ISBN 978-981-11-0008-6
Proceedings of 2016 6th International Workshop on Computer Science and Engineering (WCSE 2016)
Tokyo, Japan, 17-19 June, 2016, pp. 760-764
doi: 10.18178/wcse.2016.06.136

Prediction Model for Students' Performance in Java Programming with Course-content Recommendation System

Digna S. Evale[1], Menchita F. Dumlao[2], Shaneth Ambat[3], Melvin Ballera[3]

College of Computer Studies, AMA University Philippines

Abstract. Comparative analysis among different data mining algorithm for attribute selection and classification was conducted on this two-phase study which aims to predict the students' performance in Java Programming and be able to generate recommendations. Processes in Knowledge Discovery in Database (KDD) was followed in finding patterns among the historical data. Logistic Regression and Correlation-based Feature Selection was used for finding significant predictors. Classifiers such as CHAID, Exhaustive CHAID, CRT, QUEST, J48, BayesNet, NaïveBayes and JRip were implemented and it was found out that J48, on the context of this study, has the most straightforward rules set and the highest percentage of prediction. For the second phase evolutionary prototyping implemented through Ruby on Rails was done to develop a web-based examination module that will determine the students' index of learning style and to assess their prior knowledge in Java. A course-content recommendation presenting the learners' strengths and weaknesses in the subject with suggested method of learning style will be automatically generated by the system.

Keywords. Classification Algorithm, Prediction Model, Performance Prediction, Attribute Selection, Course-content Recommendation, Tree Classifiers, Index of Learning Style

1. Introduction

Recommender system which is a subject of intense research for almost a decade now is gaining popularity at present in the internet domain, and the network society continuously feels its increasing importance because of the drawbacks brought about by the "information overload" widely experienced by the users in the cyber world [1]. With millions of authentic and non-authentic data scattered in diverse electronic locations today, filtering what is the best need is indeed a tough task. Thus, a system, i.e., a recommender system, which can aid that selection process is undeniably beneficial. As stated by [2], e-commerce is now widely implementing the concept of using recommendation systems to attract customers but is still an area of exploration in the field of education.

Students or learners, just like customers or product consumers, are also in need of personalized recommendations for learning resources and activities that will match-up to their "personal needs, preferences, prior knowledge and current situation" to facilitate a more personalized learning environment that will promote learning outcomes [3]. And today, with the advent of electronic learning and the diversity of learners especially in Higher Education, an application that can provide a more personal approach in teaching and learning process is indeed an interesting field of study especially in improving the teaching and learning process.

Being proficient in programming have several advantages especially if you are well-versed in a programming language widely used in the field. On July 2015, the Institute of Electronics and Electrical Engineers Spectrum – the flagship publication of IEEE- published a list of top ten programming languages in the world and Java Programming is seating at the topmost position [4]. Java is a robust, platform-independent, distributed and object-oriented programming language. Java is not only a best choice for system development which requires OP concepts but for internet programming as well [5]. Indeed, there are so many career paths or job opportunities available for Java Programmers, living in a technology-driven world where almost every field needs computerized systems and applications [6].

It is a common knowledge that the students' aptitude or ability to carry-out logical analysis and competence in doing logic formulation is a great factor affecting their performance in programming subjects [7]. Several studies were already conducted regarding the challenges in teaching programming and Java Technologies [8, 9, 10, 11 & 12]. Teaching programming requires that the methodologies and strategies be appropriate with the learning styles of the target learners. Determining the students' learning style which could reveal whether they are a global or sequential, active or reflective, sensing or intuitive and visual or verbal – factors which greatly affect the way how they could be motivated to learn. This could also help the teacher since there are students too who perceived programming as a boring subject [13]. An assessment of prior knowledge about the subject is also vital in order to determine the strengths and weaknesses of students. Ignoring these factors might set off serious difficulties in learning [13 & 14]. The vast scope and complexity of Java, which if not given proper means of introduction and appropriate presentation of materials might also posit learning problems [15].

With those reasons, the researcher aimed to create a course content recommendation system for Java programming based on learners' predicted performance which will benefit the learners especially those who will be evaluated to be in need of learning assistance. First, educational data mining which is now gaining popularity in improving institutional effectiveness [17] and learners' classification to improve learning process [18] will be employed. Second, an application system will provide an assessment of the students' index of learning styles and prior knowledge in Java in order to be able to determine the recommended course-content as well as the appropriate learning style. This way, the students will be given a chance to assess their strengths and weaknesses as learners, to identify the topics that they could possibly encounter difficulties which later might cause them to fail and the good thing is that the result can be used to serve as an immediate preventive remedy to improve the students' performance.

