http://aptitude.w3c.fmi.uni-sofia.bg/

Н А У Ч Е Н О Т Ч Е Т

По проект APTITUDE

Иновативна софтуерна платформа за анализи на големи масиви от учебни и игрови данни за ориентирана към потребителя адаптация на технологично подпомогнато обучение

Фонд „Научни изследвания“, МОН,

Конкурс за финансиране на фундаментални научни изследвания по обществени предизвикателства – 2018 г.

Номер на договор: КП-06-ОПР03/1 от 13.12.2018г.

D3.2. Процедури за анализ на данни от обучението и от игрите, събрани от LMS и образователни игри

ФОНД НАУЧНИ ИЗСЛЕДВАНИЯ
МИНИСТЕРСТВО НА ОБРАЗОВАНИЕТО И НАУКАТА

I. Цел на документа

Основната цел на документа е да дефинират основните процедури при анализ на данни от обучението и образователни игрите (LA - Learning Analytics). Тези процедури определят наличието на дадено множество елементи, наречени примери, които притежават определени качества (характеристики), и целят да се предвиди с достатъчна точност поведението или класа на новопостъпил елемент на базата на неговите собствени характеристики. За тази цел се използва знание (модел), получено при анализа на информацията от съществуващите примери. Документът определя даденото множество данни за даден обучаем извлечени от различни LMS системи и образователни игри и изследва пет алгоритъма за анализ на данни.

2. Въведение

Анализът на данни е тясно свързано с извличането на знания, което е съвкупност от парадигми и методологии, които съчетават построяване на решаващи дървета, намиране на правила, класификация, клъстеризация, невронни мрежи, обучение с примери, логическо програмиране, статистически методи и др. [1].

Анализът на данни получени от LMS системи (LA) е изследователска област, която поставя ново предизвикателство, за да се анализира голямо количество данни от обучението, произведени от студентите, колкото е възможно по-ефективно и смислено. През последните години в много изследователски работи се обсъждат LA и се предлагат различни подходи в учебните процеси. Те имат силен потенциал за да се осигурят по-добри прогнози и начини за интервенция за ученици, изложени на риск от провал както в краткосрочен, така и в дългосрочен план [1].

Някои автори разглеждат връзката между LA и изследователските области на усъвършенстваното технологично обучение, като проучване на действията, извличане на образователни данни, системи за препоръки и персонализирано адаптивно обучение. Изследванията, свързана с тази област, посочват, че централизираните уеб базирани системи за обучение представляват най-широко използвания източник на данни за LA, а най-често използваните LA техники са класификация и прогнозиране [2].

Една област на въздействие на LA е използването в системите за препоръки. Knight et al. дават препоръки за прилагане на LA за въздействие върху обучението чрез разширяване на съществуващата практика, чрез централността на задачите, подходи за съвместно проектиране и внимание както към социалната, така и към техническата инфраструктура [4].

С навлизането на ИКТ във всяка една област на нашето ежедневие в повечето образователни организации се внедряват различни системи за управление на обучението (Learning Management System - LMS). LMS създават и генерират данни от различни дейности и ресурси в процеса на обучението, като логове на дейностите на обучаемите, на дейностите на курсовете, съдържание на курса или резултати от оценките на студентите. Синергията между Big Data, LA и управлението на знанието играе нарастваща роля за предоставяне на

адаптивно и персонализирано обучение, извличане на данни от образователни дейности, визуализация на данни и дава по-добра информираност на служителите и преподаватели във висшето образование [5].

Документът описва създаването на модел на най-използваните дейности от страна на обучаемите и от страна на преподавателя. за да помогнем да се изгради адаптация и препоръки за съдържание на курса или поток от дейностите в курса. За построяване на модела за анализ на данни се използват различни алгоритми за машинно обучение. Някои предишни изследователски работи изследват софтуерна архитектура за производство и предоставяне на учебни ресурси с аудио елементи в университетските курсове по програмиране [19]. Друго проучване създава уеб услугата API на Moodle, която обхваща извличането на подробности и ресурси за курса, за да се доставят всички налични ресурси на даден курс, запазвайки вътрешната структура на организацията на курса. Предлаганото решение е подходящо за търсене в реално време на ресурси и learning analytics. Всички те работят с данни от различен източник на информация [20].

3. Процедури за анализ на данни от LMS и образователни игри

Процедурите за анализ на данни от LMS и образователни игри ще използват различни алгоритми за машинно обучение за LA. За да утвърдим тези процедури, ние избираме и използваме пет алгоритми за машинно обучение за класификация.

