DUOMENŲ TYRYBOS PROGRAMINIŲ ĮRANKIŲ GALIMYBIŲ ANALIZĖ

Danutė Kaklauskienė
Šiaulių valstybinė kolegija
Lietuva

Anotacija
Straipsnyje nagrinėjamos ir lyginamos duomenų tyrybos programinių įrankių galimybės. Duomenų tyrybos programiniai įrankiai pagal savo pasirinktus požymius suskirstyti į 5 grupes. Pirmųjų trijų grupių įrankių galimybės aprašytos, likusių dviejų grupių – universaliųjų matematikos kompiuterinių sistemų – galimybės palygintos pagal šešis įverčius, kuriems priskiriami 48 kriterijai. Šiose grupėse pagal pasirinktuosius kriterijus nustatyti optimalūs duomenų tyrybos įrankiai.
Reikšminiai žodžiai: *duomenų tyryba, duomenų tyrybos programiniai įrankiai, statistinė analizė.*

Įvadas

Temos aktualumas ir naujumas. Žmonija nuo neatmenamų laikų rinko duomenis ir juos analizavo. Kuo toliau, tuo labiau buvo reikalingos greitos ir patikimos išvados. Kasdien sukuriami milžiniški kiekiai duomenų, tačiau tai nėra informacija. Kad būtų išgauta iš informacija duomenų, reikia juos apdoroti.

Duomenų apdorojimo mokslas (angl. *data science*) – tai duomenų analizė panaudojant mokslinius metodus. Didžiuliame kiekyje duomenų gali slėptis ir strategiškai svarbi, ir niekinė informacija. Svarbios informacijos paieška milžiniškuose duomenų kiekiuose ir paskatino atsirasti duomenų analizės priemonėms, aukštos kokybės taikomiesiems paketams, programavimo įrankiams, kurie padeda susiorientuoti informacijos gausoje. Duomenų ir informacijos didėjimas kelia naujus reikalavimus informacijos apdorojimo kompiuterinėms sistemoms.

Duomenų tyryba (angl. *data mining*) – naudingos informacijos ištraukimas iš sukauptų duomenų. Ji įdomi tuo, kad jos technologijos sugeba faktiškus duomenis paversti naudinga informacija ir žiniomis, tinkamomis veiklos valdymui, rinkos analizei, sprendimų priėmimui (Sekliuckis, 2006).

Duomenų tyryba – daugiareikšmė sąvoka: ją galima apibrėžti kaip struktūrų (modelių, ryšių, statistinių modelių, šablonų) radimą duomenų bazėse (Fayyad, Chaudhuri, Bradley, 1993), kaip statistikos pritaikymą tiriamųjų duomenų analizės ir prognozuojamų modelių formai, siekiant atrasti modelius ir kryptingumus (angl. *trends*) dideliuose duomenų rinkiniuose, ir kaip didelių duomenų kiekių tyrinėjimą ir analizę automatizuotu arba pusiau automatizuotu būdu, siekiant rasti naudingus modelius (angl. *patterns*) ir taisykles (Berry, Linoff, 1999).

Duomenų aibėse duomenų tyryba atliekama žinių radimo procese (angl. *knowledge discovery in databases*). Šio proceso metu didelių apimčių duomenų aibėse ieškoma naujos informacijos, kuri padėtų įgyti žinias apie analizuojamus duomenis ir priimti sprendimus (Han, Kamber, Pei, 2011).

Naudojant duomenų tyrybos būdą sprendžiami prognozavimo, klasifikavimo, klasterizavimo, susietumo taisyklių paieškos uždaviniai. Todėl svarbu turėti sistemas, kuriose būtų įvairūs duomenų tyrybos uždavinius sprendžiantys metodai (Dunham, 2002).

Tyrimo tikslas – įvertinti didelės apimties duomenų tyrybos priemones, atlikti duomenų tyrybos programinių įrankių galimybių lyginamąją analizę, išsiaiškinant galimybių skirtumus bei panašumus.

Tyrimo objektas – duomenų tyrybos programinė įranga.

Tyrimo metodai: teorinės mokslinės literatūros apžvalga, statistinė lyginamoji analizė.

