Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
55 t., 2014, 60–65

ISSN 0132-2818
www.mii.lt/LMR/

Programinės sistemos duomenų tyrybos mokymui

Olga Kurasova[1,2]

[1] *Vilniaus universitetas, Matematikos ir informatikos institutas*
Akademijos g. 4, LT-08663, Vilnius

[2] *Lietuvos edukologijos universitetas, Gamtos, matematikos ir technologijų fakultetas*
Studentų g. 39, LT-08106, Vilnius

E. paštas: olga.kurasova@mii.vu.lt

Santrauka. Šiame straipsnyje apžvelgiamos programinės sistemos, kurios gali būti naudojamos duomenų tyrybos dalykui mokyti(-s). Įprastai matematinės statistikos mokyme naudojamos tokios sistemos, kaip SPSS Modeler (Clementine), Statistica, SAS/STAT. Jos turi daug funkcijų, yra daugiau orientuotos į matematinės statistikos sprendžiamus uždavinius, todėl duomenų tyrybos mokymui ne visada tinka. Duomenų tyrybos mokymui tinkamesnės sistemos: WEKA, Orange, KNIME, RapidMiner.

Raktiniai žodžiai: duomenų tyryba, klasifikavimas, klasterizavimas, mokslinių darbų seka.

Įvadas

Šiuo metu yra sukurtos ir toliau intensyviai tobulinamos įvairios duomenų tyrybos sistemos, kurios skiriasi viena nuo kitos savo funkcionalumu, grafinėmis vartotojų sąsajomis ir kitais aspektais. Populiariausios yra WEKA, Orange, KNIME, RapidMiner. Dažniausiai tokias sistemas naudoja tyrėjai, spręsdami sudėtingus duomenų tyrybos uždavinius, tačiau jos gali būti naudojamos ir duomenų tyrybos ar su ja susijusių dalykų mokyme. Mokymui naudojamų sistemų kokybė turi būti vertinama pagal kitus kriterijus nei renkantis sistemą sudėtingiems duomenų tyrybos uždaviniams spręsti. Besimokančiajam daug svarbiau yra sistemos draugiškumas, patogumas, nuorodų aiškumas ir panašios savybės, nei tose sistemose įgyvendintų algoritmų tikslumas. Sistemų naudojimas turi padėti suprasti duomenų tyrybos algoritmus, jų veikimą. Todėl dėstytojui svarbu parinkti tinkamas sistemas mokymui. Straipsnyje pateikiami patarimai, į ką turi atkreipti dėmesį dėstytojai ar besimokantieji, renkantis duomenų tyrybos sistemas.

1 Duomenų tyrybos uždaviniai

Duomenų tyryba (angl. *data mining*) – tai procesas, kurio metu, naudojant įvairius duomenų analizės įrankius, bandoma nustatyti ir atrasti „užslėptas“ duomenų struktūras ir ryšius. Duomenų tyrybos tikslas – iš įprastai didelių duomenų aibių išgauti svarbią informaciją ir atmesti nereikšmingą. Duomenų tyrybos metodai padeda „nepasiklysti“ duomenų ir informacijos gausoje. Duomenų tyryba atliekama žinių radimo duomenų aibėse procese (angl. *knowledge discovery in databases*), kurio metu didelių

apimčių duomenų aibėse ieškoma naujos informacijos, padėsiančios įgyti žinias apie analizuojamus duomenis ir priimti tinkamiausius sprendimus [5].

Pasitelkus duomenų tyrybą sprendžiami klasifikavimo, klasterizavimo, prognozavimo, susietumo taisyklių paieškos uždaviniai. Didžioji dalis duomenų tyrybos metodų yra grindžiami matematine statistika. Duomenų tyryba praplečia matematinės statistikos galimybes. Dažnai vienu metu sprendžiami keli uždaviniai, pavyzdžiui, klasifikavimo, klasterizavimo ir prognozavimo, siekiant gauti kiek galima daugiau žinių apie analizuojamus duomenis [3]. Todėl svarbu turėti programines sistemas, kuriose įgyvendinti įvairūs duomenų tyrybos uždavinius sprendžiantys metodai.

