-Driven Constructive Induction in AQ17-PRE: A Method and Experiments / Bloedorn E. and Michalski R. // Artificial Intelligence Center, George Mason University, Fairfax, VA. – 1991.
9. Borisov A. Research of Behaviour of Fuzzy Cora on selections of different Volumes / Borisov A. and Grekov R. // 5th International Conference on Applcation of Fuzzy Systems and Soft Computing "ICAFS-2002", Milan, Italy. – 2002. – P. 181-187.
10. Breiman L. Classifications and regression trees / Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. – Belmont, CA: Wadswonth International, 1984. – 358 p.
11. Buchanan B. G. Constructing an expert system. In Building Expert Systems / Buchanan B. G., Barstow D., Bechtel R., Bennet J., Clancey W., Kulikowski C., Mitchell T. M., Waterman D. A. (Hayes-Roth F, Waterman D.A. and Levat D., eds.) // Reading, MA: Addison-Wesley. – 1983. – Chapter 5.
12. Buntine W. Learning Classification Trees / Buntine W. // Statistic and Computing. – 1992. – Vol. 2. – P. 63-73.
13. Cartler C. Assesing Credit Card Applications Using Machine Learning / Cartler C. and Catlett J. // IEEE Expert. – 1997. – P. 71-79.
14. Clark P. The *CN2* Induction Algorithm / Clark P. and Niblett T. // Machine Learning. – 1989. – Vol. 3. – P. 261-283.
15. Dietterich T. G. Machine Learning Research: Four Current Directions / Dietterich T. G. // AI Magazine, 1997 18(4). – P. 97-136
16. Donoho S. Representing and Restructuring Domain Theories: a Constructive Induction

Approach / Donoho S. and Rendell L. // Artifical Intellegence Research. – 1995. – Vol. 2. – P. 411-446.

17. Dunham M.H. Data mining. Introductory and advanced topics. Prentice Hall, 2003. – 315 p.
18. Esposito F. A Comparative Analysis of Methods for Pruning Decision Trees / Esposito F., Malerbo D and Semeraro G. // IEEE Transactions on pattern analysis and Machine Intelligence. – 1997. – No. 5. – Vol. 19. – P. 476-491.
19. Gini G. Clustering and Classification Techniques to Assess Aquatic Toxicity / Gini G., Benfenati E., Boley D. // 4th International Conference of Knowledge-Based Inetlligent Engineering Systems & Allied Technologies. – Brighton, UK. – 30. Aug. – 1. Sep. 2000. – P. 166-172.
20. Gini G. Some Results for the Prediction of Carcinogenicity Using Hybrid Systems / Gini G., Lorenzini M., Vittore A. // Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools. AAAI 1999 Spring Symposium Series; Gini, G. C.; Katritzky, A. R., Eds.; AAAI Press, Menlo Park, CA, 1999/ - P. 139-143.
21. Graf J. Credit Scoring Based on Neural and Machine Learning / Graf J. and Nakhaeizadeh G. // Frontier Decision Support Concepts, Edited by V.L.Plantamura, B.Soucek and G.Visaggio, Chapter 14, 1994, John Wiley & Sons, Inc. P. 241-256.
22. Grekov R. Efficiency Evaluation of Fuzzy Cora Algorithm with Application of Cross Validation / Grekov R. // 3rd European Interdisciplinary School on Nonlinear Dynamics for System and Signal Analysis, Warsaw. – 2002.
23. Han J. Data Mining. Concept and Techniques / Han J. and Kamber M. // Morgan Kaufman Publishers. – 2000. – 550 p.
24. Hedberg S. Emerging Genetic Algorithms / Hedberg S. // AI Expert. – 1994. - September. - P. 25-28.
25. Holsheimer M. Data Minimg. The Search for Knowledge in Databases / Holsheimer M., Siebes A. // Report CS-R9406., Amsterdam., The Netherlands. – 1991. – 78 p.
26. Hong J. AQ15: Incremental Learning of Attribute-Based Descriptions from Examples the Method and User Guide/ Hong J., Mozetic I., Michalski R. // Report of the Intelligent Systems Group, UIUCDCS-F-86-949 Department of Computer Science, University of Illinois at Urbana-Champaign, Urbana, IL, 1986.
27. Ignizio P. Introduction to Expert Systems. The Development and Implementation of Rule-based Expert Systems / Ignizio P. // McGraw-Hill, 1991. – 402 p.
28. Kohavi R. *MLC*++ Tutorial: A Machine Learning Library of C++ classes / Kohavi R. // Stendford. – 1995. – 28 p.
29. Kohavi R. *MLC*++: A Machine Learning Library in C++ / Kohavi R., John G, Long R., Manley D. and Pfleger K. // Standford, CA: Computer Science Department Standford University. – 1994. – 4 p.
30. Kornienko J. Production rules induction algorithm based on the finish learning principle / Kornienko J. & Borisov A. // Fourth International Conference on Application of Fuzzy Systems and Soft Computing ICAFS'2000. Siegen, 2000. - June 27-29. – P. 287-292.
31. Kornienko Y. Application of genetic algorithms to classifier ensemble construction / Kornienko Y. // Scientific Proceedings of Riga Technical University, Information Technology and Management Science. Riga: RTU, 2003. – Issue 5. – Vol. 14. – P. 138-146.

