Improving the Quality of Education in Higher Schools by Developing the Students Capabilities using Data Mining Techniques

1Ahmed Ashraf, 2Hazem M. El-Bakry, 3Yehia El-Mashad, 4Samir Abd-Elrazik and 5Mohammed El-Desouky

1,2,3Information Systems Dept., Faculty of Computer and Information Sciences, Mansoura University, Mansoura, EGYPT
4Computer science Dept., Faculty of Computer and Information Sciences, Mansoura University, Mansoura, EGYPT
5Delta University, EGYPT

Abstract

The main objective of this research work is to develop the performance of education in higher schools e-learning systems. This is accomplished with the aide of data mining (DM) techniques. The proposed model is applied on different students. Data is collected using online school tests, reports and quizzes. This paper applies SVM with accuracy 89\%, Decision tree with accuracy 89\%, M5-Rules with RMS error equal to 1.4621 and Linear Regression with RMS error equal to 2.0017, 3.0089 and 3.6057. Once getting both first and second grades, the presented results show a high predictive accuracy. Not only past student's evaluations affected in their academic achievement, but also other factors like father's and mother's jobs and absences. Briefly, student performance can be improved depending on predictive results and enhancing school systems.

Keywords: Data Mining, Educational Systems, Classification, Regression, Decision Trees.

I. Introduction

Students have to learn in anywhere by any way. Electronic learning (E-Learning) is the employment of computers and web to help both learners and educators to learn anytime and anywhere without restrictions. E-learning has created new markets for education. It is far from the traditional trend of education which depends on the tremendous revolution of information technology. The use of Internet, computers and networks for learning purposes is called E-learning (1). Usual E-learning systems offer knowledge and evaluation for learners but our proposed system aims to ensure that students received educational content correctly by monitoring his handling of the course content and communicate directly with him to direct it to the correct way. On the other hand, as a result of the evolution of information systems, attention towards Data Mining is constantly increasing (2). Making decisions and achieving goals need to real information. So the student's information must be very close to reality. Examples are grades, social and demographic data. Because of the limited human ability to predict, an alternative tool to analyze a big data to make a decision is needed. Student learning is the best area for applying data mining applications because of the available data like databases, pages on web and all web process (19-52). There are many questions concerning the field of E-Learning that can be answered with techniques using Data Mining algorithms: Which type of students who have a credit hours system? Who need to study the course again? How to increase the number of our students? How to handle system's errors? How to prevent student converting to another E-Learning system? What are the methods of predicting student performance? How to improve the student's performance and achievements? The main points of this paper are to predict and enhance the student's performance as well as improve his achievement (3).
II. MATERIALS AND METHODS

It is known that the secondary education consists of three years. Students study many educational materials like Sciences, Historical, Geometric and Mathematics. The method of evaluation of tests varies, but a 20-point grading is often used, Starting from 0 and ending with 20. Students are evaluated on three levels: the first evaluating (G1N), the second evaluating (G2N) and the last evaluation (G3N) as shown in table 1. In order to keep pace with the development of information technology in learning we developed website to make e learning system instead of sheets system. The traditional systems have many disadvantages such as lack of data and lack of credibility. The database was built from three dimensions data registration on our website questionnaires and results are tested for each year. We designed our site to get special data divided into four sections grades, related feature, social and demographic which affect student performance (4). During the preprocessing step some variables had to be ignored due to the lack of discriminative value.

III. DATA MINING MODELS

There are many analytics tools (Classification and Regression Tree) used for extracting the most useful variables from the dataset. The output of each tool is the difference between them (classification is a discrete and regression is a continuous). The classification shows the result as percentage while in regression as Root Mean Squared (5). Classifications to be good should present a low Correct Classifications, while regression should present a low global error. All of these results come from using the equations:

\[ \Phi(i) = \begin{cases} 1, & \text{if } y_i = \hat{y}_i \\ 0, & \text{else} \end{cases} \]  

(1)

\[ \text{PCC} = \frac{\sum_{i=1}^{N} \Phi(i)}{N} \times 100(\%) \]  

(2)

\[ \text{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} (y_i + \hat{y}_i)^2}{N}} \]  

(3)

The dataset records is tested and applied on various classification algorithms using WEKA an Open source tool such as:

- Linear Regression (finding the best-fitting straight line through the point).
- SVM Support Vector Machines.
- C4.5.
- M5-Rules Algorithm.