2 Duomenų tyrybos sistemos

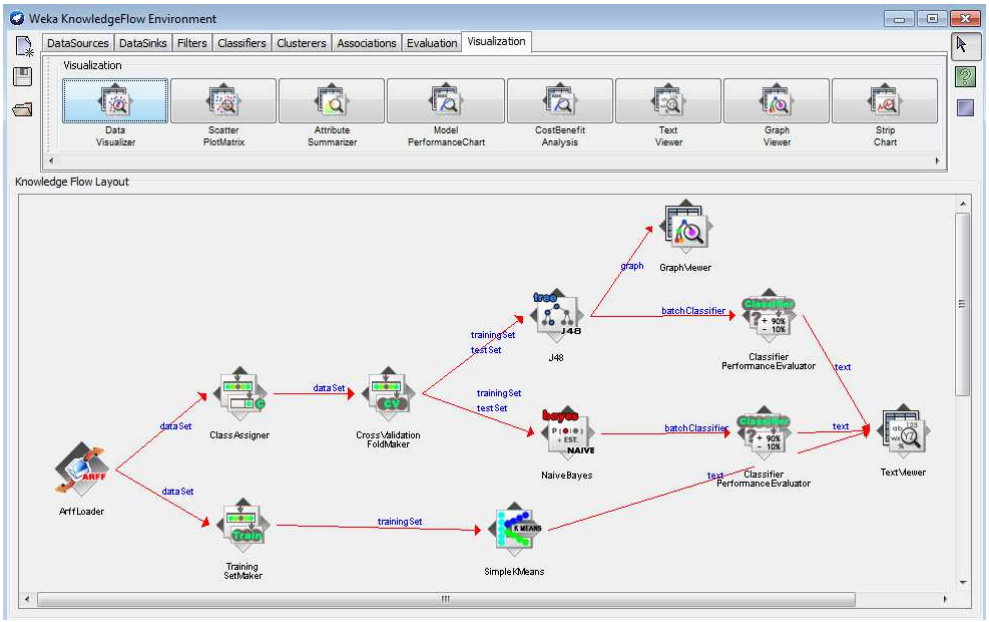
Toliau trumpai apžvelgtos populiariausios duomenų tyrybos sistemos. Įprastai kiekvienoje sistemoje įgyvendinti net keli metodai tam pačiam uždaviniui spręsti. Be to, dauguma sistemų yra pritaikytos ne vienam, o keliems duomenų tyrybos uždaviniams spręsti. Duomenų tyrybos sistemų apžvalga yra pateikta straipsnyje [9], tačiau ten nėra nagrinėjama, ar sistemos yra tinkamos duomenų tyrybai mokyti(-is). Duomenų tyrybos sistemos, taikomos mokymo(-si) procese, gali būti skirstomas į tokias grupes:

1. Matematinės statistikos paketai – IBM SPSS Modeler (Clementine), SAS/STAT, Statistica;
2. Programavimo funkcijas turintys paketai – Matlab, Octave, R;
3. Mokslinių darbų sekų principais pagrįstos sistemos – WEKA, Orange, KNIME, RapidMiner, ClowdFlows.

Pirmosios grupės sistemos – tai galingi daugiafunkciniai statistikos paketai. Jie yra tinkami spręsti sudėtingus uždavinius, reikalaujančius gilių matematinės statistikos žinių. Tai komercinės sistemos, todėl ribotai naudojamos akademiniams tikslams.

Antrosios grupės paketų privalumas yra tai, kad juose galima naudoti ne tik esamas funkcijas, bet ir sukurti naujas, reikalingas sprendžiamam uždaviniui. Tai įgalina kurti sudėtingus duomenų tyrybos projektus. Tačiau trūkumas tas, kad reikia žinoti, kokios funkcijos yra įgyvendintos, be to, reikia mokėti programuoti. Todėl mokymo procese juos ne visada įmanoma taikyti, ypač jei duomenų tyryba dėstoma neturintiems programavimo įgūdžių. Matlab sistema yra mokama, todėl mokymo tikslams nėra plačiai naudojama. Yra nemokamas Matlab analogas Octave, tačiau jame nėra naudotojui patogios grafinės sąsajos, todėl mokymo tikslams jis taip pat nėra labai tinkamas. Pastaruoju metu sukuriama vis daugiau paketo R grafinių naudotojų sąsajų (pavyzdžiui, Rattle), todėl R populiarėja ir plačiai taikomas ir mokymo tikslams.

