ПРИЛАГАНЕ НА DATA MINING ТЕХНИКИ В ЕЛЕКТРОННОТО ОБУЧЕНИЕ

доц. д-р Евдокия Сотирова, Университет "проф. д-р Асен Златаров",
доц. д-р Даниела Орозова - Бургаски свободен университет

APPLYING DATA MINING TOOLS IN E-LEARNING

Evdokia Sotirova, Daniela Orozova

Abstract: *In a series of research papers, the authors study some of the most important processes of functioning of universities and construct their models. The main focus in this paper is to analyze some possibilities for the applying of the Data Mining technology in electronic learning.*

Key words: *e-learning, electronic university, data mining, database.*

Увод

През последните години сме свидетели на широко използване на Data Mining техниките за анализ и прогнозиране в различни сфери на съвременния живот. Data Mining (DM) е процес на търсене на скрити, неочевидни закономерности (шаблони информация) в данните [4]. Технологията DM е процес на получаване на предварително неизвестни, нетривиални, практически полезни и достъпни знания, необходими за вземане на решение в различни сфери от човешката дейност [2]. Инструментите на DM могат да намират закономерности самостоятелно и да строят хипотези за взаимовръзките. Основни DM техники [2, 4] са: дървета на решенията, асоциативни правила, невронни мрежи, генетични алгоритми, OLAP (сложен статистически анализ, анализ в реално време) и други.

Някой DM средства автоматично избират най-подходящия техника (алгоритъм) и настройки, при други – те се указват от потребителя. Изборът на подходящ алгоритъм и настройки благоприятства за успешното изследване. Повечето средства на DM са разработени в рамките на теорията на изкуствения интелект.

Най-често използваните методи за описание на данните при различните приложения са: класификационен анализ, регресионен анализ, асоциативен анализ, клъстерен анализ, характеристичен анализ, сравнителен анализ, еволюционен анализ

1. Възможности за интегриране на Data Mining техники с електронното обучение

Широкото разпространение на информационните и комуникационни технологии в днешно време налага необходимостта от глобална образователна реформа в традиционните обучаващи системи. Характерни за висшето образование днес са организационни промени, увеличаване на компютърните ресурси, включващи web-базирани технологии, нарастване на образователните on-line ресурси, административни услуги, доставяни чрез мрежовите ресурси. От тук и необходимостта от повишаване квалификацията на персонала по принципа „обучение през целия живот".

Вземането на решения за процеса на обучение изисква изследване на поведението на студентите, анализиране на данни и оценяване на ефективността от педагогическите стратегии. Средствата на DM могат успешно да бъдат интегрирани с университетската информационна система на електронен университет. Предпоставки за използване на DM са например:

- нараства обема на данните, които трябва да се поддържат и обработват;
- развиват се големи, интегрирани бази от данни и като резултат се получават много данни, но малко информация за аналитика;
- недостиг на време и ресурси за анализиране на всички връзки;
- откриване на закономерности в данните, които могат да служат за разбиране и прогнозиране на академичните дейности на студентите.

Технологията DM може да се приложи в различни направления на електронен университет, някои, от които са разгледани по-долу.

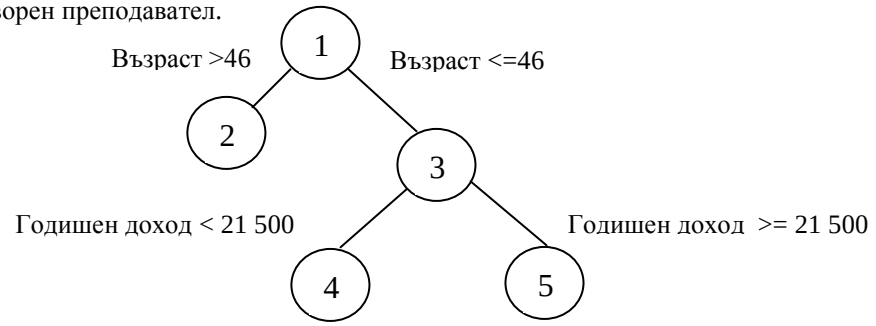
1.1. Оценяване на преподавателите

В [5] е разгледан процеса на оценяване на преподавателите в университет. Оценката на преподавателя е функция на усреднената оценка на студентите от изпита за курса, усреднената оценка от студентските отговори на анкетите за преподавателя и оценка за научната дейност на преподавателя.