Duomenų tyrybos programinių įrankių klasifikacija

Duomenų tyrybos programiniai įrankiai suklasifikuoti pagal šiuos požymius:

- uždavinių sprendimo automatizacijos lygį,
- programų funkcines galimybes,
- programų greitaeigiškumą,
- maksimalius apdorojamos imties tūrius,
- naudotojo kvalifikacijos (statistikos žinių) reikalavimus;
- kainą ir t. t.

Pagrindinis dėmesys skirtas pirmajam požymiui. Jo pagrindu analizuoti atitinkantys duomenų tyrybos programiniai įrankiai.

Programiniai įrankiai sugrupuoti į penkis klasterius:

1. Kitų programų įskiepiai.
2. Duomenų tyrybos programų bibliotekos.
3. Ekspertinės duomenų tyrybos sistemos.
4. Universalios kompiuterinės matematikos sistemos, kuriose yra duomenų analizės ir tyrybos paketai.
5. Kompiuterinės statistikos programos.

Programų įskiepiai

Programų įskiepiai kaip programų papildiniai veikia drauge su kita programa. Literatūroje galima rasti tris pagrindinius įskiepius: XLMiner, TreePlan ir Microsoft SQL Server 2008. Visi trys įskiepiai yra skaičiuoklės Microsoft Excel 997/200/2003/2007 papildiniai, dviejų jų – XLMiner, TreePlan – mokamos licencijos, o įskiepis Microsoft SQL Server 2008 yra nemokamas. Įskiepiai XLMiner, TreePlan turi panašias funkcijas, tačiau jų visų trijų grafinės naudotojo sąsajos skiriasi (tai naudotojui yra nepatogu). Plačiau šiuos įskiepius nagrinėjo Stravinskienė, Žukauskaitė ir Gudas (2010), Preidys, Sakalauskas (2010).

Duomenų tyrybos programų bibliotekos

Duomenų tyrybos programų bibliotekomis gali naudotis vartotojas, turintis atitinkamą kvalifikaciją statistikos, duomenų tyrybos ir programavimo srityse. Programų bibliotekos dažnai parašytos universaliomis programavimo kalbomis (FORTRAN, PL/1,C, PASCAL ir t. t.).

Ekspertinės duomenų tyrybos sistemos

Ekspertinių duomenų analizės sistemų grupei priklauso šios sistemos: TABLECURVE 2D, TABLECURVE 3D, MVSV, ABP.

Šiuose įrankiuose automatizuoti duomenų tyrybos etapai: užduoties formulavimas, tinkamo duomenų analizės metodo parinkimas, duomenų analizė, rezultatų interpretavimas ir išvadų formulavimas. Programa TABLECURVE turi dvi versijas – TABLECURVE 2D v5.01 ir TABLECURVE 3D v4.0.

Sistema TABLECURVE 2D, sukurta 1999 m., pateiktiems duomenims iš 3665 galimų variantų parenka geriausią regresijos lygtį ir apskaičiuoja jos koeficientų įverčius. TABLECURVE 3D sukuria empirinių duomenų teorinius modelius, rezultatus atvaizduoja arba dvimatėmis kreivėmis, arba kaip trimačius paviršius.

Universalios kompiuterinės matematikos sistemos

Universaliosioms kompiuterinėms matematikos sistemoms priklauso šie matematinės statistikos paketai: MATHCAD, MATHEMATICA, MATLAB, MAPLE, MAXIMA (atviro kodo), Scilab (atviro kodo) ir kiti, turintys aprašomosios statistikos, faktorinės analizės, laiko eilučių analizės, daugiamačių metodų, kokybės kontrolės metodų ir kt. funkcijas.

Kompiuterinė matematikos sistema MATHCAD yra specializuota matematiniams, techniniams ir ekonominiams uždaviniams spręsti sistema. Duomenų analizės paketas (angl. *Data Analysis Extension Pack*) leidžia šią sistemą taikyti ir sprendžiant duomenų tyrybos uždavinius. MATHCAD suderinama su *SmartSketch, VisSim/Comm PE, Pro/ENGINEER* programomis, palaiko 9 kalbas (anglų, vokiečių, prancūzų, italų, ispanų, japonų, tradicinė ir supaprastinta kinų, korėjiečių) (Raudys, 2008).