Trečiosios grupės sistemose įgyvendintas mokslinių darbų sekų (angl. *scientific workflow*) principas. Tokiose sistemose yra komponentės, skirtos tam tikram veiksmui atlikti, pavyzdžiui, duomenų įkėlimui, statistinių charakteristikų apskaičiavimui, klasifikavimui, klasterizavimui, rezultatų įverčiams, rezultatų vizualizavimui. Naudotojas į sistemos darbalaukį turi įkelti norimas komponentes, jas sujungti ir įvykdyti visą sukurtą darbų seką. Tai ypač svarbu mokant žinių radimo duomenų aibėse proceso žingsnių, kuriuose naudojami ir duomenų tyrybos metodai. Besimokantieji gali į sukurtą darbų seką įkelti naują komponentę, panaikinti esamą, eksperimentus atlikti pasirinkus kitą duomenų aibę, tokiu būdu ieškoti paslėptos informacijos analizuojamuose duomenyse.

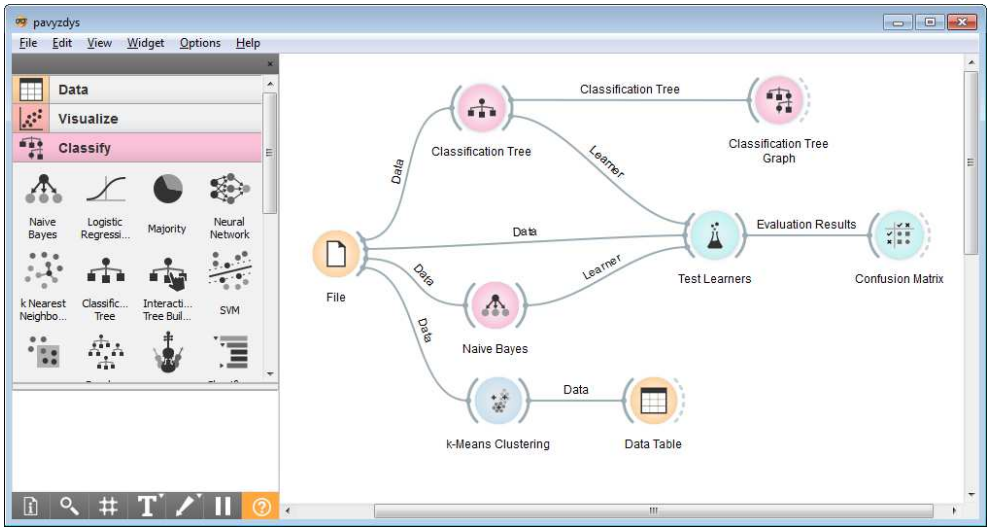


1 pav. Mokslinė darbų seka WEKA sistemoje.