В [8] е реализирано дърво на решение за оценяване на удовлетреността на преподавателите. В корена му са всички преподаватели. Така първият възел разделя данните във възли 1 и 2, базирани на преподавателската възраст. После всеки от тези възли разделя данните, базирани на условието относно “среден годишен доход”.

Номер на преподавател	Възраст	Стаж	Среден годишен доход	**Удовлетвореност**
1	51	4,5	11 500	+
2	45	15,0	3 500	-
3	36	3,4	21 500	+
4	47	6,1	36 000	+
5	42	14,5	7 000	-
6	50	2,5	15 000	+
7	40	11,9	6 000	-
8	31	4,10	2 000	-
9	38	10,2	5 500	-
10	37	7,5	31 500	+

Използвайки данните от таблицата алгоритъмът намира Max(възраст)=45, за която целевият атрибут „Удовлетвореност” има стойност “false” (-), т.е. не е удовлетворен преподавател.



Следва справка за всеки възел на дървото: свързаното с него правило, стойността на целевата функция, броя случаи и мерките **“сигурност”** и **“подкрепа”.**

Възел	Правило	Целева стойност	Брой случаи	Сигурност	Подкрепа
1		Удовлетворен	10	5/10=0.5	5/10=0.5
2	Възраст > 46	Удовлетворен	3	3/3 = 1	3/10=0.3
3	Възраст <= 46	Не Удовлетворен	7	5/7=0.7	7/10=0.7
4	Възраст <= 46 и доход < 21 500	Не Удовлетворен	5	5/5 = 1	5/10=0.5
5	Възраст <= 46 и сметка >= 21 500	Удовлетворен	2	2/2 = 1	2/10 =0.2

Прилагането на DM техника върху получените оценки (напр. невронна мрежа с право предаване) може да спомогне за изграждане на прогнозиране и формиране за цялостното развитие.

1.2. Оценяване на студентите

В [3] е създаден модел, който описва организацията на електронно обучение. Въвеждат се критерии за оценяване знанията на студентите по зададени теми, чрез прилагане на оценки от интуиционистки размити множества.

Оценките указващи нивото на усвоеност или неусвоеност на дадена единица знание представляват наредени двойки от реални числа от множеството [0,1]x[0,1]. В наредената двойка $<\mu, \nu>$, μ съответства на броя верни отговори съотнесен към общия брой отговори, а ν съответства на броя грешни отговори съотнесен към общия брой отговори. Степента на несигурност $\pi = 1 - \mu - \nu$ представя случаите, когато отговорите не могат да бъдат дефинирани точно или е допусната техническа грешка. Навсякъде наредените двойки се дефинират в смисъла на теорията на интуиционистки размитите множества. В началото, когато все още не е извлечена информация за ТО, всички оценки получават стойност <0,0>. Текущата оценка, $(k+1)$-вата по ред за $k \geq 0$, се пресмята на базата на предишните оценки, (ако има такива) по формулата:

$$<\mu_{\kappa+1}, \nu_{\kappa+1}> = <\frac{\mu_k k + m}{k+1}, \frac{\nu_k k + n}{k+1}>,$$

където $<\mu, \nu>$, е предишната оценка, а <*m,n*> е оценката от текущата задача за $m, n \in [0,1]$ и $m + n \leq 1$. Така в оценката за всяко умение се съдържа информация както от предходните събития, така и от последно решавания проблем.