32. Kornienko Y. Genetic-based decision trees / Kornienko Y. & Borisov A. // MENDEL'98 - 4th International Conference on Genetic Algorithms, Optimization Problems, Fuzzy Logic, Neural Networks and Rough Sets. Brno, 1998. – June 24-26. – P. 42-44.

33. Kornienko Y. Investigation of a hybrid algorithm for decision tree generation / Kornienko Y., Borisov A. // Proceedings of the Second IEEE International Workshop on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications, IDAACS'2003. Lviv, 2003. – September 8-10. – P. 63-68.

34. Kornienko Y. Symbolic inductive learning and genetic algorithms approach synergy / Kornienko, Y. & Borisov. A. // "Novadam un Latvijai" – Rēzeknes Augstskolas 5 gadu jubilejai veltītā zinātniski praktiskā konference. Rēzekne, 1998. – 9.-10. oktobris. – Lpp. 115-117.

35. Kornienko Y. The *CART* methodology for production rules induction / Kornienko Y. and Borisov A. // 5th International Conference on Soft Computing MENDEL'99. Brno, 1999. – June 9-12. – P. 362-366.

36. Kornienko Y. The *CART2* inductive algorithm in comparison with standard machine learning methods / Kornienko, Y. & Borisov. A. // Proceedings of the International Scientific-Technical Workshop "Problems of Transfer Technology". Ufa, 1999. – 30. September – 1. October. – P. 154-161.

37. Kornijenko J. Application of inductive diagnostic rules to intracerebral extravasation analysis / Kornijenko J., Dzenis J. & Borisov A. // Rīgas Tehniskās universitātes zinātniskie raksti. 5. sērija "Datorzinātne". 20.sējums "Informācijas tehnoloģija un vadības zinātne". Rīga: RTU, 2004. – Lpp. 36-42.

38. Kornijenko J. Using methods of inductive learning to forecast spontaneous intracerebral extravasation / Kornijenko J., Dzenis J. & Borisov A. // Rīgas Tehniskās universitātes zinātniskie raksti. 5. sērija "Datorzinātne". 5.sējums "Informācijas tehnoloģija un vadības zinātne". Rīga: RTU, 2001. – Lpp. 70-77.

39. Kornijenko Y. Application of genetic algorithms for generating decision trees / Kornijenko Y. and Borisov A. // International Conference on Parallel Computing in Electrical Engineering, PARELEC'98. Bialystok, 1998. – September 2-5. – P. 277-279.

40. Kramer S. Prediction of Ordinal Classes Using Regression Trees / Kramer S., Widmer G., Pfahringer B., De Groeve M. // Fundamenta Informaticae. – Vol. 34. – 2000. – P. 1-15.

41. Kubat M. Machine Learning for the Detection of Oil Spills in Satellite Radar Images / Kubat M., Holte R.C., and Matwin S. // Machine Learning. – 1998. – Vol. 30. – P. 195-215.