Viena populiariausių mokslinių darbų sekomis grįstų sistemų yra WEKA [4]. Jos grafinės naudotojo sąsajos, kurioje sukurta mokslinė darbų seka, vaizdas pateiktas 1 pav. Šioje sistemoje įgyvendintas didelis kiekis duomenų tyrybos metodų bei jų modifikacijų. Į WEKA sistemą labai panaši Orange sistema [2], kurios funkcionalumas šiek tiek mažesnis, tačiau grafinė naudotojo sąsaja yra draugiškesnė, todėl mokymo tikslams ji gali būti naudojama dažniau nei WEKA sistema. KNIME sistemos [1] galimybės yra platesnės nei anksčiau paminėtų WEKA ir Orange sistemų. Be to, KNIME sistemoje yra integruoti ir WEKA sistemos algoritmai. Todėl naudojant vieną sistemą, galima pasinaudoti abiejų sistemų funkcijomis. Dideles galimybes turi ir RapidMiner sistema [6], tačiau grafinė naudotojo sąsaja gana paini, ir neturintiems įgūdžių yra sudėtinga sukurti net paprasčiausią mokslinių darbų seką. Visos minėtos sistemos turi bendrą trūkumą – jos turi būti instaliuotos į kompiuterį. ClowdFlows sistema [8] šio trūkumo neturi, jos naudojimui pakanka interneto ryšio ir naršyklės, todėl puikiai gali tikti duomenų tyrybos mokymo procese. Naudotojai turi galimybę susikurti savo aplinką, kuri prieinama iš bet kurio interneto ryšį turinčio kompiuterio. Be to, sistemoje integruoti ir WEKA, ir Orange algoritmai. Visos minėtos sistemos yra nemokamos, todėl yra tinkamos mokymui(-si).

3 WEKA ir Orange sistemų palyginimas

Straipsnyje [7] nurodoma, kad duomenų tyrybos mokymui labiausia tinka WEKA ir Orange sistemos, o taip pat Matlab ir R paketai. Šiame skyriuje tiriama, kuri sistema WEKA ar Orange yra tinkamesnė duomenų tyrybos mokymui(-si). ClowdFlows sistema čia nėra nagrinėjama, kadangi ji pastaruoju metu intensyviai tobulinama,



2 pav. Mokslinė darbų seka Orange sistemoje.

1 lentelė. WEKA ir Orange sistemų komponentės.

Funkcija	WEKA	Orange
Duomenų įkėlimas	ArffLoader, ClassAssigner, Training SetMaker	File
Klasifikavimo medis	J48	Classification Tree
Naive Bayes klasifikatorius	NaiveBayes	Naive Bayes
Kryžminė patikra	CrossValidation FoldMaker	Test Learners
Klasifikavimo tikslumo nustatymas	Clasifier PerformanceEvaluator	Test Learners
Klasifikavimo medžio peržiūra	Graph Viewer	Classification Tree Graph
Klasterizavimas	SimpleKMeans	k-Means Clustering
Rezultatų peržiūra	TextViewer	Data Table, Confusion Matrix

ir kartais atlikti pakeitimai sistemą padaro laikinai nestabilia. Tai internetinė sistema, todėl nėra galimybės naudotis jos ankstesne versija, o sistemos nestabilumas gali sutrikdyti mokymo procesą.

Sistemos vertintos pagal mokslinių darbų sekų sukūrimo patogumą. 1 pav. pateikta WEKA, o 2 pav. Orange sistemose sudarytos mokslinių darbų sekos tiems patiems duomenų tyrybos uždaviniams atlikti: įkelta duomenų aibė; atliktas duomenų klasifikavimas taikant medžiais grįstą klasifikavimo metodą; sudarytas medis pavaizduojamas grafiškai; atliktas duomenų klasifikavimas taikant Naive Bayes klasifikatorių; klasifikavimo tikslumas vertinamas atliekant kryžminę patikrą; atliktas duomenų klasterizavimas.

Matome, kad tiems patiems uždaviniams spręsti abejose sistemose yra naudojamos ne vienodas skaičius komponenčių (1 lentelė).

Orange sistemoje pakanka vienos duomenų failo komponentės, o WEKA sistemoje reikia net trijų komponenčių. WEKA sistemoje kiekvienam klasifikatoriui būtina nurodyti klasifikavimo tikslumo nustatymo komponentę, o Orange sistemoje pakanka

vienos komponentės kelių klasifikatorių tikslumui nustatyti, be to, ta pati komponentė yra naudojama ir kryžminei patikrai atlikti. Orange sistemoje rezultatų peržiūros komponentės yra labiau struktūrizuotos. Darbų seka automatiškai įvykdoma, atlikus bet kokius pakeitimus. Tokio funkcionalumo WEKA sistema neturi, norint įvykdyti seką, reikia pasirinkti tam skirtą kontekstinio meniu nuorodą. Orange sistemoje yra patogesnis ir pradedančiajam suprantamesnis komponenčių parametrų nustatymo būdas.