Случайната гора (random forest) се състои от голям брой индивидуални дървета за решения и е „комбинация от предсказващи дървета, така че всяко дърво зависи от стойностите на случаен вектор, изваден независимо и със същото разпределение за всички дървета в гората“ [26]. Класификаторът Naïve Bayes е прост класификатор на вероятностите, който се основава на теоремата на Bayes за силна (естествена) независимост между отделните характеристики на обектите и се използва в теорията на вероятностите за изчисляване на вероятността от настъпване на събитие, след като част от информацията за това събитие се знае. Класификацията се извършва като резултат, класът с най-голяма вероятност при наличието на определени характеристики се приема [27].

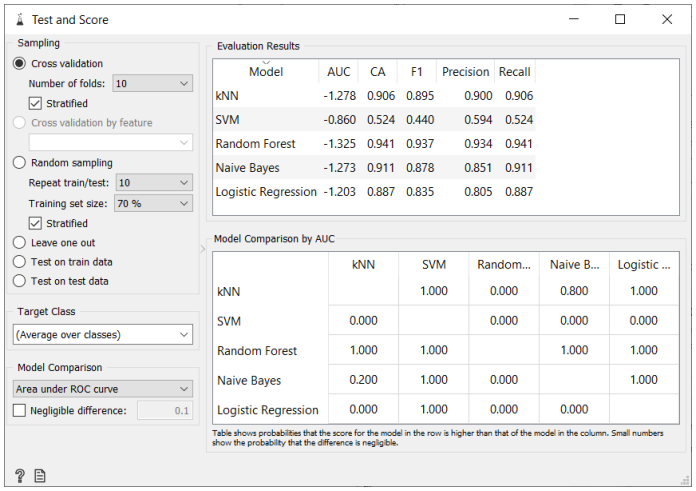
Третият алгоритъм за машинно обучение е k-най-близките съседи (KNN) и е избран, като той „намира група от k обекти в учебния набор, които са най-близо до тестовия обект, и определя възлагането на етикет върху преобладаването на конкретен клас в този квартал ”[28]. Логистичната регресия се използва за възлагане на наблюдения на дискретен набор от класове и се основава на концепцията за вероятността. Хипотезата на логистичната регресия има тенденцията да ограничава функцията на разходите между 0 и 1.

Последният алгоритъм за машинно обучение е поддържащи векторни машини (support vector machines - SVM), които изискват само дузина примери за обучение и са нечувствителни към броя измерения. В допълнение, ефективни методи за обучение. SVM намират най-добрата функция за класификация, за да разграничават членовете на двата класа в данните за обучението. Метриката за концепцията за „най-добрата“

класификационна функция съответства на разделителна хиперплоскост, която преминава през средата на двата класа, разделяйки двата. След като тази функция бъде определена, нов екземпляр от данни може да бъде класифициран и принадлежи към положителния клас [28].

4. Експериментални резултати и анализ

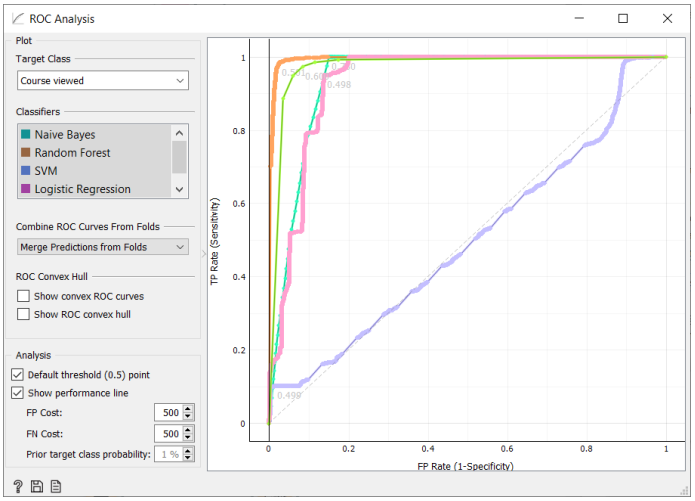
Прецизността е една от критериите за оценка на ефективността на модела и се изчислява като съотношение на истински положителни класифицирани елементи, разделени на сбор от истински положителни и невярно положителни елементи в тестовия набор. Диапазонът на точност е от 0 (най-малка точност) до 1 (най-голяма точност). Измерените резултати за прецизност, получени от избраните алгоритми за класификация, са показани на фиг. 1. Най-добрият резултат от гледна точка на точността се постига чрез класификационен алгоритъм Random Forest: 0.934.



Фиг. 1 Тестване и оценка на избраните алгоритми за анализ на данни

ROC кривите се използват за наблюдение на класификаторите и сравнение между моделите за класификация. Кривите на ROC за тестваните модели и резултатите от алгоритмите за тестване на класификацията са показани на фиг. 2.