В [7] оценките на студентите от положения изпит (тест) по конкретна дисциплина се реализират от трислойна невронна мрежа. Студентите попълват затворен тест с *m* въпроса с три възможни отговора (a, b и c), на базата на които се формира оценката. На входовете на невронната мрежа постъпват конкретните отговори на студентите, както и верните отговори за теста.

На трите й изхода a_μ, a_ν и a_π се получават интуиционистки размити оценки. Първият изход дава степента на усвояване на знанията μ; вторият – степента на неусвояване на знанията ν, и третият – степента на несигурност π = 1-μ-ν.

№	Отговори на всеки от въпросите										μ	ν	π
1	T	T	T	T	T	T	T	T	T	T	0.9978	0.0001	0.0002
2	T		T	T	T	T	T	T	F	F	0.7999	0.1002	0.0983
3	F	F	T			T	F	T	T	T	0.6008	0.2991	0.1038
4	T	T	F		T	T	T		T	F	0.7998	0.1001	0.0995
5	T	T	T	F			T	T	F	T	0.8010	0.0998	0.1016
6	F	F	T	F	F		T	F	T		0.3886	0.5093	0.1024
7	T		T	T	T	T	F	F	F	T	0.5992	0.2988	0.0997

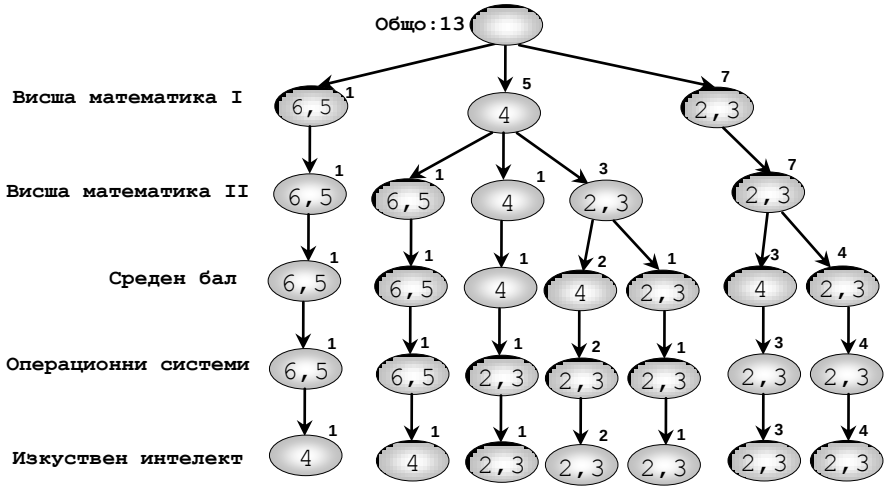
Моделирането и изучаването на процеса на електронно обучение и прилагането на средства на Data mining са необходими за процеса на персонализиране на курсове за дистанционно обучение. На базата на получените оценки могат да се въведат допълнителни мерки за анализ и промяна на курсовете на обучение. Това от своя страна е път към повишаване на качеството на обучението във висшето училище.

В това изследване прилагаме техниката "дърво на решението" с цел създаване на модел за предвиждане на студентската успеваемост по дадена дисциплина, в случая „Изкуствен интелект". Базираме се на предшестващите дисциплини „Висша математика I", „Висша математика II", „Операционни системи" и средния бал. Некогнитивните фактори като: стил на учене, мотивация и интерес и отделено време за учене, както и уменията преподавателите и обучаващата среда не са включени тук, те ще бъдат включени в бъдеща разработка.

От общо 195 записа с данни, които се отнасят за 39 студента като 31 мъже и 8 жени. Студентите са от специалност Информатика и компютърни науки. Всеки студент е свързан с 5 записа от данни. За 39 студента, това са 5×39=195 записи с данни.

Записите за студентите са разделени в две множества: за класификация (обучаващо множество) и множество за проверка (валидация). В обучаващото множество съдържа записите относно първите 13 студента по азбучен ред на имената в курса, а записите относно другите 26 студента се използват като валидиращо множество.