Abejose sistemose yra trumpi paaiškinimai apie kiekvieną komponentę, o Orange sistemoje yra galimybė pasirinkti ir išsamius paaiškinimus. Pradedantieji naudoti mokslines darbų sekas susiduria su komponenčių sujungimo problema, t. y., dažnai nėra aišku, kokios komponentės turi būti sujungtos tarpusavyje. Orange sistemoje prie kiekvienos komponentės yra nurodytos kitos komponentės, kurios gali būti sujungtos. Tai padeda suprasti ką su juo jungti.

4 Išvados

Šiame straipsnyje nagrinėtos duomenų tyrybos sistemos, kurios gali būti taikomos duomenų tyrybos mokymui(-si). Renkantis sistemą svarbu atsižvelgti į sistemos patogumą ir paprastumą. Sistemos platus funkcionalumas neturi trukdyti suprasti žinių radimo duomenų aibėse proceso. Jei sprendžiami uždaviniai betarpiškai susiję su matematine statistika, galima naudoti SPSS, Statistica, Matlab, R paketus, ar jų nemokamus analogus. Tačiau jei norima mokyti, kaip spręsti klasifikavimo, klasterizavimo, susietumo taisyklių sudarymo uždavinius, rekomenduojama naudoti WEKA, Orange, KNIME ar RapidMiner sistemas. Iš pastarųjų tinkamiausiomis savybėmis mokymui(-si) pasižymi Orange sistema.

Literatūra

[1] M.R. Berthold, N. Cebron, F. Dill, T.R. Gabriel, T. Kotter, T. Meinl, P. Ohl, Ch. Sieb, K. Thiel and B. Wiswedel. Knime: the Konstanz information miner. In *Data Analysis, Machine Learning and Applications*, Studies in Classification, Data Analysis, and Knowledge Organization, pp. 319–326. Springer, 2008. ISBN 978-3-540-78239-1.

[2] T. Curk, J. Demšar, Q. Xu, G. Leban, U. Petrovič, I. Bratko, G. Shaulsky and B. Zupan. Microarray data mining with visual programming. *Bioinformatics*, **21**(1):396–398, 2005.

[3] M.H. Dunham. *Data Mining: Introductory and Advanced Topics*. Prentice Hall PTR, 2002. ISBN 0130888923.

[4] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann and I.H. Witten. The WEKA data mining software: an update. *SIGKDD Explorations*, **11**(1):10–18, 2009.

[5] J. Han, M. Kamber and J. Pei. *Data Mining: Concepts and Techniques*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 3rd edition, 2011. ISBN 0123814790, 9780123814791.

[6] M. Hofmann and R. Klinkenberg. *Anomaly Detection*, pp. 367–394. Data Mining and Knowledge Discovery. Chapman and Hall/CRC, 2013. ISBN 978-1-48-220549-7.

[7] B.R. King and A. Satyanarayana. Teaching data mining in the era of big data. In *120th ASEE Annual Conference & Exposition*, 2013.

[8] K. Kranjc, V. Podpecan and N. Lavrac. Clowdflows: a cloud based scientific workflow platform. In *ECML/PKDD(2)*, volume 7524, pp. 816–819. Springer, 2012.

[9] O. Kurasova, V. Marcinkevičius, V. Medvedev ir A. Rapečka. Duomenų tyrybos sistemos, pagrįstos saityno paslaugomis. *Informacijos mokslai*, **65**:66–74, 2013.

SUMMARY

Data mining systems for teaching

O.Kurasova

Data mining systems suitable for data mining teaching have been investigated in the paper. Ussualy, such systems as SPSS Modeler (Clementine), Statistica, SAS/STAT are used in the mathematical statistics courses. However, they are not always suitable for data mining teaching. WEKA, Orange, KNIME, RapidMiner systems are more appropriate for this purpose.

Keywords: data mining, classification, clustering, scientific workflow.