# Analyzing Student Performance Competence Using Intricate Tool for Educational Mining

**Mrs. Ananthi[1] ,Mrs. Mythili[2]**

[1]M Phil Research Scholar, Department. of Computer Science, [2]Head & Associate Professor, Department of Information and computer Technology, Hindustan College of Arts and Science, Coimbatore.

**ABSTRACT - Now a day's Data Mining is becoming a common tool in education field. Data mining tools help in analytical methodology for detecting valuable information. This paper shows how data mining algorithms can help to discover pedagogically relevant data contained in databases available in web based educational systems.**

**Key Words:** Data mining, Association rules, Pedagogical, Clustering, Classification, Vsualization, k-means clustering, and Apriori algorithm.

## 1.INTRODUCTION

Web based educational systems collect large amounts of student's data from web logs to semantically rich data contained in student models. The focus of this research is to provide adaptation to a learner using the data stored in his/her student model. We try and explore ways to mine data in a more collective way just as a human teacher can adapt to an individual student, the same teacher can learn more about how students learn, reflect and improve his/her practice by studying a group of students.

The field of data mining is concerned with finding new patterns in large amounts of data. Widely used in business, it has scarce application to Education. Data mining can be used in the domain of education, for example to find out which alumni are likely to make larger donations. Here we are interested in mining student models in a pedagogical perspective. The goal of this paper is to define how to make data possible to mine, to identify which data mining techniques are useful and understand how to discover and present patterns that are pedagogically interesting both for learners and for teachers.

The process of tracking and mining such student data in order to enhance teaching and learning is relatively recent but there are already a number of studies trying to do so and researcher are starting to merge their ideas. The usefulness of mining such data is promising but still needs to be proven and stereotypical analysis to be streamlined. Some researchers already try and set up some guidelines for ensuring that.

Simple statistics, queries or visualization algorithms are useful in giving teachers an overall view of how a class is performing. For example, the authors use pedagogical scenarios to control interactive learning objects. Records are used to build charts that show exactly where each student is in the learning process, thus offering the teacher distant monitoring. Simple queries allow showing charts to teachers the details of all the exercises every student has taken up and successfully completed thereby making teachers aware of individual student's progress through the course. More sophisticated information visualization techniques are used to externalize student data and generate pictorial representations for course instructors to explore. Using features extracted from log data and marks obtained in the final exam, classification techniquesare used to predict student performance fairly accurately. These allow teachers to identify students at risk and provide advice ahead of the final exam. When student mistakes are recorded, association rules algorithm can be used to find mistakes often associated together. Combined with a genetic algorithm, concepts mastered together can be identified using student scores. The teacher may use these findings to reflect on his/her teaching and re-designing the course material.

The purpose of this paper is to synthesize and share various experiences of using data mining for education, especially to support reflection on teaching and learning, and to contribute to the emergence of stereotypical directions. Forthcoming section briefly explains various algorithms that are used.

## 2. ALGORITHMS AND TOOLS

Data mining involves variant algorithms that are varied in their methods and aims. It has data exploration and visualization techniques to represent results in a convenient way to the users. Set of tools and types of algorithms that areused briefly explained below. A data element will be called an individual. It is characterized by a set of variables. In most of the time, an individual is represented as a learner and variables can be marks scored, exercises attempted

by the learner, mistakes made, time spent, number of successfully completed exercises and so on. New variables may be calculated and used in algorithms, such as the average number of mistakes made per attempted exercise.

- Access tool is used to perform simple SQL queries and visualization.
- Clementine for clusteringmodels which focus on identifying groups of similar records and labeling the records according to the group to which they belong.
- Tada-Ed for classification and association rule to predict student's performance and efficiency of the course.
- SODAS to perform symbolic data analysiswhich is a relatively new field that provides a range of methods for analyzing complex datasets.

## 3. DATA EXPLORATION AND VISUALIZATION

Raw data and algorithm results can be visualized through tables and graphics such as graphs and histograms and as well as through more specific techniques such as symbolic data analysis. The aim is to display data along certain attributes and make extreme points, trends and clusters obvious to human eye. Clustering algorithms aim at finding homogeneous groups in data. *K-means*clustering and its combination with hierarchic clustering is used. Both methods rest on a distance concept between individuals. Euclidian distance is used.For example, given all the work done by a set of students, one may want to group the students into similar group based on their performance.