3.1 C4.5

The most important feature in Decision tree (DT) is the use of a tree structure as simple representation for a set of rules that diG1 Suish values hierarchically (6). It is a useful tool for the classification (7). DT Algorithm is to discover the behavior of each attribute (8). Tree classification algorithm is used to make the prediction and understood the critical distribution of the data is easily (9).

3.1.1 Confusion Matrix

A confusion matrix (an error matrix) is a specific table layout that allows visualization of the performance of an algorithm, typically a supervised learning one (in unsupervised learning it is usually called a matching matrix) (10). This matrix used in the field of machine learning and specifically the problem of statistical classification. In the table layout rows and columns have a role, each row represents the instances in an actual class and each column of the matrix represents the instances in a predicted class (11). (Fig. 1) shows the prediction outcome using a confusion matrix. The main objective from confusion matrix is to describe classification model performance which applied to set of known data.

**Sensitivity**

\[ \text{Sens.} = \frac{TP}{p} = \frac{TP}{TP+FN} \]  

(4)

**Specificity**

\[ \text{Spec.} = \frac{TN}{n} = \frac{TN}{TN+FP} \]  

(5)

**Precision**

\[ \text{Prec.} = \frac{TP}{TP+FP} \]  

(6)

**Accuracy (ACC)**

\[ \text{ACC} = \frac{TP+TN}{p+n} = \frac{TP+TN}{TP+TN+FP+FN} \]  

(7)

3.1.2 Pruning

One of the most important techniques in machine learning is a pruning which used to remove a section of the tree to reduce the size of decision trees. A tree with a lot of branching is a big problem so we need to reach to the optimal
size by removing the nodes that do not provide important information (12).

3.2 Linear Regression

Regression used as statistical analysis tools (13). To explain the link between a dependent and independent variables we have to use regression analysis, taking into consideration that variable based on a sample from a given community (14). Regression model can be written as.

\[ y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_n x_n + \varepsilon_i \]  (8)

Where,
- \( \beta_0 \): Intercept.
- \( \beta_k \): Coefficient.
- \( K \): Independent variable.
- \( \varepsilon_i \): Error term.

3.3 SMO

SMO solves the SVM QP problem by dividing it into QP sub-problems then the smallest one, including two Lagrange multipliers (15). As shown in the figure 2 in the linear case, the margin is defined by the distance of the hyper plane to the nearest of the positive and negative examples. The output of a linear SVM written as:

\[ U = \hat{w} \cdot \hat{x} - b \]  (9)

Where,
- \( \hat{w} \): The normal vector.
- \( \hat{x} \): The input vector.

The separating hyper plane is the plane \( u = 0 \). The nearest points lie on the planes \( u = \pm 1 \). The margin \( M \) is thus

\[ M = \frac{1}{\|w\|_2} \]  (10)

3.4 M5-Rules Algorithm

M5 algorithm is one of the common methods for generating rules from the trees. The M5 builds regression trees whose leaves are consist of multivariate linear models, and the nodes of the tree are chosen over the attribute that maximizes the expected error reduction as a function of the standard deviation of output parameter (16). The association rule and the classification rule are the only two rule algorithm type (17). The important of using Rule algorithms in data mining application: It offer simple and clear results, active in undirected data mining, deal with huge amount of data, using a simple computation. Create M5 tree in three steps: generates a regression tree using the training data, and calculates a linear model for each node of the tree generated, tries to simplify the regression tree deleting the nodes of the linear models whose attributes do not increase the error and reduces the size of the tree without reducing the accuracy.

IV. SIMULATION RESULTS

Classification has been performed using C4.ss (Sequential Minimal Optimization), Linear Regression and M5-Rules Algorithm m on mathematical dataset in weka tool.

4.1 Results for Classification using J48

J48 is applied on the data set and the confusion matrix is generated for class gender having two possible values i.e. PASS or FAIL. Confusion Matrix

\[ \begin{array}{cc}
\alpha & \beta \\
102 & 28 \\
250 & 15 \\
\end{array} \]

For above confusion matrix, TP for class \( \alpha = 'Fail' \) is 102 while FP is 28 whereas, for class \( \beta = Pass \), TP is 250 and FP is 15 i.e. diagonal elements of matrix 102 + 250 = 352 (the correct instances) and other elements 15 + 28 = 43 (the incorrect instance).