2. Methodologies

This study used two models for the system development: these are the Fayyad Knowledge Discovery Process Model which is also considered as one of the best Academic Research Model [19] for the data mining in education and evolutionary prototyping for the development of system prototype. There are five stages namely for KDD; selection, pre-processing, transformation, mining and interpretation. In selection, possible attributes is collected for data set while in prep-processing is the filtering and removing of irrelevant data. Transformation, on the other hand, is determining the most suited data mining technique to provide the best prediction algorithm. Mining phase discovering the pattern capture through classification rules, regression models or decision tree. Evaluation or interpretation is the process of visualization extracted from models.

Table 1 shows the different attributes relevant to model the learner's profile. The attribute was evaluated through Waikato Environment for Knowledge Analysis (WEKA) data mining tool and IBM Statistical Package for the Social Science (SPSS). There were 8 attributes namely gender, age, course, section, schedule and 3 academic performance for programming languages.

Table 1. Possible Attributes for Building Learners' Model

Attributes	Values	Description
gender	{male, female}	gender of the learner
Age	{ less than or equal 16, 17, 18, 19, 20, 21 above 21}	age of the learner
course	{BS Information Technology, BIT Computer Technology, BS Computer Science}	course of the learner
section	{A, B, C, D, E, F, G, H, I, J}	section where the learner belongs
schedule	{morning, afternoon, evening}	class schedule (7am-12nn: morning; 12:nn-5pm: afternoon; 5pm-8pm:evening)
grd_Prog1	{1.0,1.25,1.5,2.0,2.25,2.5,3.0,4.0,5.0}	grade of the learner in Turbo C programming
grd_Prog2	{1.0,1.25,1.5,2.0,2.25,2.5,3.0,4.0,5.0}	grade of the learner in C ++ programming
grd_Prog3	{1.0,1.25,1.5,2.0,2.25,2.5,3.0,4.0,5.0}	grade of the learner in Visual Basic programming

The prototype system contains two sets of modules – the first one is for prediction and the other is an examination module for evaluation of learning style and assessment of prior knowledge on Java Programming. Upon signing up, the student will be asked to enter all the information found to be significant in predicting his performance together with some other data for security process.

Prediction whether the student will pass or fail on Java Programming will be automatically generated by the system, if predicted as 'failed', the learner will be required to proceed with the examination module. The first exam is a forty one-item test which can classify the student as global or sequential, active or reflective, sensing or intuitive and visual or verbal learner. The next one is a one hundred forty – item questionnaire which will test the level of understanding of the students in Java Programming. Questions are from the concepts which are already included in their prior programming subjects, but the implementation is in Java language.

At the end of the test, necessary reports that may be used to improve the teaching and learning process will be generated both for the students and the teachers. The prototype was created using Ruby on Rails and SQL as indicated in Fig. 3. The system can be accessed via a web-browser for greater accessibility to students.

3. Results And Discussions

In order to identify the major attributes relevant in developing a data model and rule sets for predicting the performance of Java learners, historical data was filtered and analysed using different algorithms. Attribute selection was done using Standard Regression Analysis, Forward and Backward Conditional Regression, Likelihood Ratio, and WALD Test using SPSS. WEKA was also used to conduct pre-processing thru filtering by *AttributeSelection* before the data was subjected to an attribute evaluator. Table 2 shows the summary of results of the attribute selection process.

From the eight original attributes, there are two variables which were found to be insignificant. With a critical *p value* of .05 (significant predictors should have smaller critical p value), Binary Logistic Regression using SPSS found *section* and *course* as highly insignificant with .747 and .221 p value respectively. *Gender* can be interpreted as not highly significant since its p value is .016 both in Standard and Forward Regression and .053 in Backward Regression.

Table 2. Summary of Attribute Selection Result

Tool	Method		Attribute w/p-value >0.000	p-value
IBM SPSS 20	Standard Regression		Gender	.016
			Section	.747
	Forward Regression	Conditional Likelihood Ratio WALD	Gender	.016
	Backward Regression	Conditional Likelihood Ratio WALD	Course	.221
			Gender	.053
*Other attributes p-value =0.000 * Predicted Percentage Correct of Classification Table = 89.9%				
Tool	Method		Attribute Removed	merit of best subset
Weka	CfsSubsetEval(BestFirst)		course, section	.239
	* Predicted Percentage Correct of Classification Table = 76.1%			
	Method		Attribute with no. of folds !=10	No. of folds
	CfsSubsetEval (GreedyStepWise)		Gender	70%
			Course	0%
			Section	0%
	* Other attributes appeared 10 times (100%) in 10-fold validation			

Attributes were further analyzed using WEKA, in pre-processing step, Filtering thru *AttributeSelection* was done, and the result was parallel with the SPSS data – course and section was automatically removed, meaning they are found as highly insignificant.