Classification is used to predict values for some variable. For example, given all the work done by a student, one may want to predict whether the student will perform well in the final exam.

Association rules find relations between items. Rules have the following form: X -> Y, support 40%, confidence 66% which could mean if students get X incorrectly, then they also get Y incorrectly, with a support of 40% and confidence of 66%. Support is the frequency in the population of individuals that contains both X and Y. A variant of the apriori algorithm is implemented that takes temporality into account. Taking temporality into account produces a rule X -> Y only if exercise X occurred before Y.

## 4. A CASE STUDY

### 4.INTRICATE TOOL ANALYSIS

A number of queries have been performed on databases collected by Logic - ITA to assist teaching and learning. This is a web based tutoring tool used at Sydney Universityfrom 2001. Its purpose is to help students practice logical forms and to inform the teacher of the class progress.

### 4.2 PERSPECTIVE OF USE

Students used the tool to their own discretion. A consequence is that there is neither a fixednumber nor a fixed set of exercises done by all the students.

### 4.2DATA STORED IN DATABASE

The tools' teacher module collates all the student models into the database that the teacher can query and mine. Twooften-queried tables of the database are mistake and correct step. The most common variables are shown in the table

**Table 1 - Common variables in table's mistake and correct step**

| | |
|---|---|
| **Login** | **Student's login id** |
| **Question id** | **The question id** |
| **Mistake** | **Mistakes made** |
| **Rule** | **Login rule involved/used** |

| | |
|---|---|
| **Line** | **Line number of the proof** |
| **start Date** | **Date exercise was started** |
| **end Date** | **Date exercise was finished** |

## 5. DATA MINING PERFORMED

Each year of data is stored in a different database. In order to perform any clustering, classification or association rule query, the first action to take is to prepare the data for mining. In particular, we need to specify two aspects what

element we want to cluster or classify: students, exercises, mistakes? An example could be to cluster students using the number of mistakes they made and the number of correct steps they entered. Tada-Ed provides a pre-processing facility, which allows the data to be minable. For instance, the database contains the list of mistakes. If we want to group that information so that we have one vector per student, we need to choose how the mistakes should be aggregated. For example, we may want to consider the total number of mistakes per type of mistake or a flag for each type of mistake and so on.

## 6. DATA EXPLORATION

Simple SQL queries and histograms can really allow the teacher to get a first overview of the class, what are the most common mistakes, the logic rules causing the most problems? Whatis the average number of exercise per student? Is there any student not finishing the exercise? The list goes on.

To understand better, how students use the tool, how they practice and how they come to master the tool and logical proofs, we also analyzed data focusing on the number of attempted exercises per student.

In SODAS, the population is partitioned into sets called symbolic objects. Our symbolic objects are defined by the number of attempted exercises and arecharacterized by the values taken for these newly constructed variables: the number of successfully completed exercises, the average number correct steps per attempted exercises, and the average number of mistakes per attempted exercise. A number of tables are obtained to compare all these objects. An example is given in the below table which compares objects according to the number of successfully completed exercises.

For example, the second line says that, among the students who have attempted two exercises, 13% could not complete any of them, 23% could complete one and 65% could complete both. And similarly for the other lines.

Using all the tables, we could confirm that, the more students practice the more successful they become at doing formal proofs. Interestingly though, there seems to be a number of exercises which large proportion of students finish most exercises. For 2002, as little as two attempted exercises seem to put them on the safe side since 65% of the students who attempted exercises are able to finish them both.

**Table - 2 Distribution of students according to number of attempted exercises (row) and the number of completed exercises (column)**

| Finish / Attempt | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 14 | 15 | 16 | 19 | 20 | 21 |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | 46 | 54 | | | | | | | | | | | | | | | | | |
| 2 | 13 | 23 | 65 | | | | | | | | | | | | | | | | |
| 3 | 6 | 11 | 39 | 44 | | | | | | | | | | | | | | | |
| 4-6 | 4 | 8 | 27 | 19 | 29 | 10 | 2 | | | | | | | | | | | | |
| 7-10 | 3 | | 6 | 18 | 36 | 12 | 18 | 3 | 3 | | | | | | | | | | |
| 11-15 | | | 16 | 16 | 16 | 21 | 5 | 5 | | | 11 | | 5 | 5 | | | | | |
| 16 | | | 17 | | | | | | | | | | | | 17 | 17 | 33 | | 17 |