4.2 Results for Classification using SVM

SVM is applied on the data set and the confusion matrix is generated for class gender having two possible values i.e. PASS or FAIL. Confusion Matrix

\[ \begin{array}{cc}
\alpha & \beta \\
106 & 24 \\
19 & 246 \\
\end{array} \]

For above confusion matrix, TP for class \( \alpha = 'Fail' \) is 106 while FP is 24 whereas, for class \( \beta = Pass \), TP is 246 and FP is 19 i.e. diagonal elements of matrix 106 + 246 = 352 (the correct instances) and other elements 24 + 19 = 33 (the incorrect instances).

Table 2 shows the results of J48 and SVM, we notice that both have the same high accuracy which equal to 0.981.

4.3 Results for Classification using Linear Regression

To apply linear Regression, preprocessing must apply to some variables (convert nominal variables to numerical variables) (18). Linear
Regression will be applied to the available dataset on three ways:
1. All variables in dataset (G3N is the output).
2. All variables in dataset except G3N, G2N (G1N is the output).
3. All variables in dataset except G3N (G2N is the output).

Regression Results:
1. Linear Regression Model
\[ \text{G3N} = -0.53 \text{ school} + 0.2568 \times \text{age} + 0.5391 \times \text{Fjob} - 0.2845 \times \text{activities} + 0.3167 \times \text{romantic} + 0.4022 \times \text{famrel} + 0.1355 \times \text{Walc} + 0.0474 \times \text{absences} + 0.1687 \times \text{G1N} + 0.9718 \times \text{G2N} + 0.7893 \]

2. Linear Regression Model
\[ \text{G1N} = 0.8726 \times \text{sex} + 0.2222 \times \text{F edu} + 0.5466 \times \text{M job} + 0.8306 \times \text{Mjob} - 1.2776 \times \text{Mjobth} + 1.8858 \times \text{Mjob} + 1.0861 \times \text{Fjob} + 0.9411 \times \text{Fjob} + 0.6737 \times \text{study time} - 1.2542 \times \text{failures} + 2.0633 \times \text{school sup} + 0.9542 \times \text{famsup} + 1.2838 \times \text{higher} + 0.2502 \times \text{free time} - 0.4451 \times 57 \text{out} - 0.216 \times \text{health} + 6.1228 \]

3. Linear Regression Model
\[ \text{G2N} = 0.8958 \times \text{sex} - 0.2264 \times \text{age} + 0.687 \times \text{famsize} + 0.3074 \times \text{Medu} + 0.9718 \times \text{Mjob} - 1.5888 \times \text{Mjob} + 2.2183 \times \text{Mjob} + 1.4052 \times \text{Fjob} - 0.9341 \times \text{guardian} - 0.4422 \times \text{travel time} + 0.5565 \times \text{school time} - 1.4591 \times \text{failures} + 1.4613 \times \text{school sup} + 0.8055 \times \text{famsup} + 0.6754 \times \text{internet} + 0.7634 \times \text{romantic} - 0.4796 \times \text{goout} - 0.2518 \times \text{health} + 13.2063 \]

Table 3 shows Correlation coefficient (CC), Mean absolute error (MAE), Root mean squared error (RMS), Relative absolute error (RA), Root relative squared error (RRS) and Total Number of Instances (TNI). We notice that RMS has a low values (2.0017 for G3N, 3.0089 for G1N and 3.6057 for G2N).

4. Results for Classification using M5 Rules
Rule: 1
IF \[ \text{G2N} > 10.5 \]
THEN

\[ \text{G3N} = 0.0439 \times \text{age} + 0.201 \times \text{Mjob} - 0.1484 \times \text{travel time} - 0.0236 \times \text{activities} + 0.022 \times \text{romantic} + 0.0256 \times \text{famrel} - 0.1054 \times \text{Walc} + 0.03 \times \text{absences} + 0.124 \times \text{G1N} + 1.041 \times \text{G2N} - 1.0993 \]

[203/15.592%] [1]

Rule: 2
[2] IF \[ \text{absences} > 1 \]
THEN \[ \text{G2N} > 7.5 \]

\[ \text{G3N} = -0.038 \times \text{age} - 0.2381 \times \text{famsize} + 0.046 \times \text{Pstatus} - 0.0259 \times \text{Medu} - 1.8848 \times \text{Fedu} + 0.039 \times \text{Mjob} + 0.039 \times \text{Mjob} - 0.0618 \times \text{Fjob} + 0.0326 \times \text{reason} + 0.0461 \times \text{schoolsup} + 0.0745 \times \text{romantic} + 0.0912 \times \text{famrel} + 0.0314 \times \text{Walc} + 0.0074 \times \text{absences} + 0.199 \times \text{G1N} + 0.8447 \times \text{G2N} + 0.389 \]

