Work in Progress - Programming Misunderstandings Discovering Process Based On Intelligent Data Mining Tools

Paola Britos, Elizabeth Jiménez Rey, Darío Rodríguez, Ramón García-Martínez
CAPIS-ITBA, III-FI-UNLP, LSI-FI-UBA, {pbritos, drodrigu, rgm}@itba.edu.ar, ejimenez@fi.uba.ar

Abstract - **We present research work in progress that focuses on data mining tools used for helping teachers to apply a three step knowledge discovering process to diagnose students' misunderstandings (and their causes) related to their programming errors.**

Index Terms - TDIDT algorithms, bayesian networks, data mining, students' misunderstandings diagnosis.

INTRODUCTION

Data mining has been addressed as an effective way of discovering new knowledge from data sets of educational processes, data generated by learning systems or experiments, as well as how discovered information can be used to improve adaptation and personalization [1]. Among interesting problems data mining can help to solve: determining which are common learning styles or strategies [2]-[4], predicting the knowledge and interests of a user based on past behavior, partitioning a heterogeneous group of users into homogeneous clusters or detecting misconceptions in learning processes

One of the most common techniques of data mining are the decision trees (TDIDT) used for discovering knowledge in rule format which constitutes a model that represents the knowledge domain subjacent to the available examples of it.

A Bayesian network is a directed acyclic graph in which each node represents a variable and each arc represents a probabilistic dependency which specifies the conditional probability of each variable given its parents; the variable to which the arc points to is dependent (cause-effect) on the variable in the origin of this one [5].

To discover common learning misconception of learners, research described in [6], employ the association rule to data mine learner profile for diagnosing learners' common learning misconception during learning processes. The association rules that occurring misconception A implies occurring misconception B can be discovered utilizing the proposed association rule learning diagnosis approach.

In this paper we present research work in progress that focuses on data mining based tools for helping teachers to diagnose students' misunderstandings (and their causes) related to their programming errors. To do this some considerations of what to measure and why are presented in section *problem domain*, a three step *knowledge discovering process* is presented in the homonymous section, a real case taken from an introductory programming course students population is presented in section *case example*, and *finally preliminary conclusions and future research* are drawn.

PROBLEM DOMAIN

To model student's misconceptions a list of components to be evaluated has been identified. Some of these components, its justification and the rank of possible values are described in the following paragraphs.

Student applies refinement method: this component looks for to diagnose if student has acquired the analytical ability of decomposing a complex problem into more simple parts (divide and conquer strategy). The tabulated answers may be: <yes, no, incomplete>.

Student discovers algorithm: this component looks for to evaluate if student has developed the ability to put in sequence programming primitive sentences in a logic way according to problem's objective to be solved. The tabulated answers may be: <yes, no>.

Student obtains a generalized solution: this component looks for to evaluate if student has incorporated the algorithm concept as a procedure of resolution of a problem family class to which the proposed problem belongs. The tabulated answers may be: <yes, no>.

Student uses begin/end correctly: this component looks for to evaluate if student has incorporated the concept of primitive sentences block or compose sentence block that have to be executed conceptually as a unique sentence. The tabulated answers may be: <yes, no>.

Students obtain a logic solution: this component looks for to evaluate if student has the maturity to obtain a solution design with grater quality. The tabulated answers may be: <good, regular, bad>.

Student controls repetitive cycle end condition: this component looks for to evaluate if student has incorporated the algorithm concept as a procedure that has to provide a solution in a finite time verifying that an iterative structure effectively ends. The tabulated answers may be: <yes, no >.

Student uses logic connectors correctly: this component looks for to evaluate if student has detected the relation among the type of problem and the conditions which rule the problem solution generalization. Blocking conditions is an issue to be evaluated in this component. The tabulated answers may be: <yes, no>.

Student develops an infinite cycle: this component looks for to evaluate if student controls iterative cycles due to verify the value change of variables used in conditional expression associated to iterative cycle end. The tabulated answers may be: <yes, no>.

KNOWLEDGE DISCOVERING PROCESS

To discover programming misconceptions of students, this research focuses on using intelligent systems based data mining tools: rules induction by TDIDT algorithms and Bayesian networks. These tools are used in a three step knowledge discovering process:

First step deals with building a data base (manually by instructor) on a standard characterization of each student and her/his programming style and misconceptions. This characterization focuses on aspects of: student program development methodology, student developed program functionality, student program design quality.