42. Kukar M. Prognosing the femoral neck fracture recovery with machine learning / Kukar M., Kononenko I. and Silvester T. // Proc. 17th International Conference on Information Technology Interfaces ITI'95, Pula, Croatia, 1995. – P. 103-109.

43. Kumar V. R. Performance Comparison of Different Learning Methods for Weather Forecasting Operations / Kumar V. R., Guignard P. and Chung C. Y. C. // Intelligent Systems, Kluwer Academic Publishers. – 1995. – P. 275-281.

44. Kumar V. R. Toward Building an Expert System for Weather Forecasting Operations / Kumar V. R., Chung C. Y. C. and Lindlay C. A. // Expert System with Applications. – 1994. – No. 2. – Vol. 7. – P. 373-381.

45. Letourneau S. Data Mining to Predict Aircraft Component Replacement / Letourneau S., Famili F., Matwin S. // IEEE Intelligent System. – 1999. – November/December. – P. 59-

66.

46. Mehta M. SLIQ: A Fast Scalable Classifier for Data Mining / Mehta M., Agrawal R. and Rissanen J. // Proceedings of 5th Int'l Conference on Extending Database Technology. – 1996. – P. 18-32.
47. Michalski R. S. Knowledge acquisition by encoding expert rules versus computer induction from examples: a case study involving soybean pathology / Michalski R. S. and Chilaysky R. L. // Int. J. Man: Machine Studies. – 1980. – Vol. 12. – P. 63-87.
48. Mingers J. An emperical comparison of selection measures for decision-tree induction / Mingers J. // Machine Learning, - 1989. - Vol. 3. - P. 319-342.
49. Mitchell M. An Introduction to Genetic Algorithms / Mitchell M. – MIT Press. – 1996. – 205 p.
50. Mitchell T.M. Machine Learning. / Mitchell T.M. – New York: Cornegie Mellon University. The McGraw-Hill Companies, Inc., 1989. – P. 405-414.
51. Murthy S.K. Automatic construction of decision trees from data: A multi-disciplinary survey. Data Mining and Knowledge Discovery, Vol.2, Issue 4, December, 1998, Kluwer Academic Publishers. – P. 345-389.
52. Nunez M. Decison Tree Induction Using Domain Knowledge / Nunez M. // Journal of Computing in Civil Engineering. – 1994. – No. 3. – Vol. 8. – P. 276-288.
53. Nunez M. The Use of Background Knowledge in Decision Tree Induction / Nunez M. // Machine Learning. – 1991. – Vol. 6. – P. 231-250.
54. Quinlan J. Inferring Decision Trees Using the Minimum Description Lenghts Principle / Quinlan J. and Riverst R. // Information and Computation. – 1989. – Vol. 80. – P. 227-248.
55. Quinlan J. R. The Minimum Description Length Principle and Categorical Theories / Quinlan J. R. // Proceedings 11th International Conference on Machine Learning, San Francisco Morgan Kaufman, New Brunswick, 1994. – P. 233-241.
56. Quinlan J. R. Unknown Attribute Values in Induction / Quinlan J. R. // Proceedings 6th International Machine Learning Workshop, Los Altos Morgan Kaufman, CA, 1989. – P. 164-168.
57. Quinlan J.R. Bagging, Boosting and *C4.5* / Quinlan J.R. // Proceedings 13th American Association for Artificial Intelligence National Conference on Artificial Intelligence, Menlo Park AAAI Press, CA, 1996. – P. 725-730.
58. Quinlan J.R. Discovering rules from large collections of examples. A case study. / Quinlan J.R. – Edinburg: In D.Michie (Ed), Expert Systems in The Micro Electronic Age, Edinburg University press, 1979. – P. 168-201.
59. Quinlan J.R. Imporved Use of Continuous Attributes in *C4.5* / Quinlan J.R. // Journal of Artifical Intelligence Research. – 1996. Vol. 4. – P. 77-90.
60. Quinlan J.R. Induction of Decision trees / Quinlan J.R. // Machine Learning 1. Kluwer Academic Publishers. – 1986. – P. 81-106.
61. Quinlan J.R. Learning Efficient Classifications Procedures and Their Application to Chess and Games / R.S. Michalski, J.Carbonell, T.Michell, editors // Machine Learning, An Artifical Intelligence Approach. San Mateo, CA: Morgan Kaufman, 1983. – Vol. 1. – P. 463-482.
62. Quinlan J.R. The Effect of Noise on Concept Learning / R.S. Michalski, J.Carbonell,