[97/20.84%]

Rule: 3
IF \[ \text{absences} <= 1 \]
THEN \[ \text{G2N} > 6.5 \]

\[ \text{G3N} = -1.4169 \times \text{age} - 1.3755 \times \text{reason} - 0.3079 \times \text{activities} + 0.0343 \times \text{absences} + 1.1151 \times \text{G2N} + 20.2863 \]

[38/87.272%]

Rule: 4
IF \[ \text{absences} > 1 \]
THEN \[ \text{G3N} = \]

\[ 0.3999 \times \text{age} - 0.2253 \times \text{Medu} + 0.6979 \times \text{Mjob} - 0.2776 \times \text{failures} + 0.0455 \times \text{absences} + 0.7721 \times \text{G2N} - 5.2655 \]

[34/21.977%]

Rule: 5
\[ \text{G3N} = \]

\[ + 0 \]

[23]

It is noticed that RMS has a low values (1.4621).
Table 1

<table>
<thead>
<tr>
<th>The preprocessed student related variables</th>
</tr>
</thead>
<tbody>
<tr>
<td>Data type : Attribute</td>
</tr>
<tr>
<td>Gender of student : G1 S</td>
</tr>
<tr>
<td>age of student : G1 N</td>
</tr>
<tr>
<td>School of student: G1 S</td>
</tr>
<tr>
<td>student’s address : G1 S</td>
</tr>
<tr>
<td>parent: G1 S</td>
</tr>
<tr>
<td>Education of mother : G1 N</td>
</tr>
<tr>
<td>job of mother : G1 S</td>
</tr>
<tr>
<td>Education of father : G1 N</td>
</tr>
<tr>
<td>job of father: G1 S</td>
</tr>
<tr>
<td>Guardian : G1 S</td>
</tr>
<tr>
<td>Size of family : G1 N</td>
</tr>
<tr>
<td>Relationships family : G1 N</td>
</tr>
<tr>
<td>Reason to attend this school : G1 S</td>
</tr>
<tr>
<td>Travel time : G1 N</td>
</tr>
<tr>
<td>Studying time for week: G1 N</td>
</tr>
<tr>
<td>Past failures No : G1 N</td>
</tr>
<tr>
<td>Other educational support : G1 S</td>
</tr>
<tr>
<td>Family support : G1 S</td>
</tr>
<tr>
<td>Activities : G1 S</td>
</tr>
<tr>
<td>Extra classes : G1 S</td>
</tr>
<tr>
<td>Internet available: G1 S</td>
</tr>
<tr>
<td>Nursery: G1 S</td>
</tr>
<tr>
<td>Looking forward to joining higher education : G1 S</td>
</tr>
<tr>
<td>In relationship : G1 S</td>
</tr>
<tr>
<td>Free time: G1 N</td>
</tr>
<tr>
<td>Going out after school : G1 N</td>
</tr>
<tr>
<td>Alcohol: G1 N</td>
</tr>
<tr>
<td>Health: G1 N</td>
</tr>
<tr>
<td>Absences : G1 N</td>
</tr>
<tr>
<td>G1N : G1 S ( C4.s )</td>
</tr>
<tr>
<td>G1N : G1 N (M5Rules and Linear Regression)</td>
</tr>
<tr>
<td>G2N : G1 S ( C4.s )</td>
</tr>
<tr>
<td>G2N : G1 N (M5Rules and Linear Regression)</td>
</tr>
<tr>
<td>G3N : G1 S ( C4.s )</td>
</tr>
<tr>
<td>G3N : G1 N (M5Rules and Linear Regression)</td>
</tr>
</tbody>
</table>

Table 2

| J48 and SVM results |

Table 3

| Linear Regression results (Cross-validation) |

Table 4

| M5Rules (Cross-validation) |

CONCLUSION

It has been presented some ways to predict grades of students in Mathematics course for secondary schools based on previous student grades in first or/and second year and other attributes. The presented model has been tested by using four data mining methods (SVM, Decision tree M5-Rules and Linear Regression). Simulation results have proven that a high predictive accuracy can be achieved when applying Decision tree and SVM Techniques. While low Root mean squared error is obtained when applying M5-Rules and Linear Regression Techniques.
REFERENCES


47. Ahmed A. Saleh, Hazem M. El-Bakry, Mahmud Atta Alla and Nikos