Кривата на ROC демонстрира няколко неща: Тя показва компромис между чувствителност и специфичност (всяко увеличаване на чувствителността ще бъде придружено от намаляване на специфичността). Колкото по-близо кривата следва лявата граница и след това горната граница на пространството ROC, толкова по-точен е тестът. Кривата изобразява фалшиво положителна скорост на x-ос спрямо истинска положителна скорост по y-ос. Колкото по-близо кривата следва лявата граница и след това горната граница на ROC пространството, толкова по-точен е класификаторът. Като се имат предвид разходите за фалшиви положителни и фалшиви отрицания, може да се определи и оптималният класификатор, в случая това е Random Forest.



Фиг. 2 ROC анализ на класификационните модели

Матрицата на объркване в таблица I дава броя на случаите между действителния и прогнозирания клас. Матрицата е полезна за наблюдение кои конкретни случаи са класифицирани погрешно.

Таблица 1. Матрица на объркването за случай на алгоритъм random forest

No	Стойност	**Actual**	**Predicted**
1	Преглед на отчета за дейността	1	1
2	Събитието в календара е създадено	6	0
3	Събитието в календара е актуализирано	1	0
4	Модулът на курса е създаден	13	8
5	Модулът на курса е изтрит	11	13
6	Преглед на списък с екземпляри на модула на курса	7	9
7	Модулът на курса е актуализиран	95	98
8	Гледан модул на курса	26223	26245
9	Курсът се търси	6	4
10	Разделът на курса е актуализиран	2	0
11	Прегледан потребителски доклад за курса	44	0
12	Преглед на курса	30255	30653
13	Дискусията е създадена	1	0

14	Разгледана дискусия	88	67
15	Екземпляр за записване е създаден	7	0
16	Прегледан доклад за преглед на степен	131	131
17	Прегледан потребителски доклад за степен	1362	1362
18	Гледан доклад за гледане	2	2
19	Групирането е изтрито	6	0
20	Елементът е създаден	8	8
21	Елементът е изтрит	8	8
22	Гледан доклад на дневника на живо	1	0
23	Прегледан доклад за журнала	5	6
24	Прегледана скорошна активност	163	3
25	Назначена роля	1048	1146
26	Ролята не е назначена	1032	984
27	Известно съдържание е публикувано	1	1
28	Потребител, записан в курс	1048	1015
29	Преглед на потребителски списък	277	406
30	Потребителски профил е гледан	380	9
31	Прегледан потребителски доклад	510	510
32	Потребителят не е записан от курса	1032	1085

4. Заключение

Learning analytics и Gaming analytic в реално време на големите данни (big data), произведени от съвременните платформи за електронно обучение и образователни игри, за адаптиране към обучението, насочено към обучението, е едно от основните предизвикателства. Документът предлага процедури за анализ на данни от LMS и образователни игри посредством пет основни алгоритъма от машинното обучение. Основната цел на тези процедури е да се подпомогне структурирането и съхраняването на големи данни от разнородни източници, както от LMS, така и от образователна игра. Освен това да се идентифицират модели, като се анализират поведението на обучаемите и позволявайки анализи на данни с описателни, прогнозни и предписателни резултати.

Използвана литература

[1] Tempelaar, Dirk, et al. "Student profiling in a dispositional learning analytics application using formative assessment." *Computers in Human Behavior* 78 (2018): 408-420.

[2] Chatti, Mohamed Amine, et al. "A reference model for learning analytics." *International Journal of Technology Enhanced Learning* 4.5-6 (2012): 318-331.

[3] Schumacher, Clara, and Dirk Ifenthaler. "Features students really expect from learning analytics." *Computers in Human Behavior* 78 (2018): 397-407.

[4] Knight, Simon, Andrew Gibson, and Antonette Shibani. "Implementing learning analytics for learning impact: Taking tools to task." *The Internet and Higher Education* 45 (2020): 100729.

[5] J. Liebowitz, "Thoughts on recent trends and future research perspectives in big data and analytics in higher education," in *Big Data and Learning Analytics in Higher Education: Current Theory and Practice*, Springer International Publishing, 2016, pp. 7–17.

[6] H. Aldowah, H. Al-Samarraie, and W. M. Fauzy, "Educational data mining and learning analytics for 21st century higher education: A review and synthesis", *Telematics and Informatics*, vol. 37. Elsevier Ltd, pp. 13–49, 01-Apr-2019, doi: 10.1016/j.tele.2019.01.007.

[7] Romero, Cristóbal, Sebastián Ventura, and Enrique García. "Data mining in course management systems: Moodle case study and tutorial." *Computers & Education* 51.1 (2008): 368-384.

[8] Aher, Sunita B., and L. M. R. J. Lobo. "Combination of machine learning algorithms for recommendation of courses in E-Learning System based on historical data." *Knowledge-Based Systems* 51 (2013): 1-14.

