ПРОГНОЗИРАНЕ НА УСПЕВАЕМОСТТА НА СТУДЕНТИТЕ ЧРЕЗ БАЙЕСОВИ КЛАСИФИКАТОРИ

Людмила Димитрова
Университет "Проф.д-р Асен Златаров", Катедра „Компютърни и информационни технологии", бул."Проф.Я.Якимов" №1, 8010 Бургас, България, E-mail: lyudim@gmail.com

STUDENTS PERFORMANCE PREDICTION USING BAYESIAN CLASSIFIERS

Lyudmila Dimitrova
University "Prof.Asen Zlatarov", Faculty "Computer and Information Technology" Prof. Jakimov Str. 1, Bourgas, Bulgaria , E-mail: lyudim@gmail.com

ABSTRACT

This study presents an approach to predict student performance in disciplines related to the study of programming languages in the University "Prof. Ass.Zlatarov" - Bourgas. The approach is based on a Naïve Bayes classifier and other Bayes classifiers, supported by WEKA tool. A set of 159 students records during the period from September 2008 to June 2011 is used to train the classifiers. The percentage of correctly classified instances varies from 77,99% to 91.92% depending on the classifier and the number of input variables. The results suggest that basic training of the students has no significant predictive power on performance, while information about their abilities, diligence, motivation and activity in the learning process can predict their grades. The resulting forecasts can be used by the teacher in optimizing the learning process.

***Key words**: data mining, Naïve Bayes classifier, improved Naïve Bayes, student performance*

Въведение

Понастоящем информационните технологии се използват във всички области на живота. Образователната система не прави изключение. Специална област на науката – т.н. интелектуален анализ на данните в образованието (Educational data mining - EDM), се занимава с обработка и анализ на информацията, натрупана в образователния процес. Целта е извличане на нови знания от вече представените в базите данни с цел повишаване на качеството на обучението.

Всеки преподавател се нуждае от информация и инструменти за подобряване на курса по воденета от него дисциплина. Например, въз основа на анализа на вече проведените курсове, преподавателят може да адаптира учебната програма според възможностите и интересите на студентите или да раздели обучаващите се на групи - според успеваемостта, активността, базовата подготовка, мотивацията за развитие. Тези, както и много други задачи по оптимизиране на обучението се решават със средствата на EDM [1,2].

Предсказването на успеваемостта на студените е важен момент в образователния процес и е било обект на моделиране в редица разработки [6-11]. Целта на настоящата работа е: (1) – да подготви база данни за успеваемостта на студентите в дисциплини, свързани с изучаване на програмни езици (2) - да идентифицира факторите, влияещи върху овладяването на материала, респ. върху крайната оценка (3) – да конструира модел за класификация чрез средствата на data mining и (4) – да валидира получения модел чрез използването на други данни.

Интелектуален анализ на данните - класификация на данните

Интелектуалният анализ на данните (data mining) е аналитичен процес, целящ разкриването на моделите и тенденциите, скрити в големи обеми от данни [3,4]. Тази задача

се решава чрез съчетаване на множество математически инструменти (от класическия статистически анализ до новите кибернетични методи) и най-новите разработки в областта на информационните технологии. Повечето от техниките за data mining са разработени в рамките на теорията на изкуствения интелект.

Една от най-полезните техники на data mining е класификацията - техника за предсказване на значения на данните, използваща резултати, получени на базата на други данни[5]. При класификацията се конструира модел – класификатор с цел предсказване на атрибута етикет на класа. Например, при изучаване на база от данни за успеваемостта на студенти, етикети на класа ***успех*** могат да бъдат „*слаб*", „*среден*", „*добър*", „*много добър*" и „*отличен*". В общия случай, класификацията на данните е двустъпков процес. На първия етап се изгражда класификатор за описание на предварително дефинирано множество от концепции или класове. Това е обучаващият етап, при който алгоритъмът за класификация изгражда класификатора чрез изучаване на обучаващото множество данни, състоящо се от кортежи (tuples) входни данни и асоцииран с тях атрибут етикет на клас. Тъй като етикетът на класа е зададен, този етап се нарича контролирано (supervised) обучение и може да се разглежда като получаване на преобразуването или функцията y=f(**x**), която може да предскаже етиката на класа y за даден кортеж **x**. Преобразуването (mapping) се представя като правила за класификация, дърво на решенията или математични формули. На втория етап моделът се използва за класификация.

