Lietuvos matematikos rinkinys
Lietuvos matematikų draugijos darbai, ser. B
56 t., 2015, 41–45

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

kryptys*

Dovilė Stumbrienė, Audronė Jakaitienė

Vilniaus universitetas, Matematikos ir informatikos institutas
Akademijos 4, LT-08663 Vilnius
E. paštas: dovile.stumbriene@mii.vu.lt, audrone.jakaitiene@mii.vu.lt

Santrauka. Straipsnyje pateikiama sisteminė literatūros apžvalga, kurios tikslas dažniausiai naudojami duomenų šaltiniai ir taikomi duomenų analizės metodai švietimo duomenų tyryboje. Apžvalgai pasirinkta tarptautinė duomenų bazė „Web of science“. Atmetus trumpus konferencijų pranešimus bei straipsnius, kuriuose nepateikti empiriniai duomenys, išsamiai nagrinėjami 14 straipsnių. Nustatyta, kad dažniausiai tiriamos besimokančiųjų duomenų bazės, kuriuose su mokomųjų dalykų įvertinimais pateikiama ir kontekstinė informacija. Švietimo duomenų tyryboje dažniausiai taikomi klasifikavimo metodai ir kiek rečiau regresijos analizė bei klasterizavimas. Straipsnis baigiamas švietimo duomenų tyrybos apžvalga Lietuvoje.

Raktiniai žodžiai: duomenų tyryba, švietimas, klasterizavimas, klasifikavimas, regresija.

Įvadas

Švietimo duomenų tyrybą galima apibrėžti kaip informatikos, statistikos ir edukologijos mokslų persidengimą, kur atsiranda dar trys mokslinių tyrimų sritys artimos švietimo duomenų tyrybai: kompiuteriu grįstas mokymas (angl. computer-based education), duomenų tyryba ir mašininis mokymasis (*angl. machine learning*) bei mokymosi analitika (*angl. learning analytics*).

Švietimo duomenų tyrybos srityje analizuojamas ir modeliuojamas besimokančiųjų elgesys, prognozuojami tikėtini mokymosi rezultatai ir galimybė iškristi iš mokymosi proceso, nagrinėjama, kokios mokymo strategijos lemia geresnius rezultatus, tiriama šeimos ir mokymo įstaigos įtaka rezultatams, analizuojamas švietimo sistemos poveikis stratifikacijos procesui ir socialiniam mobilumui.

Siekiant nustatyti, kokie duomenų šaltiniai ir tyrimų metodai naudojami švietimo duomenų tyryboje, buvo atlikta sisteminė literatūros analizė. Literatūros šaltinių paieška atlikta 2015 m. kovo mėn. tarptautinėje duomenų bazėje „Web of science“. Naudota angliška paieškos frazė – „educational data mining“, pasirinkta anglų kalba ir laikotarpis 2010–2014 m. Iš viso buvo rasta 319 literatūros šaltinių, kuriuose nagrinėjami švietimo duomenys, naudojant duomenų tyrybos metodus. Atmetus trumpus konferencijų pranešimus bei straipsnius, kuriuose nepateikti empiriniai duomenys, išsamiai buvo nagrinėjami 14 straipsnių. Straipsnis pradėtas įvadu, kurį tęsia duomenų šaltinių ir tyrybos metodų skyriai. Straipsnį baigia apibendrinimas ir diskusija.

* Tyrimas yra remiamas Lietuvos mokslo tarybos projekto Nr. MIP-024/2015 lėšomis.

1 lentelė. Sisteminė literatūros analizė.

Straipsnis	Metodai				Duomenų šaltiniai		
	1	2	3	4	A	B	C
Narli ir Ozelik, 2010 [11]	X					X	
Dogan ir Camurcu, 2010 [5]			X		X		
Costantini ir kt., 2010 [4]				X			X
Guo, 2010 [7]	X	X			X		
Thai-Nghe ir kt., 2011 [17]	X						X
Kotsiantis, 2012 [8]	X	X			X		
Lahane ir kt., 2012 [9]			X				X
Marquez-Vera ir kt., 2013 [10]	X				X		
Chau ir Phung, 2013 [3]	X	X				X	
Ocumpaugh ir kt., 2014 [12]	X				X		
Reimann ir kt., 2014 [13]				X	X		
Sahebi ir kt., 2014 [15]	X	X				X	
Goes ir kt, 2014 [6]	X					X	
Saarela ir Kärkkäinen, 2014 [14]			X	X	X		

Metodai: 1 – klasifikavimas, 2 – regresija, 3 – klasterizavimas, 4 – kita. Duomenų šaltiniai: A – besimokančiųjų duomenų bazė su įvertinimais ir kontekstine informacija, B – besimokančiųjų duomenų bazė tik su įvertinimais, C – kita.