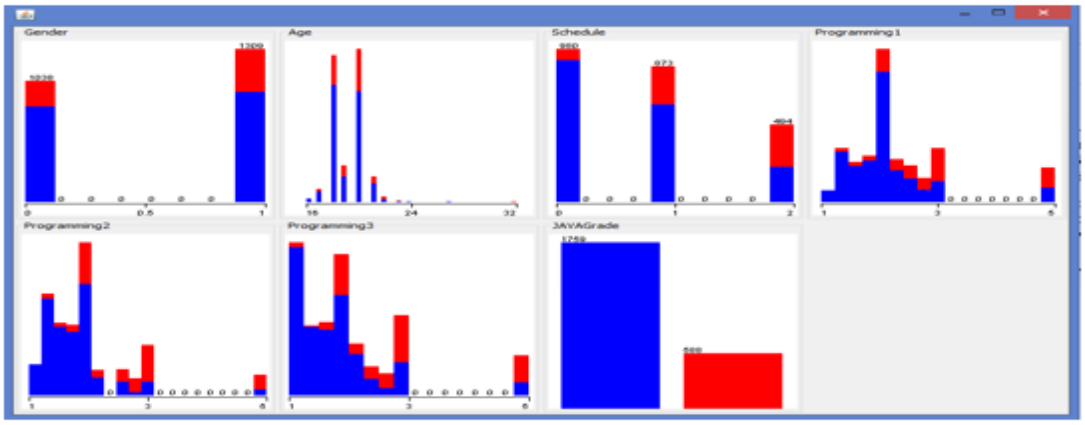

Fig. 1 – Visualization of Different Attributes

Figure 1 shows the visualization of each attributes of experimental data. To further verify the significance of attribute gender, *CfsSubsetEvaluation* was performed. In *BestFirst* method, gender was found significant with 0.239 value of merit of best subset found (value is from 0 to 1 representing the incorrectly classified instances) which means that there is 76.1% of correctly classified instances. In *GreedyStepWise* search method selected through Cross Validation, course and section are not found in any of the ten folds while gender appeared in 7 out of 10 folds (70%). With those data, the researcher came up with following attributes as significant predictors: *age, gender, schedule, grade in Programming 1, grade in Programming 2, and grade in Programming 3.*

3.1 Determining the Best Data Mining Algorithm

The six significant attributes was used in determining the best model and rule sets for prediction. Classification was done using several data mining algorithm, the one which gave the highest percentage of correct prediction was used. Table 3 summarizes the results.

Table 3. Summary of the Accuracy Result of Different Algorithm

Technique	Algorithm	Accuracy	Kappa / *std. error
Tree Classifiers	CHAID	87.5	.007
	Exhaustive CHAID		
	CRT	89.8	.006
	QUEST	89.4	.006
	J48	94.74	.8464
Bayes Classifiers	BayesNet	92.84	*.7938
	Naïve Bayes	84.63	*.5642
Rule Classifier	JRip	94.21	*.8302

The result shows that in predicting the performance of students in Java Programming subject, J48 is the best algorithm to be used since it has the highest percentage of accuracy in making predictions and at the same time has the highest Cohen's Kappa value which means that the prediction is strongly reliable with 64% to 81% reliability as indicated in Table 4 [20].

Table 4. Cohen's Kappa Equivalent Value

Value of Kappa	Level of Agreement	% of Data that are Reliable
0-.20	None	0-4%
.21-.39	Minimal	4-15%
.40-.59	Weak	15-35%
.60-.79	Moderate	35-63%
.80-.90	Strong	64-81%
Above .90	Almost Perfect	82-100%

J48 which is an improved implementation of C4.5 is one of the most widely-used classification algorithm today among the other tree classifiers that can be used for predicting performance. It is gaining popularity because of its high percentage of correct prediction, optimized decision tree diagram and straightforward rule sets which doesn't need complicated interpretation [21].