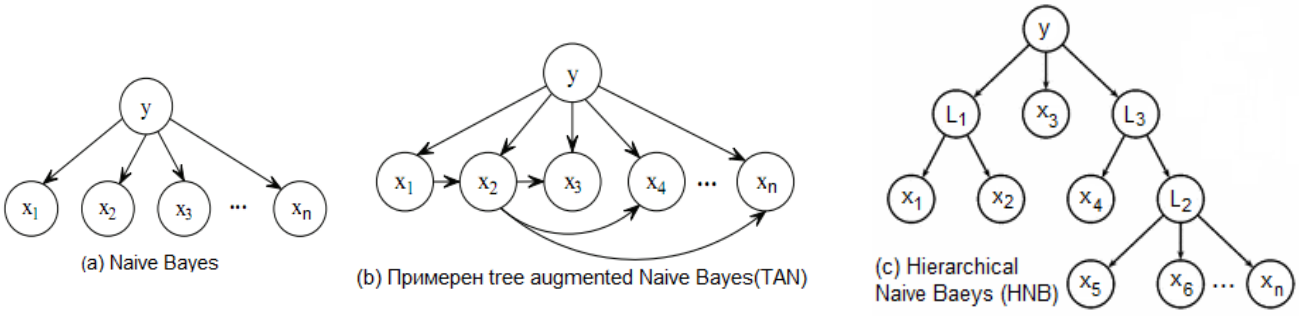
Байесови мрежи и класификация

Байесовата мрежа (BN) [12] показва причинно-следствените вероятностни връзки между променливите в дадена предметна област. Вероятностните връзки между случайните променливи $\{x_1, x_2, \ldots, x_N\}$ в байесовата мрежа се представят чрез насочен ацикличен граф (directed acyclic graph, DAG). Количествената част на модела включва множество от таблици на условните вероятности за възлите на модела. Те задават вероятността на всяко състояние на дадения възел при всички възможни комбинации на състоянията на родителските възли .

Основно свойство на байесовите мрежи е логическият извод (inference). Това е процесът на получаване на оценката на модела – съвместното разпределение на вероятностите при наличните наблюдения за някои от възлите в мрежата в конкретния момент. Ако означим родителските възли на възела x_i с $Pa(x_i)$, съвместното разпределение на вероятностите се представя като произведение от множителите (factors), дефинирани от структурата на мрежата:

$$P(x) = \prod_{i=1}^{n} P(x_i \mid Pa(x_i)) \qquad (1)$$

За дадено множество данни D={(**x**, y)}, където y е променливата – клас, BN класификаторите (BNCs) описват D чрез съвместното разпределение на вероятностите P(**x**,y) и го преобразуват в разпределение на условните вероятности P(y|**x**) за определяне на етикета на класа.



Фиг.1. Наивен байесов класификатор (a) и две от подобренията му (b, c). L_1,L_2 и L_3 са латентни променливи

Задачата за класифициране на нова инстанция на базата на значенията на кортежа от атрибутите ѝ $D_a=\{x_1, x_2, \ldots, x_n\}$ към един от класовете $y_i \in y$ се свежда до определяне на значението с максимална постерирорна вероятност (MAP):

$$y_{MAP} = \arg\max_{y_j \in y} P(y_j \mid x_1, x_2, ..., x_n) = \arg\max_{y_j \in y} \frac{P(x_1, x_2, ..., x_n \mid y_j)P(y_j)}{P(x_1, x_2, ..., x_n)}$$
$$= \arg\max_{y_j \in y} P(x_1, x_2, ..., x_n \mid y_j)P(y_j) \qquad (2)$$

Показаният на фиг.1,a наивен байесов класификатор (NBC) допуска, че всяка променлива - атрибут зависи само от променливата-клас, т.е.

$$P(x_1, x_2, ..., x_n \mid y) = P(x_1 \mid y)P(x_2 \mid y)\ldots P(x_n \mid y) \qquad (3)$$

при което $P(x_i \mid y)$ за категорийни атрибути се определя като относителната честота на извадките, имащи стойност x_i като i-ти атрибут в класа y.

Тъй като на практика условието за независимост на променливите атрибути не се спазва, са разработени много подобрения на наивния байесов класификатор, които отчитат връзките между атрибутите [13-14]. Два от тях са показани на фигури 1,b и 1,c.