## 7. ASSOCIATION RULES

Association rules are used to find mistakes often occur together while solving exercises. The purpose of looking for these associations is for the teacher to ponder and maybe to review the course material or emphasize subtleties while explaining concepts to students. Thus, it makes sense to have a support that is not too low. The strongest rule for 2004 is shown in table below. The first association rule says that if students make mistake, Rule can be applied but deduction incorrect while solving an exercise, then they also made a mistake. Wrong number of line references given while solving the same exercise. Findings are quite similar across the years

***Table – 3 Association Rules***

| | |
|---|---|
| ***M11 ➔ M12 [sup: 77%, conf: 89%]***<br>***M12 ➔ M11 [sup: 77%, conf: 87%]***<br>***M11 ➔ M10 [sup: 74%, conf: 86%]***<br>***M10 ➔ M12 [sup: 78%, conf: 93%]***<br>***M12 ➔ M10 [sup: 78%, conf: 89%]***<br>***M10 ➔ M12 [sup: 74%, conf: 88%]*** | ***M10: Premise set incorrect***<br>***M11: Rule can be applied,but deduction incorrect***<br>***M12: Wrong number of line reference given*** |

## 8. CLUSTERING AND VISUALIZATION

Clustering isused to group and characterize students with difficulties. For example, to characterize the students, thosewho failed to complete the attempted exercise. To do so, clustering is performed using sub-population both using k-means and Tada-Ed and combination of k-means and hierarchical clustering of Clementine because there is neither a fixed number nor a fixed number of exercises to compare students, determining a distance between individuals isnot obvious. A new variable, total number of mistakes made per student in an exercise are used. As a result, students with similar frequency of mistakes are out into same group. Histograms showing the different clusters revealed interesting patterns. Consider the below histogram obtained with Tada-Ed. There are three clusters 0, 1, and 4. From other windows we could observe that students in cluster 0 made many mistakes per exercise not finished, students in cluster 1 made few mistakes and students in cluster 4 made an intermediate number of mistakes. Students make many mistakes also use many different logic rules while solving exercises; this is shown with vertical, almost solid lines.

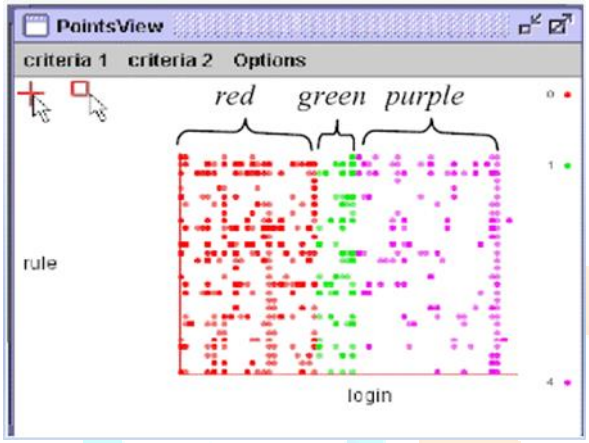


**Figure 1**. Histogram showing, for each cluster of students, the rules incorrectly used per student

On the other hand, another histogram displays exercises against students tells us that students from 0 to 4 have not attempted more exercises that students from cluster 1, who make few mistakes. This suggests that these students try out the logic rules from pop-up menu of the tool one after the other while solving exercises till they find the one the solves the problem.