Sistema MATHEMATICA naudojama kaip galinga statistinių uždavinių sprendimo priemonė. MATHEMATICA yra daugiaplatformė programinė įranga, ji veikia *Linux, Apple Macintosh, MS-DOS, NeXT, OS/2, Unix, VMS ir Windows* sistemose. Plačiau šios sistemos galimybės nagrinėtos Varian (1996) darbe, sistemos taikymas duomenų tyrybos srityje apžvelgtas Bouchard (2014) darbe.

Naudojant sistemą MATLAB, galima analizuoti duomenis, kurti algoritmus, modelius ir taikomąsias programas. MATLAB įgalina analizuoti, valdyti, filtruoti ir vizualizuoti duomenis, atlikti tiriamąją duomenų analizę, siekiant atskleisti tendencijas, bandymų prielaidas ir kurti aprašomuosius modelius. MATLAB leidžia pasiekti duomenis iš failų, kitų programų, duomenų bazių ir išorinių įrenginių.

Nemokamu MATLAB analogu gali būti sistemos *Scilab* ir *Octave*. Statistiniams skaičiavimas atlikti naudojamas paketas *Statistics Toolbox*. Čia yra duomenų analizės ir modeliavimo algoritmai ir instrumentai: prognoziniam modeliavimui galima taikyti regresiją ar klasifikaciją, Monte-Karlo metodo modeliavimui taikyti atsitiktinių skaičių generavimą, taip pat patikrinti hipotezes. Paketas leidžia visus duomenis – skaitinius, tekstinius ar metaduomenis –

saugoti viename kataloge. Įterptieji metodai leidžia sujungti duomenų rinkinius, konvertuoti duomenis, kategoriniai masyvai naudoja nedaug kompiuterio atminties (Himberg, Alhoniemi, Parhankangas, 2000; Blasius, Greenacre, 2014).

Sistema MAPLE skirta aukštosios matematikos uždaviniams spręsti bei sudėtingesniems matematiniams tyrimams atlikti. Ši sistema apima įvairius matematikos skyrius, turi patogią vartotojo sąsają, dideles simbolinių skaičiavimų galimybes ir daug paketų, galinčių spręsti atskirus matematikos taikymo uždavinius, pavyzdžiui, tiesinio optimizavimo, statistikos, diferencialinių lygčių ir t. t. Ši sistema turi paketą *stats*, skirtą statistinei analizei atlikti. Šis paketas turi savo popakečius statistiniam uždaviniui išspręsti (Hrebícek, Trenz, Chvátalová, Soukopová, 2015).

Kompiuterinės statistikos programos

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 (Dunham, 2002).

Šios grupės programos turi dviejų lygių programines priemones bendravimui su vartotoju:

- tipinių duomenų analizės uždavinių sprendimo posistemį. Jį dažniausiai sudaro naudojamų duomenų analizės procedūrų rinkinys ir programinės priemonės, sukuriančios programų vartotojui patogią darbo aplinką. Ją gali naudoti nemokantis programuoti vartotojas, bet turintis statistikos žinių;
- specializuotą programavimo kalbą, skirtą duomenų analizės uždavinių programavimui.

Statistikos programų grupei priskiriamos tokios programos: SAS, SPSS, STATISTICA, R, Insightful, KXEN, Excel XL, BMDP, STATGRAPHICS, GENSTAT, S-PLUS, Vortex, SIGAMD, DataScope, STADIA, COMI, COPPA-2, CITO.

Pagal funkcionalumą kompiuterinės statistikos programos dar skirstomos į 3 grupes:

1. **Universalios arba bendrosios paskirties programos** (SPSS, STATA, STATISTICA, S-PLUS, Stadia, STATGRAPHICS, SYSTAT, Minitab). Šios programos nėra skirtos konkrečios srities duomenims analizuoti. Jos turi nemažai statistinių metodų, gana paprasta grafinė naudotojo sąsaja. Šiomis programomis gali dirbti pradedantieji naudotojai ir turintys pagrindinių (teorinių ir praktinių) žinių apie statistines duomenų analizės programas, taip pat patyrę naudotojai.

SPSS (*Statistical Package for the Social Sciences*) paketas – universali programa. Tai modulinė programa, turinti bazinį modulį SPSS Base, kuris įgalina valdyti duomenis ir taikyti duomenų statistinės analizės metodus. Detalesniems duomenų tyrimams galima taikyti ir kitus papildomus modulius. IBM SPSS Statistics 19 turi 16 papildomų modulių.