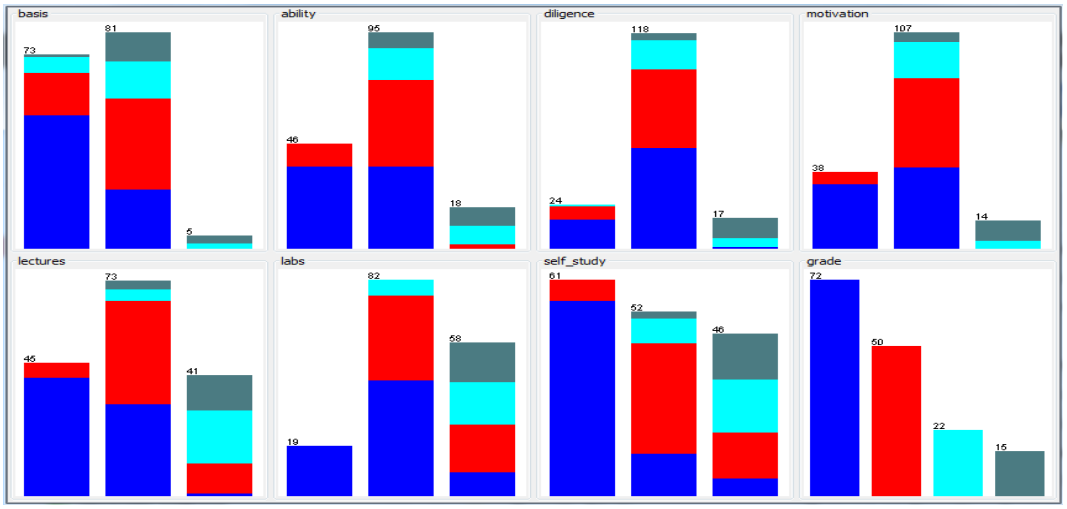
Съществуват редица свободни софтуерни продукти, реализиращите техниките на data mining: TANAGRA, WEKA, пакета e1071 на R, KNIME, Orange и много други[15]. В настоящата работа са използвани първите два от изброените.

Разработване и използване на модела за класификация на студентите по успех

Като източници на данни за процеса data mining в са използвани записите за присъствие и активността на студените на учебните занятия по програмиране, наблюденията за техните качества и изпитните протоколи. На този етап е въведена информацията за 204 студента. На фиг. 2 е показана извадка от базата данни на Excel, която се използва директно в TANAGRA**.** Тъй като базата данни ще служи както за обучение на класификатора, така и за прогнози, т.е. за класификация на нови записи, в таблицата като допълнителна колона е добавена "*status*". Като значение на това поле в обучаващата извадка от 159 записа е зададено значението „*learning"*, а на останаващите 44 тестови записи – значение „*to_classify"* (фиг 2). Значението ***grade*** на записите 160-204, предназначени за тестване на модела, не се използва, и може да има произволно значение. В случая за сравнение с резултатите от класификацията за записани оценките по протокол.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P
1	ID	name	year	degree	specialty	course	age	basis	ability	diligence	motivation	lectures	labs	self	grade	status
158	157	Георги Вичев	2011	master	KST	V	21-25	high	average	moderate	high	average	good	good	<6>	learning
159	158	Мариан Панков	2011	master	KST	V	21-25	mid	average	high	high	good	good	good	<6>	learning
160	159	Драгомир Стамболи	2011	master	KST	V	21-25	mid	average	high	high	good	good	good	<6>	learning
161	160	Велина Стоянова	2011	master	KST	V	21-25	low	average	high	average	good	good	average	<5>	to_classify
162	161	Желязко Желев	2011	master	KST	V	21-25	low	average	low	average	average	average	average	<4>	to_classify
163	162	Здравко Златев	2011	master	KST	V	21-25	mid	average	high	high	good	good	good	<6>	to_classify

Фиг. 2. Данни за оценка на успеваемостта на студентите в изучаването на програмни езици



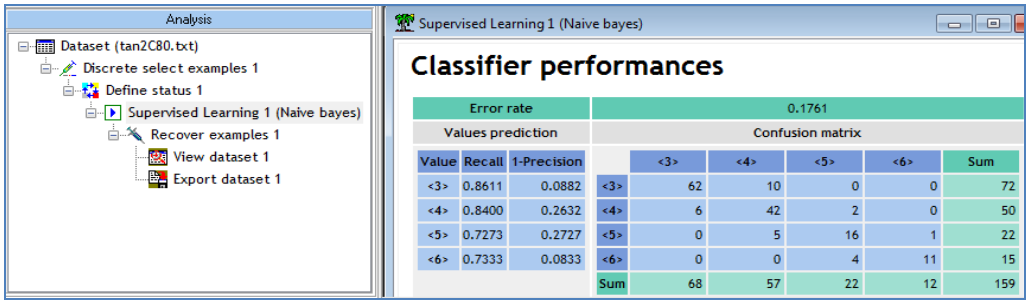
Фиг. 3. Разпределение на оценките за значенията на 7-те основни входни променливи и класа- grade