1 Švietimo duomenų šaltiniai

Švietimo duomenų tyryboje naudojami tarptautinių švietimo sistemos tyrimų duomenys, nacionalinės mokinių duomenų bazės, universitetų ir mokyklų turimos duomenų bazės, nuotolinių kursų ir interaktyvios mokymosi aplinkos generuojamų duomenų bazės. Visuose šiose duomenų bazėse yra besimokančiųjų įvertinimo duomenys, tačiau tik dalyje iš jų yra kontekstinė informacija apie besimokantįjį (demografinės charakteristikos, atsakymai į apklausos klausimus, veiksmai, daryti mokymosi metu ir pan.). Esant šiai informacijai galima atlikti sudėtingesnius tyrimus, nei nagrinėjant duomenų bazes, kuriose yra tik mokomųjų dalykų įvertinimai.

Pusėje nagrinėtų straipsnių (detaliau 1 lentelė) yra naudojamos besimokančiųjų duomenų bazės, kuriuose yra kontekstinė informacija. Dalyje straipsnių (28,6 proc.) nagrinėjami besimokančiųjų įvertinimai be kontekstinės informacijos. Likusiuose straipsniuose analizuojami specializuoti duomenų rinkiniai, kurie naudojami naujai pasiūlytų metodų efektyvumui įvertinti, taip pat analizuojami studentų nuomonės tyrimo duomenys.

2 Duomenų tyrybos metodai švietime

Straipsniuose taikyti duomenų tyrybos metodai buvo suskirstyti pagal savo pobūdį į tris grupes: klasifikavimo, regresinės analizės ir klasterizavimo metodus (detaliau 2 lentelė). Nustatyta, kad švietimo duomenų tyryboje dažniausiai taikomas – klasifikavimas (47,4 proc.), regresija (21,1 proc.) ir klasterizavimas (15,8 proc.). Taip pat taikomas sąryšio taisyklių (angl. association rule) tyrybos metodas, sekos ir proceso (angl. sequential pattern and process) tyrybos metodas [13] bei pagrindinių komponenčių metodas [4], leidžiantis daugiamačius duomenis atvaizduoti mažesnio skaičiaus matmenų erdvėje.

Klasifikavimo metodai. Švietimo duomenys nagrinėtuose straipsniuose dažniausiai klasifikuojami, taikant dirbtinius neuroninius tinklus, genetinius algoritmus

2 lentelė. Švietimo duomenų tyryboje taikomi metodai ir algoritmai.

Metodai	Taikomi algoritmai (angl.)
Klasifikavimas (1)	Rough set data analysis, Neural Network, M5 Model trees, Artificial Neural Networks, Three-mode tensor factorization, Genetic Algorithms, Support Vector Machines, Bayesian Knowledge Tracing, Bayesian Probabilistic Tensor Factorization, Bayesian Probabilistic Matrix Factorization, JRip, NNge, OneR, Prism, Ridor, J48, C4.5, SimpleCart, ADTree, RandomTree, REPTree, Na ıve Bayes, Support Vector Machine, Neural Network, K-nn, C4.5, Bagging with SVM, Boosting with SVM, Random Forest, K∗
Regresija (2)	Linear Regression, Logistic Regression, Multilayer Regression Analysis, Performance Factor Analysis
Klasterizavimas (3)	Two Phase Clustering, Cluster analysis using robust prototypes, k-means, fuzzy c-means
Kita (4)	Nonlinear principal component analysis with optimal scaling, categorical principal component analysis, Sequential pattern mining and process mining, Association rule mining

ir atraminių vektorių klasifikatorių (angl. support vector machines): Guo [7] analizuoja studentų nuomonės tyrimo duomenis ir pasiūlo modelį dėstomo kurso klausytojų pasitenkinimo lygio prognozavimui, Marquez-Vera ir kt. [10] nagrinėja mokinių tikimybę iškristi iš mokymosi proceso ir sudaro modelį mokymosi sėkmei prognozuoti, Goes ir kt. [6] sukonstruoja mokyklų kokybės vertinimo modelį, o Chau ir Phung [3] bei Kotsiantis [8] prognozuoja tikėtinus mokymosi rezultatus. Švietimo duomenų tyryboje taip pat taikoma faktorinė analizė [17, 15] ir grubių aibių (angl. rought set) teorija [11]. Klasikiniai klasifikavimo algoritmai (J48, REPTree, JRip ir kt.) švietimo duomenų tyryboje taikomi retai [12], dažniau naudojami palyginimui su kitais algoritmais [10, 3].