Sistema R dažnai traktuojama kaip alternatyva statistinei programai SPSS. Ji suderinama su operacinėmis sistemomis *Linux* ir *Windows*, turi duomenų keitimosi su elektroninėmis lentelėmis galimybę, statistinių procedūrų programavimo nuosavą kalbą R, kuri yra faktiškai standartinė. R – tai objektiškai orientuota atviro kodo programinė įranga, skirta statistinei analizei finansų sektoriuje. Tačiau jos galimybėmis gali pasinaudoti ir kitų sričių atstovai.

Programa STATA – universali statistinė kompanijos StataCorp programa, turinti modulinę struktūrą. Galima taikyti plačiai: visuomenės mokslų (ekonomika, politologija ir t. t.), medicinos mokslų (biostatistika, epidemiologija ir t. t.) ir pan. duomenų analizei.

Programa Minitab yra universali programa, sukurta 1972 metais Pensinvalijos valstybiniame universitete. Paskutinė versija Minitab 16 išleista 7 kalbomis (anglų, prancūzų, vokiečių, japonų).

Standartinę STATISTICA programą sudaro 3 moduliai, kurie gali būti taikomi ir atskirai:

- Bazinis paketas STATISTICA Base. Jame yra visų statistinės analizės tipų galimybės.
- Tiesinio ir netiesinio modelio modulis (*Advanced Linear/NonLinearModels*) turi modeliavimo ir prognozavimo instrumentų rinkinius, tarp jų ir automatinį modelio parinkimą bei interaktyvias vizualizacijos galimybes.
- Analizės daugiamačių technologijų modulis (*Multivariate Exploratory Techniques*) leidžia taikyti skirtingų tipų duomenų tiriamąją analizę kartu su interaktyvia vizualizacija.

Programa taip pat turi specializuotus modulius, pavyzdžiui, techniniams ir pramoniniams uždaviniams spręsti: kokybės kontrolės žemėlapiai, procesų analizės modulis, eksperimento planavimo modelis ir pan.

Programa STATGRAPHICS (*STATistical GRAPHICS System*) – universalus statistinis paketas, suderinamas su kita programine įranga (skaičiuokle, duomenų bazėmis). Jame integruota grafika.

Programos SAS (*Statistical Analysis System*) paskutinė versija SAS Statistics jau traktuojama kaip statistinis paketas, skirtas šiuolaikinėms įmonėms. Tai daugiau vieno tipo statistinis paketas – marketingo duomenų analizei, klinikiniams tyrimams, medicininių-sanitarinių tyrimų analizei ir pan.

2. **Profesionalios statistinės programos** (SAS, BMDP). Profesionalios statistinės programos skiriasi nuo universaliųjų tuo, kad galima dirbti su itin didelės apimties duomenimis, naudoti labai specializuotus analizės metodus, sukurti savo duomenų apdorojimo sistemą.

3. **Specializuotos kompiuterinės statistinės programos** (BioStat, MESOSAUR, DATASCOPE).

Specializuotos programos leidžia atlikti duomenų analizę su ribotu statistinių metodų, skirtų tam tikrai sričiai, skaičiumi. Pavyzdžiui, programa BioStat sukurta duomenų analizei biomedicinoje ir medicinoje. Rusų sukurta programa MESOSAUR skirta vienmačių ir daugiamačių laiko eilučių ir regresinių modelių analizei. Rusų kalba programa DATASCOPE taikoma daugiamačių duomenų analizei.

Programa BioStat skirta medicinos, biologijos ir chemijos duomenų statistinei analizei. Programa naudoja elektroninę lentelę, suderintą su Microsoft Excel, kaip duomenų formatą ir kaip sąsają (Yu Hong, Zhanguo Liu, Guoyin Wang, 2014; Michelland, Combes, Cauquil, 2014).

