# Duomenų tyrybos sistemų galimybių tyrimas įvairių apimčių duomenims analizuoti

**Kotryna Paulauskienė**

Vilniaus universiteto Matematikos ir informatikos instituto doktorantė
Vilnius University, Institute of Mathematics and Informatics, Doctoral student
Akademijos g. 4, LT-08663 Vilnius
El. paštas: kotryna.paulauskiene@mii.vu.lt

**Olga Kurasova**

Vilniaus universiteto Matematikos ir informatikos instituto vyresn. mokslo darbuotoja
Vilnius University, Institute of Mathematics and Informatics, Senior researcher
Akademijos g. 4, LT-08663 Vilnius
El. paštas: olga.kurasova@mii.vu.lt

*Tobulėjant šiuolaikinėms informacinėms ir komunikacinėms technologijoms, sparčiai didėja apdorojamų ir saugomų duomenų kiekiai, todėl duomenų analizės uždavinys tampa vis sudėtingesnis, sunku daryti greitus, efektyvius ir teisingus sprendimus. Duomenų analizei dažnai pasitelkiama duomenų tyryba. Duomenų tyryba – tai procesas, kurio metu iš duomenų išgaunamos naudingos žinios. Duomenims apdoroti bei žinioms išgauti reikalingos duomenų tyrybos sistemos, leidžiančios apdoroti įvairios apimties duomenis. Tyrime siekiama nustatyti, kokios apimties duomenis per priimtiną laiką sugeba apdoroti populiariausios duomenų tyrybos sistemos. Nagrinėjamas ir lyginamas trijose atvirojo kodo duomenų tyrybos sistemose (WEKA, KNIME, ORANGE) įgyvendintų klasifikavimo ir klasterizavimo algoritmų skaičiavimo laikas, analizuojant skirtingos apimties duomenų aibes. Vertinant sistemas svarbus ne tik algoritmų skaičiavimo laikas, bet ir klasifikavimo bei klasterizavimo tikslumas, kurį pavyksta pasiekti per tą laiką, todėl straipsnyje pateikiamos ir eksperimentiniuose tyrimuose gauto tikslumo matų reikšmės.*

## Įvadas

Šiandieniame pasaulyje įvairiose srityse kaupiami dideli, nuolat augantys duomenų kiekiai. Šiuo metu net asmeniniai kompiuteriai leidžia saugoti tokius duomenų kiekius, kurių anksčiau nebuvo įmanoma saugoti dėl nepakankamos disko vietos. Duomenų vis daugėja, o santykinė dalis, kurią žmonės pajėgūs suprasti, gręsmingai mažėja (Witten, Frank, 2005). Analizuojant duomenis, dažnai pasitelkiama duomenų tyryba. Duomenų tyryba – tai procesas, kurio metu iš duomenų išgaunama informacija ir žinios, būtinos reikiamiems sprendimams priimti (Han, Kamber, 2006). Duomenų tyryba taikoma įvairiose srityse: versle ir komercijoje, inžinerijoje, telekomunikacijose, bankinėse ir draudimo sistemose, elektroninėje prekyboje, medicinoje ir kt. Duomenims apdoroti bei žinioms išgauti dažnai naudojamos duomenų tyrybos sistemos, leidžiančios apdoroti įvairios apimties duomenis. Kyla klausimas, kokie duomenys gali būti vadinami didelės apimties. Vienareikšmišką atsakymą į šį klausimą sunku rasti. Duomenys, kurie prieš kelerius metus buvo didelės apimties, atsiradus greitesniems duomenų apdorojimo įrenginiams ir metodams, tampa nedidelės apimties. Viena iš didelių apimčių duomenų apibrėžčių yra tokia: didelės apimties duomenimis galima laikyti tuos, su kuriais per priimtiną laiką nesusidoroja įprastos duomenų tyrybos sistemos, ir būtinos specialios, pritaikytos didelės apimties duomenims analizuoti, pavyzdžiui, pasitelkiant lygiagrečiuosius ir paskirstytuosius skaičiavimus bei debesų kompiuterijos technologijas. Priimtino laiko nustatymo uždavinys taip pat nėra elementarus. Tai priklauso nuo

sprendžiamo uždavinio specifikos ir norimo rezultatų tikslumo. Pavyzdžiui, analizuojant medicininius duomenis, labai svarbus tikslumas, todėl duomenų analizės rezultato tikslinga laukti kelias valandas ar net paras. Jei sprendžiamo uždavinio tikslumas nėra tiek svarbus, kiek rezultato radimo laikas, tik kelios sekundės gali būti laikomos priimtinu laiku.

Šio tyrimo objektas – įvairių apimčių duomenys ir duomenų tyrybos sistemos. Tyrimo tikslas – nustatyti, kokių apimčių duomenis per priimtiną laiką geba ištirti populiarios duomenų tyrybos sistemos, sprendžiant klasifikavimo ir klasterizavimo uždavinius. Tyrime taikoma informacijos paieškos, sisteminimo, analizės, lyginamosios analizės ir apibendrinimo metodika. Nagrinėjama klasifikavimo ir klasterizavimo algoritmų greitaveika naudojant skirtingos apimties duomenų aibes. Be sistemų skaičiavimo laiko, vertinami klasifikavimo ir klasterizavimo algoritmų tikslumo matai.

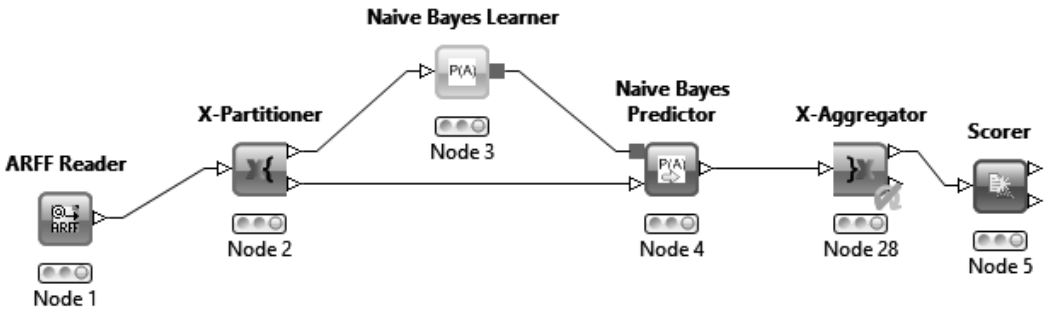
## 1. Duomenų tyrybos sistemos ir metodai

Straipsnyje nagrinėjamos ir lyginamos trys atvirojo kodo duomenų tyrybos sistemos:

- WEKA (Waikato Environment for Knowledge Analysis) (Hall et al.., 2009),
- KNIME (Konstanz Information Miner) (Berthold, 2007),
- ORANGE (Curk, 2005).