Regresinė analizė. Švietimo duomenims analizuoti taikoma tiesinė, logistinė regresijos. Visais atvejais regresinės analizės algoritmai buvo lyginami su klasifikavimo algoritmais ir tikrinamas jų tinkamumas mokymosi rezultatų prognozavimui [8, 3, 15] bei pasirenkamo studijų dalyko pasitenkinimo prognozavimui [7].

Klasterizavimo metodai. Švietimo duomenų tyryboje klasterizavimas taikomas tiriamų objektų grupavimui į homogenines grupes. Du nehierarchinius klasterinės analizės algoritmus (k – vidurkių ir fuzzy c – vidurkių) taiko Dogan ir Camurcu [5] mokinių suskirstymui į grupes, remiantis egzaminų rezultatais. Darbe [14] taip pat taikomi nehierarchiniai klasterizavimo algoritmai mokinių grupavimui, remiantis tarptautinio švietimo sistemos tyrimo duomenis. Tuo tarpu Lahane ir kt. [9] pristato naują dviejų fazių klasterizavimo algoritmą (naudojama klasterių skaidymo strategija) švietimo duomenims klasterizuoti.

3 Švietimo duomenų ištekliai ir jų analizė Lietuvoje

Lietuvoje yra sukauptos nacionalinės ir tarptautinės besimokančiųjų duomenų bazės kartu su kontekstine informacija: nacionalinis mokinių pasiekimų matematikos ir lietuvių kalbos srityse tyrimas 4, 6, 8 ir 10 klasės mokiniams, tarptautinis penkiolikmečių tyrimas PISA, tarptautinis matematikos ir gamtos mokslų gebėjimų tyrimas 4 ir 8 klasės mokiniams TIMSS, tarptautinis 4 klasės mokinių skaitymo gebėjimų tyrimas PIRLS. Taip pat yra nacionalinės besimokančiųjų duomenų bazės be kon-

tekstinės informacijos (brandos egzaminų rezultatai, matematikos ir lietuvių kalbos pagrindinio ugdymo pasiekimų programos duomenys) bei bendrieji Lietuvos švietimo sistemos statistiniai duomenys.

Lietuvoje nėra kiekybinių mokslinių tyrimų, kuriuose švietimo sistema būtų tyrinėjama kaip visuma, ir būtų taikomi duomenų tyrybos metodai. Brandos egzaminų rezultatai nagrinėjami skirtingų mokomųjų disciplinų kontekste. Salienė [16] atliko lietuvių kalbos žinių ir gebėjimų analizę, remdamasi brandos egzaminų rezultatais. Buvo skaičiuojamas Pirsono korealiacijos koeficienas, tiriant santykį tarp žinių, įgūdžių ir gebėjimų. Blonskis ir kt. [2] nagrinėjo informacinių technologijų valstybinio brandos egzamino struktūrą, praktinių užduočių ypatumus ir mokinių rezultatus.

Tarptautinių tyrimų duomenis Lietuvos mokslininkai analizuoja naudodamiesi antriniais duomenimis, duomenų analizei netaiko statistinių metodų. Želvys [18], naudodamas PISA 2012 antrinius duomenis, nagrinėja mokyklų ir jų vadovų savarankiškumo sąsają su ugdymo kokybe. Ališauskas [1], analizuodamas švietimo sistemos raidą, naudoja tarptautinio tyrimo TIMSS antrinius duomenis, nacionalinius ir ES šalių statistinius švietimo sistemos duomenis, atlieka duomenų lyginamąją analizę ir taiko aprašomosios statistikos metodus.

4 Apibendrinimas ir diskusija

Atlikus sisteminę literatūros apžvalgą, nustatyta, kad švietimo duomenų tyrybos srityje dažniausiai yra tiriamos besimokančiųjų duomenų bazės, kuriuose yra ne tik mokomųjų dalykų įvertinimai, bet ir kontekstinė informacija. Duomenų analizei dažniausiai taikomi įvairūs klasifikavimo metodai, kurie dažnai derinami su regresijos modeliais. Tiriamųjų grupavimui į homogenines grupes taikyti nehierarchiniai klasterinės analizės metodai. Apžvelgus Lietuvoje sukauptas švietimo duomenų bazes ir atliktus tyrimus, galima teigti, kad Lietuvoje švietimo duomenų tyryba – dar neatrasta mokslinių tyrinėjimų sritis.