Duomenų tyrybos įrankių vertinimo kriterijai

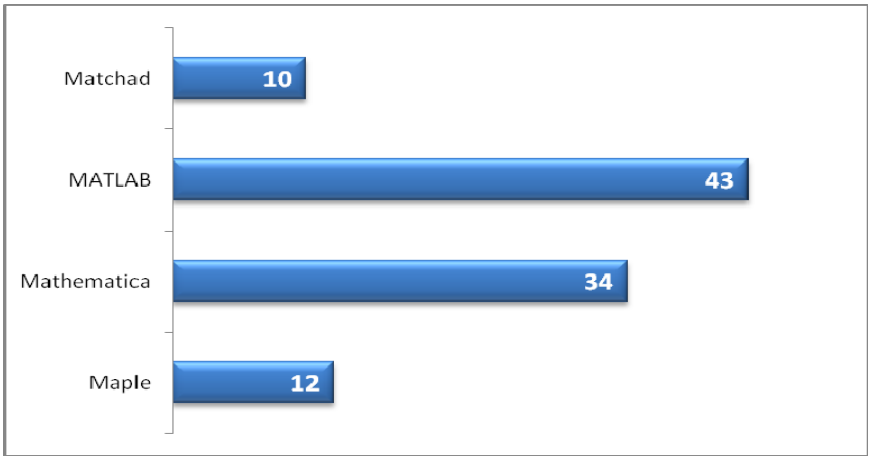
Detalesniam pasirinktos programinės įrangos įvertinimui taikytas literatūros analizės metodas. Pasirinktų programų įvertinimui naudoti tokie šeši įverčiai:

- sistemos ir operacinės sistemos (OS) suderinamumas,
- regresijos metodų panaudojimo galimybės,
- laiko eilučių metodų panaudojimo galimybės,
- Anova metodų panaudojimo galimybės,
- vizualizacijos galimybės.

Kiekvienas įvertis apima nuo šešių (vizualizacijos įvertis) iki trylikos (regresiniai metodai) kriterijų. Tokiu būdu viso vertinti 48 kriterijai. Įvertinimai koduoti taip: jei programa vykdo įvertinamą parametrą, koduojama 1, jei nevykdo – 0.

Duomenų tyrybos įrankių vertinimo rezultatai

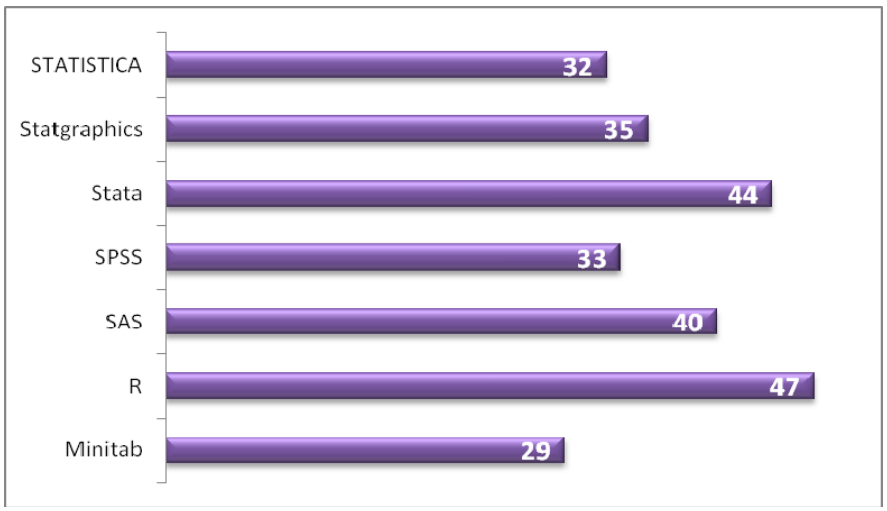
Galimybių vertinimui pagal pasirinktus kriterijus atrinktos 4 populiarios universalios kompiuterinės matematikos sistemos: MATHCAD, MAPLE, MATHEMATICA ir MATLAB (18). Vertinti 48 kriterijai, priklausantys anksčiau minėtiems įverčiams. Rezultatai pateikiami diagramoje (1 pav.).



1 pav. Universalių kompiuterinių matematikos sistemų įverčių diagrama

Iš 1 paveikslo matyti, kad daugiausia įverčių tenkina sistema MATLAB (43 iš 48), mažiausia – sistema MATHCAD (tik 10 iš 48). Vadinasi, žinių gavimui iš duomenų labiausiai tinka matematinė sistema MATLAB.

Palyginti statistikos programas pagal anksčiau apibrėžtus kriterijus gaunama, kad iš nagrinėtų kompiuterinės statistikos programų daugiausia parametrų tenkina sistema R (2 pav.).