Tai vienos populiariausių duomenų tyrybos sistemų. Nors jos nėra pritaikytos didelės apimties duomenų apdorojimui ir analizei, jose įgyvendinti duomenų tyrybos metodai pajėgūs susidoroti su nemažomis duomenų aibėmis. Be to, šioms sistemoms naudoti nereikia specialių informatikos žinių. Greitai suvokiami jų naudojimo principai leido šioms sistemoms tapti populiariomis tarp įvairių sričių tyrėjų. Būtent dėl šių priežasčių minėtos sistemos pasirinktos tolesnei analizei. Šiose sistemose įgyvendintų klasifikavimo algoritmų gebėjimas teisingai klasifikuoti duomenis tiriamas darbe (Wahbeh et al., 2011), sistemų analizė atlikta darbe (Zupan, Demsar, 2008), tačiau ten nėra nustatyta, kokių apimčių duomenis sistemos pajėgios apdoroti ir analizuoti. Atvirojo kodo duomenų tyrybos sistemų taikymo sritys, vartotojų grupės, įgyvendinti algoritmai, vizualizavimo būdai ir kitos savybės vertinamos darbe (Madasamy, Tamilselvi, 2012), bet analizė naudojant įvairias duomenų aibes neatlikta. Autoriai (Chen, Williams, Xu, 2007) nurodo, kad WEKA, KNIME, ORANGE sistemos susidoroja su vidutinio dydžio duomenų aibėmis, tačiau ten nėra nurodoma, kokie duomenys vadinami vidutinio dydžio.

Tyrimuose naudotos šios sistemų versijos: WEKA 3.6.7, KNIME 2.7.3, ORANGE 2.6.1. Visose trijose sistemose yra įgyvendintas darbo eigos (angl. *workflows*) modulis, suteikiantis sistemoms patrauklumo. Vartotojas gali iš esamų mazgų (angl. *nodes*) sudaryti norimą schemą savo eksperimentams. Yra mazgų, skirtų įvairiems duomenų tyrybos algoritmams, pradiniam duomenų apdorojimui, rezultatų vizualizavimui ir kt. Sistemų intuityvios vartotojo sąsajos leidžia lengvai keisti darbo eigos modulius, įtraukiant ar šalinant mazgus, bei interaktyviai stebėti darbo eigos būseną ir analizės rezultatus. Vienos iš analizuojamų sistemų darbo eigos modulio pavyzdys pateikiamas 1 paveiksle.




*1 p a v.* ***KNIME sistemos darbo eigos modulio pavyzdys***

WEKA – atvirojo kodo programa, realizuota Java programavimo kalba (Hall et al., 2009). Ši sistema paprasta naudoti pradedančiajam vartotojui. WEKA sistemoje realizuoti įrankiai: duomenų pradinis apdorojimas, klasterizavimas, klasifikavimas, loginės taisyklės, regresija, vizualizavimas. Sistemos pagrindinė vartotojo sąsaja yra *Explorer*, be jos, dar įgyvendinta darbo eigos moduliu paremta sąsaja *Knoweldge Flow* ir komandų eilutė. Vartotojo sąsaja *Experimenter* leidžia vartotojui palyginti tarpusavyje kelių eksperimentų rezultatus, kai analizuojamos skirtingos duomenų aibės (Bouckaert et al., 2012).

KNIME – vartotojui draugiška atvirojo kodo duomenų apdorojimo, analizės ir vizualizavimo sistema, kurios veikimas taip pat grindžiamas darbo eigos moduliu. Sistemą sudaro per 1000 mazgų, kuriuos jungiant sukuriamos darbo eigos schemos. Be to, sistemoje yra integruoti visi WEKA sistemos moduliai (Berthold et al., 2007). KNIME sistema naudojasi daugiau nei 3000 organizacijų daugiau nei 60 pasaulio šalių.

ORANGE – atvirojo kodo duomenų analizės sistema, skirta ir pradedantiesiems, ir ekspertams (Curk, 2005). Sistemoje duomenų tyryba vykdoma naudojant darbų eigos sudarymo įrankį *Orange Canvas* arba programuojant *Phyton* kalba. ORANGE sistemoje realizuotas duomenų pradinis apdorojimas bei populiarūs klasifikavimo, klasterizavimo, vizualizavimo, loginių taisyklių, mokymo be mokytojo, regresijos metodai.

Toliau trumpai aptariami klasifikavimo ir klasterizavimo algoritmai, naudojami eksperimentiniuose tyrimuose. Pasirinkti populiariausi klasifikavimo ir klasterizavimo algoritmai, kurie yra įgyvendinti visose arba bent dviejose sistemose.

Naudojami šie klasifikavimo metodai:

- Bajeso klasifikatorius (angl. *Bayes classification*),
- *k* artimiausių kaimynų (angl. *k- nearest neighbours*),
- sprendimų medis (angl. *decision tree*),
- daugiasluoksnis neuroninis tinklas (angl. *multilayer perceptron*),
- atraminių vektorių klasifikatorius (angl. *support vector machine*).

Naudojami šie klasterizavimo metodai:

- *k* vidurkių (angl. *k- means*),
- hierarchinis klasterizavimas (angl. *hierarchical clustering*).