Literatūra

[1] R. Ališauskas. Lietuvos švietimas: sėkmės įrodymų paieškos. *Pedagogika*, (95):13–23, 2009.

[2] J. Blonskis, R. Burbaitė ir V. Dagienė. Informacinių technologijų valstybinio brandos egzamino praktinių užduočių ypatumai. *Informacijos mokslai*, (50):136–141, 2009.

[3] V.T.N. Chau and N.H. Phung. Imbalanced educational data classification: an effective approach with resampling and random forest. In *Computing and Communication Technologies, Research, Innovation, and Vision for the Future (RIVF), 2013 IEEE RIVF International Conference*, pp. 135–140. IEEE, 2013.

[4] P. Costantini, M. Linting and G.C. Porzio. Mining performance data through nonlinear PCA with optimal scaling. *Appl. Stoch. Models Bus. Ind.*, **26**(1):85–101, 2010.

[5] B. Dogan and A. Y. Camurcu. Visual clustering of multidimensional educational data from an intelligent tutoring system. *Comput. Appl. Eng. Educ.*, **18**(2):375–382, 2010.

[6] A.R.T. Góes, M.T. Arns Steiner and P.J. Steiner Neto. Education quality measured by the classification of school performance using quality labels. *Appl. Mech. Mater.*, **670**:1675–1683, 2014.

[7] W.W. Guo. Incorporating statistical and neural network approaches for student course satisfaction analysis and prediction. *Expert Syst. Appl.*, **37**(4):3358–3365, 2010.

[8] S.B. Kotsiantis. Use of machine learning techniques for educational proposes: a decision support system for forecasting students grades. *Artif. Intell. Rev.*, **37**(4):331–344, 2012.

[9] S.V. Lahane, M. Kharat and P.S. Halgaonkar. Divisive approach of clustering for educational data. In *Emerging Trends in Engineering and Technology (ICETET), 2012 Fifth International Conference*, pp. 191–195. IEEE, 2012.

[10] C. Márquez-Vera, A. Cano, C. Romero and S. Ventura. Predicting student failure at school using genetic programming and different data mining approaches with high dimensional and imbalanced data. *Appl. Intell.*, **38**(3):315–330, 2013.

[11] S. Narli and Z.A. Ozelik. Data mining in topology education: Rough set data analysis. *Int. J. Phys. Sci.*, **5**(9):1428–1437, 2010.

[12] J. Ocumpaugh, R. Baker, S. Gowda, N. Heffernan and C. Heffernan. Population validity for Educational Data Mining models: a case study in affect detection. *British J. Educat. Techn.*, **45**(3):487–501, 2014.

[13] P. Reimann, L. Markauskaite and M. Bannert. e-Research and learning theory: what do sequence and process mining methods contribute? *British J. Educat. Techn.*, **45**(3):528–540, 2014.

[14] M. Saarela and T. Kärkkäinen. Discovering gender-specific knowledge from Finnish basic education using PISA scale indices. In *Educational Data Mining 2014*, 2014.

[15] S. Sahebi, Y. Huang and P. Brusilovsky. Predicting student performance in solving parameterized exercises. In *Intelligent Tutoring Systems*, pp. 496–503. Springer, 2014.

[16] V. Salienė. Dvyliktos klasės mokinių lietuvių kalbos žinių ir gebėjimų analizė. *Pedagogika*, (76):100–106, 2005.

[17] N. Thai-Nghe, T. Horváth and L. Schmidt-Thieme. Factorization models for forecasting student performance. In *EDM*, pp. 11–20, 2011.

[18] R. Želvys. Mokyklų savarankiškumas ir ugdymo kokybė Lietuvoje. *Pedagogika*, **114**(2):54–63, 2014.

SUMMARY

Educational data mining: overview and research trends

D. Stumbrienė, A. Jakaitienė

The article presents a systematic literature review about the most commonly used data sources and data mining methods in education. International database Web of Science was selected. Excluding short conference proceedings and articles without empirical data, 14 papers were analyzed in detail. It was obtained that the most explored databases of learners consisted of subjects evaluation results together with contextual information. Classification methods were the most commonly used; to a lesser extent regression analysis and clustering. An educational data mining research overview in Lithuania ends the article.

Keywords: data mining, education, clustering, classification, regression.