Софтурните пакети за data mining позволяват филтриране на атрибутите за конкретната задача, т.е. избор и използване за конкретната задача само на част от променливите в базата данни. Използваните за класификацията в TANAGRA променливи и възможните им значения са показани в таблица 1. Разпределението по значения на тези основни променливи е показано на фиг 3.

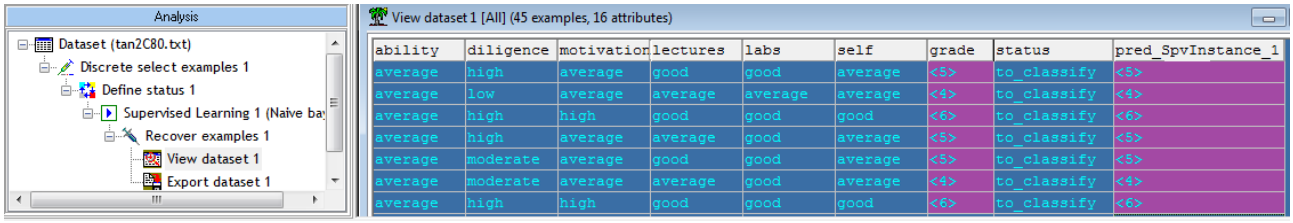
Таблица 1. Променливи, използвани в класификацията по успех в Tanagra

Променлива	Описание	Възможни значения
1: base	Ниво на студента в началото на обучението по дадената дисциплина	{low, mid, high}
2: ability	Възможности на студента в изучаването програмните езици – в редки случаи се срещат студенти, за които това не е подходящ избор	{low, average, high}
3: diligence	Усърдие на студента в учебните часове и при изпълнение на самостоятелните задания	{low, moderate, high}
4: motivation	Мотивация на студента за изучаване на дадения програмен език	{low, average, high}
5: lectures	Посещаемост на лекциите по дисциплината	{poor, averare, high}
6: labs	Посещаемост на лабораторните занятия	{poor, averare, high}
7: self-study	Степен на самоподготовка	{poor, averare, high}
8: **grade - клас**	Оценка от изпита	{3, 4, 5, 6}

На фиг.4 е показана класификацията с използване на наивен байесов класификатор в TANAGRA. Грешката в класификацията при така зададените данни за класификация е 17.61%. Матрицата на неточностите (confusion matrix) показва грешките за всеки от етикетите на класа. С най-голяма точност – 86.11% се класифицират тройките - от 72 тройки по протокол правилно са класифицирани 62, останалите 10 са класифицирани като четворки. Фиг. 5 илюстрира част от резултатите от квалификацията на тестовата извадка.



Фиг. 4. Резултати от класификацията с наивен байесов класификатор в Tanagra



Фиг. 5. Използване на класификатора за прогнози

Софтуерният пакет за data mining WEKA включва няколко алгоритъма за байесова класификация. В таблица 2 е показана точността на класификацията, получена при прилагането им за разглежданата задача при два варианта на входните данни: *(1)* –основните 7, използвани и в TANAGRA (табл.1) и *(2)* - всички 12 променливи (без *ID* и *name*) от изходната база (фиг. 1). По-чувствително увеличение на точността на квалификацията се получава при използване на подобрен алгоритъм за класификация – 88.68 % при HNB срещу 81.86% за Naïve Bayes. В използваната извадка добавянето да допълнителни 5 променливи: година на обучение, образователна степен, специалност, курс и години няма значима роля в прогнозирането на успеваемостта. Като атрибути с най-голяма сила за предсказването на класа WEKA определя *ability, diligence, motivation, lectures, labs* и *self-study* – т.е. използваните на този етап данни показват, че основата, получена в училище, не е от решаващо значение за успешното усвояването на програмните езици.