[9] Chen, Lei, et al. "Personalized itinerary recommendation: Deep and collaborative learning with textual information." *Expert Systems with Applications* 144 (2020): 113070.

[10] Zhu, Haiping, et al. "A multi-constraint learning path recommendation algorithm based on knowledge map." *Knowledge-Based Systems* 143 (2018): 102-114.

[11] Monfil-Contreras, Erick Ulisses, et al. "RESYGEN: A Recommendation System Generator using domain-based heuristics." *Expert systems with applications* 40.1 (2013): 242-256.

[12] Zhou, Pingyi, et al. "Is deep learning better than traditional approaches in tag recommendation for software information sites?" *Information and software technology* 109 (2019): 1-13.

[13] Gutiérrez, Francisco, et al. "LADA: A learning analytics dashboard for academic advising." *Computers in Human Behavior* (2018): 105826.

[14] Moodle, "Analytics - MoodleDocs," 2019. [Online]. Available: https://docs.moodle.org/36/en/Analytics. [Accessed: 13-Mar-2020].

[15] C. Romero, S. Ventura, and E. García, "Data mining in course management systems: Moodle case study and tutorial," *Comput. Educ.*, vol. 51, no. 1, pp. 368–384, Aug. 2008, doi: 10.1016/ j.compedu.2007.05.016.

[16] A. Bovo, S. Sanchez, O. Heguy, and Y. Duthen, "Clustering moodle data as a tool for profiling students," in *2013 2nd International Conference on E-Learning and E-Technologies in Education*, ICEEE 2013, 2013, pp. 121–126, doi: 10.1109/ICeLeTE.2013.6644359.

[17] De Medio, Carlo, et al. "MoodleREC: A recommendation system for creating courses using the moodle e-learning platform." *Computers in Human Behavior* 104 (2020): 106168.

[18] Khaled, Abdelaziz, Samir Ouchani, and Chemseddine Chohra. "Recommendations-based on semantic analysis of social networks in learning environments." *Computers in Human Behavior* 101 (2019): 435-449.

[19] Milen Petrov, Asen Asenov, Adelina Aleksieva-Petrova, "A as in Audio: Facilitating the Automatic Generation of Audio Lectures ", *Proceedings of the International Conference on E-Learning* in the Workplace New York, NY, USA June 15-17, 2016 (ICELW), editor: David Guralnick, Ph.D., 2016

[20] Petrov M., Aleksieva-Petrova A., Design Of Rest Client Architecture For Course Resources Download And Package, *10th International Technology, Education and Development Conference*, editors: L. Gómez Chova, A. López Martínez, I. Candel Torres, IATED Academy, 2016, pp.6513-6521, doi:doi:10.21125/inted.2016.0535

[21] Adelina Aleksieva-Petrova, Veska Gancheva and Milen Petrov "Software Architecture for Adaptation and Recommendation of Course Content and Activities Based on Learning Analytics" in *Proc of Int. Conf. on Applied Mathematics & Computational Science*, Venice, Italy, March 21-23, 2020, to be published.

[22] Aleksieva-Petrova, A., I. Chenchev, and M. Petrov. "LMS Data-Collection, Processing and Compliance with EU GDPR.", EDULEARN19 Proceedings, IATED, ISBN: 978-84-09-12031-4 / ISSN: 2340-1117, 2019

[23] Lean Yu, Shouyang Wang, and K. K. Lai, "An integrated data preparation scheme for neural network data analysis," *IEEE Trans. Knowl. Data Eng*., vol. 18, no. 2, pp. 217–230, Feb. 2006, doi: 10.1109/TKDE.2006.22.

[24] I. Iguyon and A. Elisseeff, "An introduction to variable and feature selection," *J. Mach. Learn. Res*., vol. 3, pp. 1157–1182, 2003, doi: 10.1162/153244303322753616.

[25] S. I. Gallant, "Perceptron-Based Learning Algorithms," *IEEE Trans. Neural Networks*, vol. 1, no. 2, pp. 179–191, 1990, doi: 10.1109/72.80230.

[26] Orange Data Mining, [Online]. Available: https://orange.biolab.si/

[27] L. Breiman, "Random Forests." *Machine Learning*. 45 (1): 5–32. doi:10.1023/A:1010933404324.

[28] A. Aleksieva-Petrova, M. Petrov, P. Georgonikos, "Web application for document classification with Naïve Bayes Algorithm." in *Proceedings of the International Scientific Conference Computer Science'2018*, Kavala, Greece, 2018.

[29] Wu, Xindong, et al. "Top 10 algorithms in data mining." *Knowledge and information systems* 14.1 (2008): 1-37.