Naïve Bajeso klasifikatorius remiasi Bajeso taisykle. Laikoma, kad visi duomenų požymiai yra nepriklausomi ir kiekvienas iš požymių daro įtaką klasifikavimo rezultatui. Klasifikatorius skaičiuoja aposteriorines tikimybes kiekvienai klasei. Objektas priskiriamas tai klasei, kuri įgyja didžiausią aposteriorinę tikimybę (Dunham, 2003).

Sprendimų medžio algoritmo rezultatą galima pavaizduoti struktūra, panašia į medį, kurio kiekvienas išsišakojimas reiškia vienos ar kitos sąlygos tenkinimą. Taip sudaromos taisyklės, kurios leidžia nagrinėjamą duomenų aibę suklasifikuoti atsižvelgiant į požymių savybes (Dunham, 2003).

*k* artimiausių kaimynų metodo idėja yra naujo objekto palyginimas su mokymo aibės objektais, kurie yra panašūs į jį (Han, Kamber, 2006). Norint naują objektą priskirti kuriai nors klasei, skaičiuojami atstumai nuo to objekto iki visų mokymo aibės objektų. Dažniausiai naudojamas Euklido atstumas. Naujas objektas priskiriamas tai klasei, kuriai priklauso dauguma iš artimiausių *k* jo kaimynų.

Dirbtinio neuroninio tinklo struktūra primena biologinius neuroninius tinklus. Daugiasluoksnis neuroninis tinklas sudarytas iš kelių sluoksnių: įvesties, išvesties ir vieno ar daugiau paslėptų neuronų. Be kitų uždavinių, neuroniniai tinklai naudojami ir klasifikavimo uždaviniui spręsti. Tuomet įvesties sluoksnyje pateikiama požymius aprašanti informacija, o išvesties sluoksnyje gaunamas rezultatas – priklausymas klasėms.

Atraminių vektorių klasifikatorius – algoritmas, kuris transformuoja pradinius duomenis į didesnę dimensiją, kur randama hiperplokštuma, skirianti dvi klases kiek galima didesniu atstumu tarp klasifikuojamų duomenų (Han, Kamber, 2006). Radus šią hipreplokštumą, duomenis galima suskirstyti į dvi atskiras klases.

Hierarchinių klasterizavimo metodų rezultatai nusako klasterių tarpusavio hierarchiją, t. y. visi objektai laikomi vienu dideliu klasteriu, kurį sudaro mažesni klasteriai, šiuos – dar mažesni ir t. t. Taikant šiuos metodus, nustatoma bendra visų klasterių tarpusavio priklausomybių struktūra ir tik paskui sprendžiama, koks klasterių skaičius optimalus. Hierarchinis jungimo metodas smulkius klasterius jungia vis į stambesnius, kol galų gale lieka vienas (Čekanavičius, Murauskas, 2002).

*k* vidurkių metodas yra vienas iš nehierarchinių klasterinės analizės metodų. Nehierarchiniai metodai paprastai taikomi tada, kai iš anksto žinomas (pasirenkamas) klasterių skaičius ir norima klasterizuoti tiriamus objektus. Klasterizavimo procedūrą sudaro tokie žingsniai: 1) objektai suskirstomi į *k* pradinių klasterių; 2) paeiliui apskaičiuojamas kiekvieno objekto atstumas iki klasterių centrų (atstumas paprastai skaičiuojamas naudojantis Euklido metrika arba jos kvadratu); objektas priskiriamas artimiausiam klasteriui; perskaičiuojami klasterių centrai; 3) algoritmas kartojamas tol, kol daugiau nėra perskirstymų (Čekanavičius, Murauskas, 2002).

Visi nurodyti duomenų tyrybos metodai turi tam tikrus valdymo parametrus. Pirmoje lentelėje pateikiamos tyrime naudojamos parametrų reikšmės. Tos pačios parametrų reikšmės naudojamos visose tiriamose duomenų tyrybos sistemose. Pakeitus parametrų reikšmes, klasifikavimo ir klasterizavimo rezultatų absoliutūs dydžiai pasikeistų, tačiau rezultatų, gautų skirtingomis sistemomis, santykiai išliktų tie patys.

Klasifikavimo algoritmų rezultatams įvertinti naudojamas *q* blokų kryžminio patikrinimo metodas (angl. *q-fold cross validation*). Duomenų aibė yra suskaidoma į *q* nesusikertančių blokų. Algoritmas yra apmokomas naudojant *q*–1 bloko duomenis, o likusi duomenų dalis yra naudojama algoritmui testuoti, fiksuojamos klasifikavimo matų reikšmės. Ši procedūra atliekama *q* kartų, mokymui imant vis kitus *q*–1 blokus, pabaigoje randamos klasifikavimo matų vidutinės reikšmės (Han, Kamber, 2006; Witten, Frank, 2005). Tyrime pasirinktas blokų skaičius *q* yra 10. Klasifikavimo tikslumui nustatyti vertinami šie matai: jautrumas (angl. *sensitivity*), bendras klasifikavimo tikslumas (angl. *accuracy*), bendra klasifikavimo klaida (angl. *error*).

Apibrėžkime pagrindines sąvokas:

- tikrai teigiamas (TT) (angl. *true positive – TP*) – objektas $X_i$ priskirtas klasei $C_j$ ir iš tiesų jis jai priklauso;
- tikrai neigiamas (TN) (angl. *true negative –TN*) – objektas $X_i$ nepriskirtas klasei $C_j$ ir iš tiesų jis jai nepriklauso;
- klaidingai neigiamas (KN) (angl. *false negative – FN*) – objektas $X_i$ nepriskirtas klasei $C_j$, bet iš tiesų jis jai priklauso;
- klaidingai teigiamas (KT) (angl. *false true – FT*) – objektas $X_i$ priskirtas klasei $C_j$, bet iš tiesų jis jai nepriklauso.