Second step deals with discovering rules (by using TDIDT algorithms) which establish a relation among programming misconceptions with possible causes. The obtained rules are applied to confirm or deny courseware structure teachers hypothesis. Experiment results indicate that applying this step, teacher can correctly discover learners' common misconception causes.

Third step deals with discovering the weight each cause has on each misconception (by using Bayesian networks). This allows establishing a rank of importance of the causes of misconceptions in order to propose recovering teaching strategies.

CASE EXAMPLE

We have carried on a preliminary experiment on a first course of programming with a population of 45 students. Professor of this course wish to explore which misconceptions are related to the fact that student doesn't discover correctly the exercise associated algorithm ("student discovers algorithm = no").

From the database developed in Step 1 of the course on programming examinations, we apply Step 2: and using TDIDT algorithm we obtain the following set of rules:

IF	student applies refinement method = no
THEN	student develops an infinite cycle =yes
IF	student uses logic connectors correctly = no
THEN	student develops an infinite cycle =yes
IF	student develops an infinite cycle =yes
THEN	student uses logic connectors correctly = no
IF	student uses begin/end correctly =no
THEN	student uses logic connectors correctly = no
IF	student uses logic connectors correctly = no
THEN	student controls repetitive cycle end condition = no
IF	student controls repetitive cycle end condition = no
THEN	student obtains a logic solution = bad
IF	student obtains a logic solution = bad
THEN	student obtains a generalized solution = no
IF	student obtains a generalized solution = no
THEN	student discovers algorithm = no

From this set of rules we can identified that "student discovers algorithm = no" has three possible causes: "student applies refinement method = no", "student uses logic connectors correctly = no" and "student uses begin/end correctly =no". We apply Bayesian Networks (BN) in Step 3 to determine which of the three identified causes has the greater impact on the considered problem: "student doesn't discover correctly the exercise associated algorithm". The results obtained are:

CAUSE	BN WEIGHT
"student applies refinement method = no",	84,88%
"student uses logic connectors correctly = no"	56,49%
"student uses begin/end correctly =no"	09,34%

PRELIMINARY CONCLUSIONS AND FUTURE RESEARCH

Data mining applied to help teachers to discover students' misunderstandings causes is a promising issue to explore as a new diagnosis tool area. Current results are promising but not conclusive on a basis of a set of 45 student's examination records of a Pascal first level course. Next step will be to tune the process over a population integrated by 300 students of the introductory programming course (Pascal programming language) of the Engineering School of the University of Buenos Aires.

REFERENCES

[1] Beck, J. , Calders, T., Pechenizkiy, M., Viola, S. 2007. *Workshop on Educational Data Mining*. ICALT'05: 933-934.

[2] Schulte, C., Bennedsen, J. 2006. *What do teachers teach in introductory programming?*. ICERW´06: 17-28.

[3] Salgueiro, F., Cataldi, Z., Britos, P., Sierra, E. y García Martínez, R. 2006. *Selecting Pedagogical Protocols using SOM*. Research in Computing Science Journal, 21: 205-214.

[4] Britos, P., Cataldi, Z., Sierra, E., García-Martínez, R. 2008. *Pedagogical Protocols Selection Automatic Assistance*. LNAI 5027 (in press).

[5] Felgaer, P., Britos, P. and García-Martínez, R. 2006. *Prediction in Health Domain Using Bayesian Network Optimization Based on Induction Learning Techniques*. International Journal of Modern Physics C 17(3): 447-455.

[6] Chen , C., Hsieh, Y. 2005. *Mining Learner Profile Utilizing Association Rule for Common Learning Misconception Diagnosis*. ICALT'05: 588-592.

AUTHOR INFORMATION

Paola Britos, Associate Professor, Software & Knowledge Engineering Center, Buenos Aires Institute of Technology.

Elizabeth Jiménez Rey, Assistant Professor, Engineering School; Postgraduate Student, Intelligent Systems Laboratory, University of Buenos Aires.

Darío Rodriguez, Assistant Professor, Software & Knowledge Engineering Center, Buenos Aires Institute of Technology.

Ramón García-Martínez, Full Professor, Software & Knowledge Engineering Center, Buenos Aires Institute of Technology; Director, Intelligent Systems Laboratory, University of Buenos Aires.