Таблица 2. Резултати от класификацията по байесовите алгоритми, включени във WEKA

Класификатор	Входни променливи: 1-7 от таблица 1		Входни променливи: колони C:N фиг. 2	
	Коректно класифицирани инстанции	Некоректно класифицирани инстанции	Коректно класифицирани инстанции	Некоректно класифицирани инстанции
NaiveBayes	130 - 81.76%	29-18.24%	134 – 84,28%	25- 15,73%
NaiveBayesSimple	124 - 77.99%	35 - 22.01%	134 – 84,28%	25- 15,73%
BayesNet	130 – 81.76%	29-18.24%	134 – 84,28%	25- 15,73%
NaiveBayesUpdateable	130 – 81.76%	29-18.24%	134 – 84,28%	25- 15,73%
AODE	139 - 87.42%	20 -12.58%	140 – 88.05%	19 – 11.95%
AODEsr	141 – 88.68%	18 – 11.32%	146 – 91.82%	13 – 8.18%
HNB	141 - 88.68%	18 - 11.32%	145 – 91.20%	14 – 8.80%
WAODE	140 – 88.05%	19 – 11.95%	146 – 91.82%	13 – 8.18%

Изводи

Получени са модели за класификация с използването на наивен байесов класификатор и модификациите му, включени в софтуерния пакет WEKA. Моделите класифицират студентите по успех с точност от 77,99% до 91.82% в зависимост от вида на класификатора и броя на използваните входни променливи. Повишаване на точността на предсказване може да се получи чрез стандартната за класификацията предварителна обработка на данните, като и чрез добавяне на допълнителни информативни атрибути. Получените въз основа на модела прогнози могат да се използват от преподавателя за адаптиране на процеса на преподаване по време на семестъра и като база при оценяването на студентите на изпитите.

Литература

1. Peña-Ayala, "*Educational data mining: A survey and a data mining-based analysis of recent works*", Expert Systems with Applications, Volume 41, Issue 4, Part 1, March 2014, Pages 1432-1462

2. Romero, S. Ventura, "*Educational data mining: A survey from 1995 to 2005*", Expert Systems with Applications, Volume 33, Issue 1, pp 135–146 (2007)
3. Nettleton, „*Commercial Data Mining*", Morgan Kaufmann, 2014
4. Y.Zhao, Y.Cen, „*Data Mining Applications with R*", Academic Press, 2013
5. A.Giusti, G. Ritter, M. Vichi (editors), "*Classification and Data Mining*", Springer, 2013
6. S. K. Yadav, S. Pal „*Data Mining: A Prediction for Performance Improvement of Engineering Students using Classification*" , World of Computer Science and Information Technology J. (WCSIT), vol. 2 (2), pp 51-56, 2012
7. K.S.Priya, A.V. Senthil Kumar , "*Improving the Student's Performance Using Educational Data Mining*", Int. J. Advanced Networkong and Applications, Vol. 4, Iss. 4, pp 1680-1685 (2013)
8. B.K. Baradwaj, S.Pal, "*Mining Educational Data to Analyze Student's Performance*", Int.J. of Advanced Computer Science and Applications, Vol.2, No. 6, pp 63-69, 2011
9. K.Itoh, H.Itoh, K. Funahashi "*Forcasting student's grades using Bayesian network model and an evaluation of its usefulness*", 13th ACIS Int. Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing", pp 332-336, 2012
10. F.A. Dorca. L.V. Lima, M.A.Fernandes, C.R. Lopes, "*Comparing strategies for modeling student learning styles [through reinforcement learning in adaptive and intelligent educatiolal system: An experimental analisys*", Expert Systems with Applications, 40, 2092-2101, 2013
11. R. Torabi, P. Moradi, A. R. Khantaimoori, „*Predict Student Scores Using Bayesian Networks*", Procedia - Social and Behavioral Sciences, Volume 46, 2012, Pages 4476-4480
12. Koller Daphne, Nir Friedman, "*Probabilistic Graphical Models: Principles and Techniques*", MIT Press, 2009
13. N. Friedman, D. Geiger, M. Goldszmidt, „Bayesian Network Classifiers", Machine Learning, November 1997, Volume 29, Issue 2-3, pp 131-163
14. S. F.Shazmeen1, M. M.Ali Baig, M.R. Pawar, „*Performance Evaluation of Different Data Mining Classification Algorithm and Predictive Analysis*", IOSR Journal of Computer Engineering, vol. 10 (6), pp 1-6, 2013
15. *"33 Top Free Data Mining Software"*, онлайн: http://www.predictiveanalyticstoday.com/top-15-free-data-mining-software/