Tada klasifikavimo kokybė yra apskaičiuojama pagal šias formules:

$$\text{jautrumas} = \frac{\text{TT skaičius}}{\text{TT skaičius} + \text{KN skaičius}},$$

*1 lentelė.* ***Metodų parametrai***

| Metodas | Parametrai |
|---|---|
| Bajeso klasifikatorius | Naïve Bajeso klasifikatorius |
| *k* artimiausių kaimynų | *k* = 3 |
| Sprendimų medis | Medžio tipas – C4.5, be genėjimo (sumažinimo), minimalus stebėjimų skaičius lape – 2 |
| Neuroninis tinklas | Vienas paslėptas sluoksnis su 10 neuronų, mokymosi epochų – 50 |
| Atraminių vektorių klasifikatorius | Naudotas tiesinis branduolys |
| *k* vidurkių metodas | Klasterių skaičius – 2; mokymosi epochų – 50 |
| Hierarchinis klasterizavimas | Klasterių skaičius – 2; atstumas tarp objektų – Euklido; atstumas tarp klasterių – vienetinė jungtis |

$$\text{bendras klasifikavimo tikslumas} = \frac{\text{TT skaičius} + \text{TN skaičius}}{\text{visų objektų skaičius}},$$

$$\text{klasifikavimo klaida} = 1 - \text{bendras klasifikavimo tikslumas}.$$

Klasterizavimo kokybei įvertinti parinktas klasterizavimo rezultatų su stebimomis klasėmis (angl. *classes to clusters evaluation*) patikrinimo metodas. Rezultatuose pateikiama neteisingai klasterizuotų objektų dalis procentais.

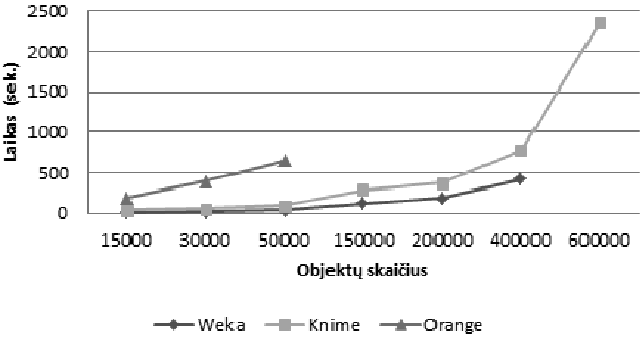
## 2. Eksperimentinių tyrimų rezultatai

Eksperimentams atlikti naudotas kompiuteris, kurio pagrindinės charakteristikos yra šios: operacinė sistema – Windows 8, operatyvioji atmintis (RAM) – 4 GB, procesorius – Intel i5-3317U, kurio taktinis dažnis – 1,7 GHz (Max Turbo dažnis 2,6 GHz). Atlikus eksperimentus naudojant kitų charakteristikų kompiuterį rezultatų skaitinių išraiškų absoliutūs dydžiai pasikeistų, tačiau išliktų toks pat santykis tarp skirtingomis sistemomis gautų rezultatų.

Eksperimentiniame tyrime siekiama išnagrinėti duomenų tyrybos sistemų galimybes analizuoti įvairaus dydžio duomenis ir nustatyti, kokių apimčių duomenų analizė negalima naudojant šias sistemas. Todėl buvo naudotos ne etaloninės duomenų aibės, skirtos duomenų tyrybos algoritmams vertinti, bet dirbtinai generuotos įvairių apimčių duomenų aibės, kurių požymių reikšmės tolygiai pasiskirsčiusios intervaluose (0; 1) ir (0,8; 2,2). Požymių skaičius fiksuotas – 100, objektų skaičius įvairus – 5000, 15 000, 30 000, 50 000, 150 000, 200 000, 400 000, 600 000. Objektai iš pirmo intervalo priskiriami pirmajai klasei, iš antro – antrajai. Duomenų intervalai parinkti tokie, kad skirtingų klasių duomenys iš dalies susiklotų, kaip dažniausiai yra realiose situacijose. Pasirinktas vienodas visų duomenų aibių požymių skaičius (lygus 100), nes toliau aprašytais eksperimentais norėta parodyti, kaip duomenų tyrybos sistemų pajėgumai priklauso nuo analizuojamų objektų skaičiaus, o ne nuo juos charakterizuojančių požymių skaičiaus. Pasirinkus kitą požymių skaičių, rezultatų absoliutūs dydžiai pasikeistų, tačiau skirtingomis sistemomis gautų rezultatų santykiai išliktų tie patys.

### 2.1. Klasifikavimo rezultatai

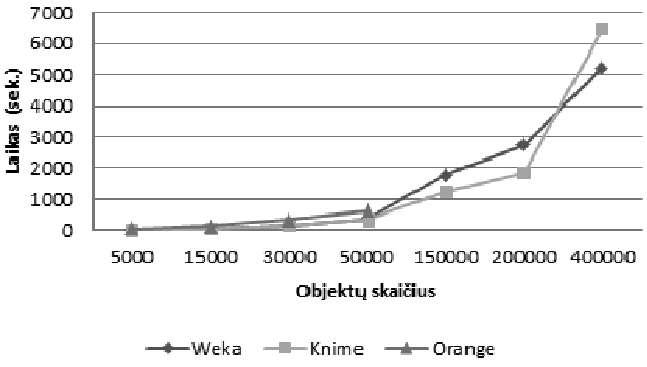
Dėl savo paprastumo Naïve Bajeso klasifikatorius visose lyginamose sistemose gana greitai gauna klasifikavimo rezultatą. Suklasifikuoti 50 000 objektų WEKA sistema užtrunka 31 sek., KNIME – 93 sek., tačiau ORANGE prireikia beveik 11 min. (2 pav.). Toliau didinant objektų skaičių iki 150 000, ORANGE sistema išsijungia dėl kompiuterio operatyviosios atminties trūkumo. Padidinus objektų skaičių iki 400 000, WEKA sistema suklasifikuoti duomenis užtrunka šiek tiek daugiau nei 7 min., o KNIME – kiek mažiau nei 13 min. WEKA sistema naudojant 600 000 objektų sudaro Naïve Bajeso modelį, tačiau pradėjus kryžminį patikrinimą sistema praneša apie klaidą, kad nepakanka kompiuterio operatyviosios atminties, o KNIME susidoroja su duomenimis per 39 min. Kaip matyti iš pateiktų rezultatų (2 pav.), WEKA ir KNIME sistemos gerai susidoroja su duomenimis, sudarytais iš maždaug 400 000 objektų, kai sprendžiamas klasifikavimo uždavinys naudojant Naïve Bajeso klasifikatorių. Šį faktą paaiškina tai, kad klasifikatorius nėra iteracinis, todėl rezultatas gaunamas gana greitai – skaičiavimai užtrunka ne daugiau kaip 13 minučių.