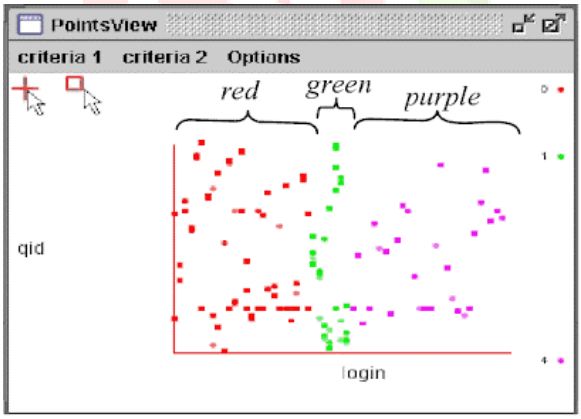


**Figure 2.** Histogram showing, for each cluster of students, the exercise attempted per student.

## 9.CLASSIFICATION

To predict exam marks, decision trees are built. The Decision Tree algorithm produces a tree like representation of the model it produces. From the tree, it is then easy to generate rules in the form of IF condition THEN outcome. Using as a training set, the previous year of student data, we can build and use a decision tree that predicts exam marks according to the attributes so that they can be used on the following year to predict the mark that student is likely to obtain.

***Table – 4 Some Results of decision tree processing. Accuracy of mark prediction using simple rounding of the mark***

| Attributes and type of pre-processing | Accuracy of mark | Accuracy of pass/fail | Diff. Avg (sd) real/predicted | Rel. error |
|---|---|---|---|---|
| Number of distinct rules in each exercise* Number of exercises per performance type^ | 51.9% | 73.4% | -0.2 (1.7) | 11% |
| Number of distinct rules* Sum of lines entered correctly in each exercise | 46.8% | 87.3% | -0.5 (1.9) | 18% |
| Number of exercises per nb of rules (interval)* Different performance achieved ^ | 45.6% | 86.1% | -0.4 (1.8) | 14% |
| Number of different length of exercises# Different performance achieved ^ | 43% | 88.6% | 0.14 (1.5) | 8% |
| Number of exercises per performance type^ Sum of lines entered correctly in each exercise | 44.3% | 86.1% | -0.3 (1.7) | 13% |
| Number of exercises per performance type^ Sum of rules used correctly (incl. repetition) | 44% | 86.1% | 0.1 (1.9) | 10% |
| Sum of rules used correctly (incl. repetition) | 43% | 87.3% | -0.22 (1.8) | 13% |
| Sum of lines entered correctly in each exercise | 43% | 87.3% | -0.22 (1.8) | 13% |
| Mistakes, in any form of pre-processing | <20% | | | |

* in order to avoid overfitting we have grouped number of rules into intervals: [0-5], [6-10], [10+] .
# for the same reason, the number of steps in exercises was grouped into intervals of 5.
^ Performance types were grouped into 3 types: unfinished, finished with mistakes, finished without mistake.

## 10.CONCLUSION

In this paper, we have shown how the discovery of different data patterns through different data mining algorithms and visualization techniques suggest to us a simple pedagogical policy. With the help of Data exploration focused on the number of attempted exercises we will be able to identify students at risk, those who have not trained enough. Clustering and cluster visualization led us to identify a particular behavior among failing students, when students try out the logic rules of the pop-up menu of the tool. A timely and appropriate warning to students at risk could prevent failing in the final exam. Therefore, it seems to us that data mining has a lot of potential for education and can bring a lot of benefits in the form of sensible, easy to implement pedagogical policies as above.

## 11. REFERENCES

[1].JoostBreaker,Bertbredewing,Chee-Kit-Looi,GordMccalla" Artificial intelligence in education" Supporting learning through intelligence and socially informed technology, IOS press Netherland, isbn 1-58603-530-4,2005.

[2].Merceron, A & K Yaccef "Web based Tutoring tool with mining facilities to improve learning and teaching" in Proceedings of 11th International Conference on Artificial Intelligence in Education".IOS Press 2003.

[3].Michel C. Desmarais, Ryan S. Baker, KalinaYacef,"Journal of education data mining" Vol 6,No 1(2014),vol 7,NO 1(2015),Vol 8,N0 1(2016),editorial acknowledgement JEDM

[4]. R. Baker "Data Mining for Education",In McGraw, B., Peterson, P., Baker, E. (Eds.) International Encyclopedia of Education (3rd edition), vol. 7, pp. 112-118. Oxford, UK: Elsevier. (2010)

[5].Abdulmohsonalgarni"Data mining in education",international journal of advanced computer science and application(IJACSA),VOL 7,NO 6,2016.

[6].C. Romero, S. Ventura, "Educational data mining: A review of the state of the art", IEEE Trans. Syst. Man Cybern. C Appl. Rev., vol. 40, pp. 601-618, Nov. 2010.