T.Michell, editors // Machine Learning, An Artifical Intelligence Approach. San Mateo, CA: Morgan Kaufman, 1986. – Vol. 2. – P. 149-166.

63. Rastogi R. PUBLIC: A Decision Tree Classifier that Integrates Building and Pruning / Rastorgi R. and Shirnu K. // VLDB'98, Proceedings of 24rd International Conference on Very Large Data Bases, August 24-27, 1998, New York, Morgan Kaufmann, 1998. – P. 404-415.
64. Shafer J. SPRINT: A Scalable Parallel Classifier for Data Minings / Shafer J., Agrawal R. and Mehta M. // Proceedings of the 22nd VLDB Conference. – 1996.
65. Tan M. Cost-Sensitive Learning of Classification Knowledge and its Applications in Robotics / Tan M. // Machine Learning. – 1993. – P. 7-33.
66. Tan P.N., Steinbach M., Kumar V. Introduction to data mining. Addison Wesley, 2005. – 769 p.
67. Turney P. Cost-Sensitive Classification: Emperical Evaluation of a Hybrid Genetic Decision Tree Induction Algorithm / Turney P. // Artifical Intelligence Research. – 1995. – Vol. 2. – P. 369-409.
68. Vafaie H. Genetic Algorithms as a Tool for Feature Selection in Machine Learning / Vafaie H. and De Jong K. // Fourth International Conference on Tools with Artificial Intelligence (ICTAI '92), November 10-13, Arlington, USA, 1992. – P. 200-203.
69. Weiss S.M. An Empirical Comparison of Pattern Recognition, Neural Nets, and Machine Learning Classification Methods / Weiss S.M. and Kapouleas I. // Proceedings of the Eleventh Internation Joint Conference of Artificial Intelligence. - 1989. - P. 177-183.
70. White R. L. Object Classification as a Data Analysis Tool / White R. L // Astronomical Data Analysis Software and Systems IX. – 2000. – Vol. 216. – P. 577-586.
71. Widmer G. Automatic Knowledge Base Refinement: Learning from Examples and Deep Knowledge in Rheumatology / Widmer G., Horn W., Nagele B. // Artificial Intelligence in Medicine. – 1993. – P. 225-243.
72. Yang J. Feature Subset Selection Using a Genetic Algorithms / Yang J. and Honavar V.
73. Бонгард М. М. Использование обучающейся программы для выявления нефтеносных пластов / Бонгард М. М., Вайнцвайг М.Н., Губерман Ш. А., Извекова М. Л., Смирнов М. С. // Геология и геофизика. – 1966. - № 6. – С. 96-109.
74. Бонгард М. М. Проблемы узнавания / Бонгард М. М. – М.: Наука, 1967. - 320 с.
75. Дзенис Ю. Синдром окклюзивной гидроцефалии при супратенториальных опухолях мозга срединной локализации (диагностика и хирургическое лечение). Автореферат на соискание ученой степени кандидата медицинских наук, Тарту, 1988.
76. Дюк В. Data Mining. Учебный курс / Дюк В. и Самойленко А. // Питер. – 2001. – 368 с.
77. Корниенко Ю.В. Экспериментальный анализ эффективности индуктивного алгоритма *CART2* / Корниенко Ю.В. и Борисов А.Н. // Автоматика и вычислительная техника. Рига: Институт электроники и вычислительной техники. – 2001. – № 1. – С. 34-42.
78. Реброва О. Ю. Статистический анализ медицинских данных: Применение пакета прикладных программ STATISTICA / Реброва О. Ю. // Медиа Сфера, Москва, 2002. – 312 с.