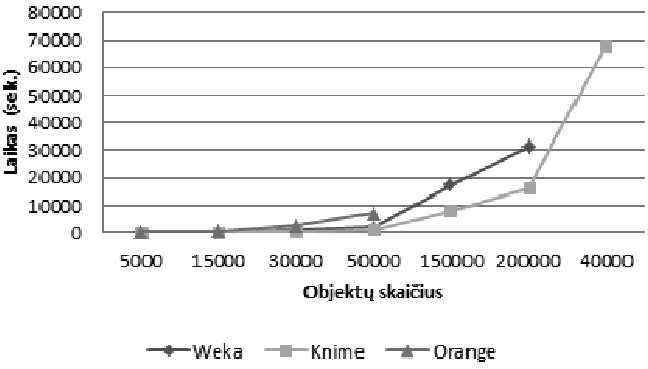

*2 p a v.* ***Naïve Bajeso klasifikatoriaus vykdymo laiko priklausomybė nuo klasifikuojamų objektų skaičiaus***

Analizuojant duomenų aibes iki 50 000 objektų visos nagrinėjamos sistemos sprendimų medžiui sudaryti ir duomenims suklasifikuoti užtrunka iki 10 min. (3 pav.). Didinant duomenų apimtį, klasifikavimo laikas didėja. Naudojant 400 000 objektų aibę WEKA sistema klasifikavimo rezultatus gauna po 1 val. 27 min., KNIME – po 1 val. 48 min., o ORANGE sistema jau nesusidoroja su 150 000 objektų aibe. 600 000 objektų aibė yra per didelė, ir WEKA bei KNIME sistemose pranešama apie kompiuterio operatyviosios atminties trūkumą.

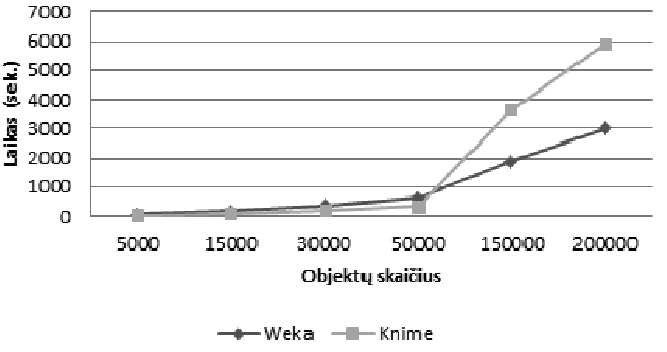
KNIME sistemai prireikė mažiausiai laiko *k* artimiausių kaimynų metodui ($k = 3$) įvykdyti (4 pav.). Nustatyta, kad KNIME užtrunka beveik 19 min., kol suklasifikuoja 50 000 objektų, WEKA – šiek tiek daugiau nei pusvalandį, o ORANGE – beveik dvi valandas. Objektų skaičių padidinus tris kartus, t. y. iki 150 000 objektų, KNIME vykdymo laikas pailgėja 7,2 karto, WEKA – beveik 9 kartus, o ORANGE sistema išsijungia dėl kompiuterio operatyviosios atminties trūkumo. KNIME ir WEKA sistemos suklasifikuoja ir 200 000 objektų aibę, tačiau tai trunka atitinkamai 4 val. 30 min. ir 9 val., ir toks laikas jau dažnai nėra priimtinas tyrėjui. Naudojant 400 000 objektų aibę KNIME sistema objektus suklasifikuoja per 18 val. 55 min., o WEKA sistema po 20 val. darbo įvykdo tik 50 proc. skaičiavimų, taigi toliau vykdyti eksperimentinius skaičiavimus naudojant 600 000 objektų aibę nebuvo prasminga.




*3 p a v.* ***Sprendimų medžio vykdymo laiko priklausomybė nuo klasifikuojamų objektų skaičiaus***

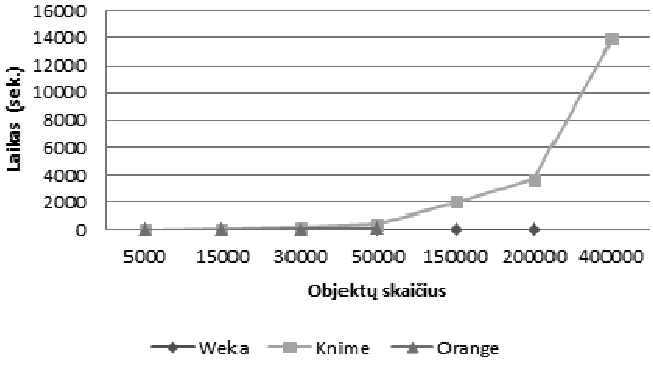



*4 p a v.* ***k artimiausių kaimynų metodo vykdymo laiko priklausomybė nuo klasifikuojamų objektų skaičiaus***




*5 p a v.* ***Neuroninio tinklo vykdymo laiko priklausomybė nuo klasifikuojamų objektų skaičiaus***

Neuroninio tinklo metodas yra įgyvendintas tik KNIME ir WEKA sistemose. Lyginant šio metodo vykdymo laiką nustatyta, kad naudojant duomenų aibes iki 50 000 objektų, sistemų vykdymo laikai skiriasi nedaug, tačiau peržengus 50 000 objektų aibės ribą, KNIME vykdymo laikas gerokai padidėja ir 200 000 objektų suklasifikuoja per 1 val. 38 min., WEKA – 49 min. (5 pav.). Naudojant 400 000 objektų aibę tiek WEKA, tiek KNIME sistemos praneša apie kompiuterio operatyviosios atminties trūkumą.