4. Conclusion

Attributes such as *age, gender, schedule, grade in Programming 1, grade in Programming 2,* and *grade in Programming 3* are found to be the significant predictors in this study, while *section* and *course* has no significant effect in predicting the learners' performance in Java Programming. Assessment of the learners in terms of their learning styles and level of understanding in the subject area is recognized by the researcher as among the important factors in improving the learning process, thus, a web-based examination module for thatis integrated in this study. For future studies, it is suggested that aside from system evaluation using ISO 9126, prediction reliability of the system be tested against the actual data by implementing Cohens' Kappa statistics.

5. References

[1] J. Itmazi., & M. Megias. Using Recommendation System in Course Management System to Recommend Learning Objects. The International Arab Journal of Information Technology, 5(3), 2008. 234- 240. Retrieved from http://ccis2k.org/iajit/PDF/vol.5,no.3/4-169.pdf

[2] R. Sikka., A. Dhankhar., & C. Rana. A survey paper on e-learning recommender system. *International Journal of Computer Applications, 47*(9), 2012, 27-30.

[3] H. Drachsler., H. G. Hummel., & R. Koper, R. Recommendations for learners are different: applying memory-based recommender system techniques to lifelong learning. Paper presented at the SIRTEL workshop at the EC-TEL 2007 Conference. September, 17-20, 2007, Crete, Greece.

[4] B. Cass. (2015, July 20). The 2015 Top Ten Programming Languages. IEEE Spectrum. Retrieved from http://spectrum.ieee.org/computing/software/the-2015-top-ten-programming-languages

[5] D. Kafura.. *Object-Oriented Software Design, and Construction with Java*, Prentice Hall. 2000.

[6] S. Nelson. (2013, August 15). *The Benefits of learning Computer Programming.* [Web-log post]. Retrieved September 15, 2015 from http://www.collegeamerica.edu/blog/the-benefits-of-learning-computer-programming.

[7] I. Milne and G. Rowe, "Difficulties in Learning and Teaching Programming – Views of Students and Tutors," *Education and Information Technologies*, Vol. 7, No. 1, 2002, pp.55-66.

[8] N. Mehic., Y. Hasan., & B. KPMG. Challenges in teaching java technology. *Informing Science*, 2001. 365-371.

[9] J. Carter and T. Jenkins, "The Problems of Teaching Programming: Do They Change with Time?" *The Higher Education Academy, Information and Computer Sciences, 11th Annual Conference*, August 24-26, Durham, UK, 2010.

[10] D. Clark, C. MacNish and G. F. Royle, "Java as a Teaching Language – Opportunities, Pitfalls and Solutions," *Proceedings of the Third Australian Conference on Computer Science Education, ACM*, July 4-6, Brisbane, Australia, 1996.

[11] M. Pendergast. Teaching introductory programming to IS students: Java problems and pitfalls. *Journal of Information Technology Education: Research*, *5*(1), 2006. 491-515.

[12] K. Ala-Mutka. Problems in learning and teaching programming. *Codewitz Needs Analysis*. 2012.

[13] T. Jenkins, "On the Difficulty of Learning to Program," *The Higher Education Academy, Information and Computer Sciences, 2nd Annual Conference*, March 22, Wolverhampton, UK, 2002.

[14] D. Adair., & M. Jaeger. Difficulties in teaching and learning the Java programming language. In *Proceedings of the 17th International Conference on Engineering Education* 2011. pp. 21-26.

[15] R. Lister, "Teaching Java First: Experiments with a Pigs-Early Pedagogy," Australasian Computer Education, 6th Conference, January 18-22, Dunedin, NZ, 2004.

[16] Higher Education in Numbers. *Higher Education Institutions.* Retrieved June 15, 2015 from http://www.ched.gov.ph/index.php/higher-education-in-numbers/higher-education-institutions/

[17] R. A. Huebner. A survey of educational data-mining research. *Research in Higher Education Journal*, Vol. 19, 2013, pp. 1-13.

[18] C. Romero., S. Ventura., P. G. Espejo., & C. Hervás. Data mining algorithms to classify students. In *Educational Data Mining 2008*.

[19] G. Piatetsky-Shapiro., U. Fayyad., P. Smith., & R. Uthurusamy. Advances in Knowledge Discovery and Data Mining, AAAI. 1996.

[20] M. L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, *22*(3),2012. 276-282.

[21] A. Rajput., R.P. Aharwal., M. Dubey., S. P. Saxena & M. Raghuvanshi. J48 and JRIP Rules for E-Governance Data. *International Journal of Computer Science and Security (IJCSS)*, *5*(2), 2011. 201.