2 pav. Kompiuterinių statistikos programų įverčių diagrama

Išvados

1. Taikant literatūros analizės metodą atrinkti duomenų tyrybos programiniai įrankiai suskirstyti į penkis klasterius: kitų programų įskiepiai, duomenų tyrybos programų bibliotekos, ekspertinės duomenų tyrybos sistemos, universalios kompiuterinės matematikos sistemos ir kompiuterinės statistikos programos.
2. Detalesnei pasirinktųjų sistemų analizei suformuoti 48 vertinimo kriterijai, suskirstyti į 6 grupes, kurios pavadintos įverčiais.
3. Remiantis lyginamosios statistikos rezultatais geriausios duomenų tyrybos priemonės yra šios: universali kompiuterinė matematikos sistema MATLAB ir kompiuterinė statistikos programa R.

ANALYSIS OF OPPORTUNITIES DATA MINING SOFTWARE TOOLS

The article examines and compares the options of data mining software tools. Data mining software tools according to their custom attributes are divided into 5 groups. The first three groups of tools options are described, the remaining two groups – the universal mathematics of comput

er systems the possibilities compared according the six estimates consisting of 48 criteria. In these groups according to the criteria set elect optimal data mining tools.

Key words: *data mining, data mining software tools, statistical analysis.*

Literatūra

1. Berry M. J. A., Linoff G. S. (1999). *Mastering Data Mining: The Art and Science* (text only).
2. Blasius J., Greenacre M. (Eds.) (2014).*Visualization and verbalization of data.* CRC Press.
3. Bouchard K. et al. (2014). *Analysis and Knowledge Discovery of Moving Objects Equipped with RFID Tags.* Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence.
4. Dunham M. H. (2002). *Data Mining: Introductory and Advanced Topics.* Prentice Hall PTR.
5. Fayyad U., Chaudhuri S., Bradley P. (1993). *Data mining andits role in database systems,* vol. 5, no. 6, 914–925.
6. Han J., Kamber M., Pei J. (2011). *Data Mining: Concepts and Techniques.* 3rd edition. Morgan Kauf-mann Publishers Inc., San Francisco, USA.
7. Himberg J., Alhoniemi E., Parhankangas J. (2000). *SOM toolbox for Matlab 5.* Helsinki: Helsinki University of Technology.
8. Hong Yu, Liu Zhanguo, Guoyin Wang. (2014). *An automatic method to determine the number of clusters using decision-theoretic rough set.* International Journal of Approximate Reasoning, 55 (1), 101–115.
9. Hrebícek J., Trenz O., Chvátalová Z., Soukopová J. (2015). *Optimization of corporate performance using data envelopment analysis with Maple.* Engineering optimization, IV. London: Taylor & Francis Group.
10. Michelland R. J., Combes S., Cauquil L. (2014). *OligoSpecificitySystem: global matching efficiency calculation of oligonucleotide sets taking into account degeneracy and mismatch possibilities.* International Journal of Data Mining and Bioinformatics, 9 (4), 417–423.

11. Preidys S., Sakalauskas L. (2010). *Analysis of Students' Study Activities in Virtual Learning Environments Using Data Mining Methods.* Technological and Economic Development of Economy, 16 (1). Vilnius: Technika.

12. Raudys Š. (2008). *Žinių išgavimas iš duomenų:* vadovėlis. Klaipėda: Klaipėdos universitetas.

13. Sekliuckis V., Gudas S., Garšva G. (2006). *Informacijos sistemos ir duomenų bazės: informacijos sistemų ir reliacinių duomenų bazių kūrimo pagrindai:* vadovėlis. Kaunas: Technologija.

14. Stravinskienė A., Žukauskaitė A., Gudas S. (2010). *Duomenų gavybos įrankių pritaikymas mažose įmonėse.* Information Technologies, 16th International Conference on Information and Software Technologies Kauno technologijos universitetas, pp. 19–25.

15. Varian Hal R. (Eds.) (1996). *Computational economics and finance: modeling and analysis with Mathematica.* Vol. 2. Springer Science & Business Media.

Įteikta: 2015 m. kovo 9 d.
Priimta publikuoti: 2015 m. gegužės 25 d.