*6 p a v.* ***Atraminių vektorių klasifikatoriaus vykdymo laiko priklausomybė nuo klasifikuojamų objektų skaičiaus***

WEKA sistemoje atraminių vektorių klasifikatoriaus vykdymo laikas, naudojant tyrimo duomenų aibes iki 200 000 objektų, yra labai trumpas. Net 200 000 objektų suklasifikuojama per 27 sek. (6 pav.). Analizuojant 400 000 objektų aibę, WEKA sistemai pritrūksta kompiuterio operatyviosios atminties. KNIME sistema suklasifikuoja 400 000 objektų, nors tai užtrunka 3 val. 51 min. Naudojant 600 000 objektų aibę ir KNIME sistemai nepakanka kompiuterio operatyviosios atminties. Atraminių vektorių klasifikatoriaus vykdymo laikas ORANGE sistemoje yra trumpesnis nei KNIME, analizuojant tyrimo aibes iki 50 000 objektų. 50 000 objektų aibę KNIME suklasifikuoja per 6 min. 31 sek., ORANGE sistema užtrunka kiek mažiau nei 2 min. Naudojant 150 000 objektų aibę ORANGE sistema išsijungia dėl kompiuterio operatyviosios atminties trūkumo.

Vertinant skaičiavimo laiką, tikslinga vertinti ir klasifikavimo kokybę. Klasifikavimo kokybės matai parodė, kad WEKA sistema *k* artimiausių kaimynų, atraminių vektorių ir neuroninio tinklo klasifikatoriais visus duomenis klasifikuoja 100 proc. tikslumu (2 lentelė). Naïve Bajeso klasifikatoriaus teisingai suklasifikuotų stebėjimų dalis kito nuo 96,48 proc. iki 97,60 proc., o sprendimų medžio – nuo 99,40 proc. iki 99,97 proc.

KNIME sistema *k* artimiausių kaimynų ir atraminių vektorių klasifikatoriais visus duomenis suklasifikuoja 100 proc. tikslumu (3 lentelė). Naïve Bajeso klasifikatoriaus teisingai suklasifikuotų objektų dalis kito nuo 92,22 proc. iki 97,50 proc., sprendimų medžio ir neuroninio tinklo – atitinkamai 99,02–99,97 proc. ir 99,66–99,87 proc.

*2 l e n t e l ė.* ***WEKA sistemos klasifikavimo kokybės matų reikšmės***

| Metodas | Klasė | Jautrumas | Bendras klasifikavimo tikslumas, % | Bendra klasifikavimo klaida, % |
|---|---|---|---|---|
| Naïve Bajeso klasifikatorius | I klasė | 1 | 96,48–97,60 | 2,4–3,52 |
| | II klasė | 0,934–0,968 | | |
| *k* artimiausių kaimynų klasifikatorius | I klasė | 1 | 100 | 0 |
| | II klasė | 1 | | |
| Neuroninis tinklas | I klasė | 1 | 100 | 0 |
| | II klasė | 1 | | |
| Sprendimų medis | I klasė | 0,996–1 | 99,40–99,97 | 0,03–0,60 |
| | II klasė | 0,991–1 | | |
| Atraminių vektorių klasifikatorius | I klasė | 1 | 100 | 0 |
| | II klasė | 1 | | |

*3 lentelė.* ***KNIME sistemos klasifikavimo kokybės matų reikšmės***

| Metodas | Klasė | Jautrumas | Bendras klasifikavimo tikslumas, % | Bendra klasifikavimo klaida, % |
|---|---|---|---|---|
| Naïve Bajeso klasifikatorius | I klasė | 1 | 92,22–97,50 | 2,50–7,78 |
| | II klasė | 0,89–0,97 | | |
| *k* artimiausių kaimynų klasifikatorius | I klasė | 1 | 100 | 0 |
| | II klasė | 1 | | |
| Neuroninis tinklas | I klasė | 0,998–0,999 | 99,66–99,87 | 0,13–0,34 |
| | II klasė | 0,995–0,998 | | |
| Sprendimų medis | I klasė | 0,996–1 | 99,02–99,97 | 0,23–0,98 |
| | II klasė | 0,984–1 | | |
| Atraminių vektorių klasifikatorius | I klasė | 1 | 100 | 0 |
| | II klasė | 1 | | |

*4 lentelė.* ***ORANGE sistemos klasifikavimo kokybės matų reikšmės***

| Metodas | Klasė | Jautrumas | Bendras klasifikavimo tikslumas, % | Bendra klasifikavimo klaida, % |
|---|---|---|---|---|
| Naïve Bajeso klasifikatorius | I klasė | 1 | 97,34–97,62 | 2,38–2,66 |
| | II klasė | 0,947–0,952 | | |
| *k* artimiausių kaimynų klasifikatorius | I klasė | 1 | 100 | 0 |
| | II klasė | 1 | | |
| Sprendimų medis | I klasė | 0,997–0,999 | 99,06–99,89 | 0,11–0,94 |
| | II klasė | 0,984–998 | | |
| Atraminių vektorių klasifikatorius | I klasė | 1 | 100 | 0 |
| | II klasė | 1 | | |

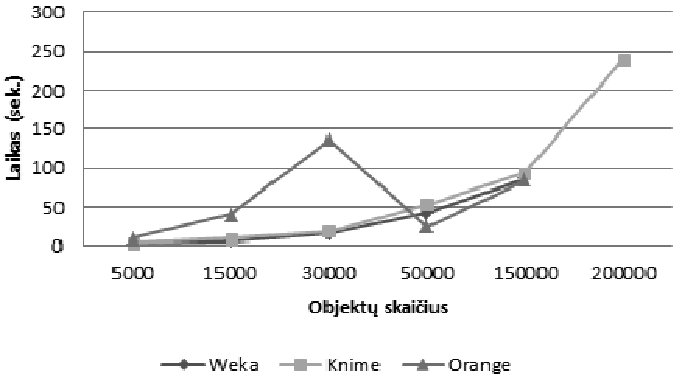
ORANGE sistemoje Naïve Bajeso klasifikatoriaus teisingai suklasifikuotų objektų dalis kito nuo 97,34 proc. iki 97,62 proc., o sprendimų medžio – nuo 99,06 proc. iki 99,89 proc. (4 lentelė). *k* artimiausių kaimynų ir atraminių vektorių klasifikatoriais klaidingai suklasifikuotų objektų visai nėra analizuojant visas duomenų aibes iki 50 000 objektų.