Алгоритъмът „дърво на решението [1] е приложен за построяване на дърво на решението. Максималната дълбочина на дървото е 5. Максималният брой клонове от един възел е 3, защото критерият във всеки разделя оценките в следните възможности: оценка 2 или 3; оценка 4; оценка 5 или 6.



Генерирани са множество правила, от които представяме 5 с най-висока степен на доверие (сигурност на предвиждането).

Правило I: ***IF*** Висша_математика_I = 2 или 3 ***AND***
Висша_математика_II = 2 или 3 ***AND*** Среден_бал = 2 или 3
THEN Изкуствен_интелект = 2 или 3 Сигурност: 7/7=1 = 100%

Правило II: ***IF*** Висша_математика_II = 5 или 6 ***AND*** Среден_бал = 5 или 6
THEN Изкуствен_интелект = 4 Сигурност: 2/2=1 = 100%

Правило III: ***IF*** Висша_математика_II = 2 или 3 ***AND***
Операционни_системи = 2 или 3
THEN Изкуствен_интелект = 2 или 3 Сигурност: 10/11=0.91 = 91%

Правило IV: ***IF*** Среден_бал = 4 ***AND*** Висша_математика_II = 2 или 3
THEN Изкуствен_интелект = 2 или 3 Сигурност: 6/10=0,60 = 60%

Правило V: ***IF*** Операционни_системи = 2 или 3
THEN Изкуствен_интелект = 2 или 3 Сигурност: 11/11=1 = 100%

За проверка на модела се прилага валидиращото множество от данни. Резултатите показват, че точността на предвиждането за "Изкуствен интелект" е за 22 студента (84.61%) от 26 студента.

Заключение

В доклада се дискутират възможности за интегриране на Data Mining техники с електронното обучение. Приложена е техниката "дърво на решението" с цел създаване на модел за предвиждане на студентската успеваемост по дадена дисциплина, в случая „Изкуствен интелект", базирайки се дисциплините „Висша математика I", „Висша математика II", „Операционни системи" и средния бал. Генерирано е множество от специфични "if-then" правила, като точността на предвиждането е 85 % на база валидиращо множество.

Литература:

1. Han, Jiawei and Micheline Kamber. 2007. *Data Mining: Concepts and Techniques*. San Fransisco, CA: Morgan Kaufmann publishers.
2. Fayyad, Piatetsky-Shapiro, Smyth, and Uthurusamy, Advances in Knowledge Discovery and Data Mining, AAAI/MIT, 1996.
3. Orozova D., H. Aladjov, K. Atanasov, Generalized Net For Machine Learning With Current Estimations, Advanced Studies in Contemporary Mathematics, 2002, pp. 61-76.
4. Parsaye K., A Characterization of Data Mining Technologies and Processes, The Journal of Data Warehousing. 1998.
5. Shannon A., D. Dimitrakiev, E. Sotirova, M. Krawczak, T. Kim, Towards a Model of the Digital University: Generalized Net Model of a Lecturer's Evaluation with Intuitionistic Fuzzy Estimations, Cybernetics and Information Technologies, Vol 9, No 2. Sofia, 2009, 69-78.
6. Shannon, A., D. Langova-Orozova, E. Sotirova, K. Atanassov, P. Melo-Pinto, T. Kim, Generalized Net Model for Adaptive Electronic Assessment, Using Intuitionistic Fuzzy Estimations, Computational Intelligence, Theory and Applications, Computer Science I, Springer, (2005), 291-297.
7. Sotirov, S., E. Sotirova, D.Orozova, Neural network for defining intuitionistic fuzzy sets in e-learning, Ituitionistic Fuzzy Sets,Thirteen Int. Conf., Sofia, 2009, 33-36.
8. Sotirova E., D. Orozova, Generalized net model of the phases of the data mining process, System Research Institute, Polish Academy of Sciences, Warsaw, 2009.
9. Tang, T., & McCalla, G. Student modeling for a web-based learning environment: A data mining approach. Eighteenth national conference on artificial intelligence, Menlo Park, CA, USA (2002), p. 967.