### 2.2. Klasterizavimo rezultatai

WEKA, KNIME, ORANGE sistemos, naudodamos hierarchinio klasterizavimo metodą, 500 ir 1000 objektų aibes suklasterizuoja per kelias sekundes (5 lentelė), o klasterizavimo rezultatai 100 proc. sutampa su stebimomis objektų klasėmis, t. y. visus pirmos klasės objektus priskiria prie vieno klasterio, antros – prie kito. Daug ilgiau užtrunka 5000 objektų klasterizavimas, be to, klasterizavimo kokybė yra ypač prasta, nes vienas objektas priskiriamas pirmajam klasteriui, o likę 4999 objektai – antrajam. Naudojant 15 000 objektų aibę hierarchinio klasterizavimo metodo veikimas WEKA ir ORANGE sistemose sustoja dėl kompiuterio operatyviosios atminties trūkumo, o KNIME sistema po 9 val. darbo įvykdo tik 14 proc. skaičiavimų, todėl laikoma, kad tai nėra priimtinas laikas, ir skaičiavimai sustabdomi. Hierarchinio klasterizavimo metodo trūkumas – atstumų matricai apskaičiuoti bei jos elementams išsaugoti reikia daug išteklių. Dideliems masyvams (>300) klasterizuoti dažnai naudojami nehierarchiniai klasterizavimo metodai (Čekanavičius, Murauskas, 2002).

*5 lentelė.* ***Hierarchinio klasterizavimo metodo vykdymo laiko sekundėmis priklausomybė nuo klasterizuojamų objektų skaičiaus***

| Objektų skaičius | WEKA | KNIME | ORANGE |
|---|---|---|---|
| 500 | 2 | 4 | 1 |
| 1000 | 6 | 30 | 4 |
| 5000 | 1059 | 4554 | 265 |




*7 p a v.* ***k vidurkių metodo vykdymo laiko priklausomybė nuo klasterizuojamų objektų skaičiaus***

*k* vidurkių metodu nagrinėjamų duomenų aibių objektai suklasterizuojami į du klasterius visomis trimis sistemomis labai greitai (7 pav.). WEKA ir KNIME sistemoms suklasterizuoti 15 0000 objektų prireikia apie 90 sek. WEKA sistema 200 000 objektų aibės nesuklasterizuoja dėl kompiuterio operatyviosios atminties trūkumo. KNIME sistema 200 000 objektų suklasterizuoja per 4 min., o naudojant 400 000 objektų aibę, pritrūksta kompiuterio operatyviosios atminties. ORANGE sistema nepajėgia klasterizuoti 200 000 ir daugiau objektų aibių dėl kompiuterio operatyviosios atminties trūkumo, be to, ir mažesnės apimties duomenis ji klasterizuoja lėčiau nei kitos sistemos, išskyrus atvejus, kai objektų skaičius 50 000 ir 150 000.

*6 l e n t e l ė.* ***k vidurkių metodo klasterizavimo rezultatų palyginimas su duomenų klasėmis: neteisingai klasterizuotų stebėjimų dalis (proc.)***

| Objektų skaičius | Sistema | | |
|---|---|---|---|
| | WEKA | KNIME | ORANGE |
| 5 000 | 1,9 | 1,9 | 4,0 |
| 15 000 | 1,7 | 1,7 | 2,0 |
| 30 000 | 1,6 | 1,6 | 1,9 |
| 50 000 | 2,5 | 2,4 | 4,3 |
| 150 000 | 1,7 | 1,7 | 2,0 |
| 200 000 | * | 2,3 | * |

***** trūksta kompiuterio operatyviosios atminties

Palyginus *k* vidurkių metodo klasterizavimo rezultatus su duomenų klasėmis pastebėta, kad KNIME ir WEKA sistemų neteisingai klasterizuotų objektų dalis, analizuojant visas duomenų aibes, vienoda arba beveik vienoda – kinta nuo 1,6 proc. iki 2 proc. (6 lentelė). ORANGE sistemos neteisingai klasterizuotų stebėjimų dalis didesnė (1,9–4,3 proc.). Čia neteisingai klasterizuotų objektų dalis buvo apskaičiuojama taip: pradžioje suskaičiuojama, kiek vienam klasteriui yra priskirta objektų iš kitos klasės nei dauguma to klasterio objektų; tuomet apskaičiuojama procentinė dalis nuo visų tos klasės objektų skaičiaus; skaičiavimai atliekami abiem klasteriams ir gauti rezultatai sumuojami.

## Išvados

Atliktas tyrimas parodė, kad ORANGE sistemą galima naudoti kaip duomenų tyrybos įrankį analizuojant duomenų aibes iki 50 000 objektų, kai kiekvieną objektą charakterizuoja 100 požymių. Norint atlikti didesnės aibės analizę, vertėtų rinktis WEKA arba KNIME sistemą. Analizuojant duomenų aibes iki 50 000 objektų, nustatytas panašus WEKA ir KNIME sistemų skaičiavimo laikas vykdant visus nagrinėtus algoritmus, o ORANGE sistema užtrunka ilgiau tiems patiems skaičiavimams atlikti. Nors WEKA sistemai reikia mažiau laiko atlikti skaičiavimams taikant didžiąją dalį nagrinėtų algoritmų, tačiau KNIME sistema tam tikrais atvejais pajėgi apdoroti didesnės apimties duomenis nei WEKA. Galima teigti, kad turint tik ORANGE sistemą didelės apimties duomenys yra tie, kurie sudaryti iš daugiau nei 50 000 objektų. Analizuojant duomenis WEKA ar KNIME sistemomis, didesnės nei 200 000 objektų duomenų aibės jau yra didelės apimties, nors naudojant nesudėtingus klasifikavimo metodus pastarosios dvi sistemos pajėgios apdoroti

ir didesnės apimties duomenis – 400 000 objektų, o KNIME – dar ir 600 000 objektų. Jei duomenų apimtys yra didesnės, būtinos didelėms duomenų aibėms pritaikytos duomenų tyrybos sistemos, pajėgios pasitelkti lygiagrečiuosius ir paskirstytuosius skaičiavimus. Be abejo, tyrimams naudojant kitų charakteristikų kompiuterį, rezultatai skirtųsi, tačiau bendros tendencijos išliktų, t. y. KNIME ir WEKA sistemos būtų pranašesnės už ORANGE.

Tyrimo rezultatai parodė, kad taikyti klasifikavimo metodai duoda tikslius klasifikavimo rezultatus, sprendžiant testinį uždavinį, kai klasės tik šiek tiek susikloja. Įprastai praktikoje kylančiuose uždaviniuose klasių sanklota būna didesnė, todėl klasifikavimo rezultatai gali būti ir šiek tiek blogesni. Prieš pasirenkant duomenų tyrybos sistemą derėtų ne tik atsižvelgti į turimų duomenų aibės dydį, bet ir įvertinti pasirinktų algoritmų sudėtingumą, kuris daro įtaką skaičiavimo laikui, nes mažai skaičiavimų reikalaujantis algoritmas gali susidoroti ir su didesnės apimties duomenimis, skaičiavimams imlus algoritmas gali užstrigti analizuojant ir mažesnę duomenų aibę.

Ateityje būtina atlikti eksperimentinius tyrimus naudojant įvairesnius, daugiau nei dviejų klasių duomenis, sudarytus iš įvairių požymių skaičių. Taip pat verta nagrinėti ir kitas populiarias duomenų tyrybos sistemas. Tai leistų daryti tikslesnes išvadas apie sistemų galimybes analizuoti įvairių apimčių duomenis.

**LITERATŪRA**


BERTHOLD, M.; CEBRON, N.; DILL, F.; GABRIEL, T., KOTTER, T.; MEINL, T., et al. (2008). KNIME: The Konstanz Information Miner. *Data Analysis, Machine Learning and Applications*. In: *Studies in Classification, Data Analysis, and Knowledge Organization (GfKL).* Freiburg: Springer, p. 319–326.

BOUCKAERT R. R.; FRANK E.; HALL M.; KIRKBY R.; REUTEMANN P.; SEEWALD A.; SCUSE D. (2012). *WEKA Manual for Version 3-6-7* [interaktyvus] [žiūrėta 2013 m. kovo 15 d.]. Prieiga per internetą: <http://www.cs.waikato.ac.nz/ml/weka/documentation.html/>.

CHEN, X.; YE, Y.; WILLIAMS, G.; XU, X. (2007). A Survey of Open Source Data Mining Systems. Emerging Technologies in Knowledge Discovery and Data Mining. *PAKDD 2007, International Workshops, Nanjing, China, May 22–25, 2007, Revised Selected Papers,* Lecture Notes in Computer Science, vol. 4819, p. 3–14.

CURK, T.; DEMŠAR, J.; XU, Q.; LEBAN, G.; PETROVIČ, U.; BRATKO, I., et al. (2005). Microarray data mining with visual programming. *Bioinformatics*, vol. 21(3), p. 396–398.

ČEKANAVIČIUS, V.; MURAUSKAS G. (2002). *Statistika ir jos taikymai*. II dalis. Vilnius: TEV. 268 p. ISBN 9955-491-16-7.

DUNHAM, M. H. (2003). *Data Mining Introductory and Advanced Topics*. New Jersey: Pearson Education, Inc. Prentice Hall. 315 p. ISBN 0-13-088892–3.

HALL, M.; FRANK, E.; HOLMES, G.; PFAHRINGER, B.; REUTEMANN, P.; WITTEN, I. H. (2009). The WEKA Data Mining Software: An Update. *SIGKDD Explorations*, vol. 11(1), p. 10–18.

HAN, J.; KAMBER, M. (2006). *Data Mining Concepts and Techniques*. Second Edition. San Francisco, CA: Morgan Kaufmann Publishers. 743 p. ISBN 13:978-1-55860-901-3.

MADASAMY, B.; TAMILSELVI, J. J. (2012). Assesement of Freeware Data Mining Tools over Some Wide-Range Characteristics. In: *International conference on information processing, Wireless Networks and Computational Intelligence, ICIP 2012,* Communications in Computer and Information Science*,* vol. 292, p. 529–535.

WAHBEH, A. H.; AL-RADAIDEH Q. A.; AL-KABI, M. N.; AL-SHAWAKFA, E. M. (2011). A Comparison Study between Data Mining Tools over some Classification Methods. *International Journal of Advanced Computer Science and Applications, Special Issue on Artificial Intelligence*, vol. 0(3), p. 18–25.

WITTEN, I. H.; FRANK E. (2005). *Data Mining: Practical Machine Learning. Tools and Techniques*. Second Edition. San Francisco, CA: Morgan Kaufmann Publishers. 525 p. ISBN 0-12-088407-0.

ZUPAN, B.; DEMSAR, J. (2008). Open-Source Tools for Data Mining. *Laboratory and Clinical Medicine*, vol. 28, p. 37–54.

**INVESTIGATION OF THE ABILITIES OF DATA MINING SYSTEMS TO ANALYSE VARIOUS VOLUME DATASETS**

**Kotryna Paulauskienė, Olga Kurasova**

Summary

The aim of the paper is to determine what volume of data the popular data mining systems are able to analyse within a reasonable period of time, when solving classification and clustering problems. Three open source data mining systems are investigated: WEKA, KNIME, and ORANGE. The experiments have been carried out with eight datasets, where the number of attributes was fixed – 100 and the number of instances ranged between 5000 and 600 000. The experimental investigation has shown that when the ORANGE system is used, the data of more than 50 000 instances are of too large volume. In order to analyse larger datasets, the WEKA and KNIME systems need to be used. The data of more than 200 000 instances are of too large volume for WEKA and KNIME, however, when simple classification methods are used, both systems are able to handle 400 000 instances, and KNIME – 600 000 instances. The results have showed that KNIME can handle larger datasets than WEKA, when applying some classification methods. The accuracy of classification is high enough, when the classification methods, implemented in the systems, are used.