PREDICTING STUDENT PERFORMANCE IN FOREIGN LANGUAGES WITH A SERIOUS GAME

A. Illanas Vila, J.R. Calvo-Ferrer, F. Gallego Durán, F. Llorens Largo

University of Alicante (SPAIN)
aillanas@ua.es, jr.calvo@ua.es, fgallego@dccia.ua.es, faraon.llorens@ua.es

Abstract

In this digital age, many statements have been made regarding the use of technology for teaching purposes. In this sense Serious Games are gaining ground considering that, besides their technological advantages, they provide fun, which allegedly engages students in their training.

Much research has been carried out to show how Serious Games improve teaching methodologies and student learning outcomes in various subjects. This research focuses on the field of digital game-based learning from a different perspective: Namely, the work carried out does not focus on the use of Serious Games for teaching and learning, but on the use of such tools for the prediction of learning outcomes. Accurately predicting future student performance lets teachers give customized advice to them.

The approach is undertaken by means of machine learning and data mining techniques, and educational data mining techniques in particular. These techniques are applied to data collected from games played by students. For such purposes, The Conference Interpreter (CoIn), a Serious Game which simulates a context of simultaneous interpreting has been developed and used as a data mining tool. Following this, the experiment carried out is described and machine learning/data mining results are presented and discussed.

Keywords: serious games, machine learning, data mining, student performance, foreign languages.

1 INTRODUCTION

Technology is changing our lives. We use technology from the moment we wake up until we go back to bed at night, sometimes without noticing it. This applies even more often to children than to adults. This generation of individuals, widely know as digital natives, have supposedly developed a different set of skills that have a particular impact on learning [1][2]. We assume students are becoming more technology-friendly, although according to research [3][4], traditional teaching has become a disengaging activity and teachers lack basic technology skills.

In an attempt to bring students and teachers closer in such daunting scenario, computer games for educational purposes have gained popularity in the educational process. It is stated that computer games favour skills such as mental agility, reflexes and visual capabilities [3][4]. It is also said that multi-player computer games help develop competition skills and favour teamwork [5]. In addition, they also exercise the coordination of the body thanks to the new peripherals available in the market [6][7]. But most importantly, computer games are funny. Using computer games for educational purposes implies that the student can learn contents and skills and enjoy while doing so [8]. Although games have often been criticised regarding violence and addiction issues, the truth is that the immersive context they provide turns them into an active learning method where players can learn contents and skills by themselves [9]. As proof of this, digital game-based learning has gained enormous acceptance in research in recent years and Serious (and commercial) Games have become widely accepted tools for teaching and learning [10].

Nonetheless, this research wishes to be more than further research into the benefits of Serious Games in certain skills or into its learning outcomes. Thus, it uses the results obtained and the data gathered from the serious game CoIn to predict the learning outcomes of the players with regards to a different set of contents. This is, the purpose of the study is to use CoIn not to determine its adequacy as a tool for teaching and learning the contents thereof, but as a means of predicting student behaviour in subjects not strictly related to the contents of the video game.
Predicting students’ final grades while the learning process is taking place may prove very useful: teachers can analyse the progress of each student any given time in the learning process in order to adapt teaching contents and skills to the each student according to the prediction obtained.

As already stated, this research aims to predict learning outcomes and grades to individually design learning contents for students. Thus, fifty-seven students from the degree in Translation and Interpreting from the University of Alicante were invited to play CoIn, and the data they produced were automatically gathered following existing research that proves such video game as an effective tool for terminology acquisition in second language learning [11].

Consequently, such data were compared with the different grades in the subject Lengua B(III): Inglés obtained by the students that played CoIn, aiming to establish correspondences between them by means of data mining techniques. Such techniques are used to train an algorithm, which is reasonably able to predict students’ final grades in the subject.

This paper explains the different stages of the research carried out: how the data were gathered, what data mining techniques were used and the results obtained in the experiment. Finally, conclusions and future work is also brought up.

2 DATA COLLECTION

As already said, the object of the investigation is to use data mining techniques to predict final grades in the subject Lengua B(III): Inglés of students playing CoIn, using the data the game provides. In order to train an algorithm, real game data and real final marks of real students are needed.

2.1 CoIn. A serious game for simultaneous interpreting

CoIn is a conceptual mini-game [12] devised for the training of simultaneous interpreters. In CoIn the player is represented by a conference interpreter who needs to simultaneously interpret various speeches. The screen simulates a room where a speaker is giving a speech in English. At the bottom of the screen, the Spanish translation is shown simultaneously as the speaker intervenes. However, the Spanish translation is not complete. Some words are replaced by an empty gap and the task of the player is to choose one of the four options that the game offers for every single missing word. A screenshot of the game is provided in Figure 1 below.
If an empty gap disappears without any answer being selected by the player, the game moves on to the next missing word. The player can seek help using a limited number of power-ups during the game, such as a 50% option which eliminates two wrong options.

The player must also keep his level of stamina up during the game. There is an energy bar that rises if the correct answer is chosen but decreases if the selected answer is wrong or the word is ignored. If the player loses all of his stamina, the game is then over, although it can be restarted from the beginning of the level that was being played.

2.2 Collecting data from students

There are two determining factors that have made this research possible. Firstly, we have used a serious game developed by ourselves. That means that we know very well the way it works and it could have been adjusted if we had needed to. Secondly, the CoIn game has been used in a real context at the University of Alicante, by students whose subject grades were also made available for the purposes of this research.

Thus, a game session was organised, where 57 students from the degree in Translation and Interpreting played CoIn for two hours. All of the input data generated were collected at the end of the session to be confronted with the Lengua B(III): Inglés subject grades.

2.3 Data collected

As said before, data from 57 students of English were collected. Such data reflect two levels of information about how the user played the game:

- **General data about the level played.**

A register is recorder every time the user plays a level of the game. The record of each level has information as: date, starting hour, duration of the game, result (whether the level is successfully completed or not) final score and percentage of life at the end of the game, total of empty words in the level, number of right, wrong and ignored words, the time the game was paused, and the type of power-up (if any) used. Figure 2 provides an example of these data.

<table>
<thead>
<tr>
<th>partida</th>
<th>nivel</th>
<th>pantalla</th>
<th>fecha_inicio</th>
<th>hora_inicio</th>
<th>duracion</th>
<th>resultado</th>
<th>puntuacion</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>transcript_apple_1_1</td>
<td>speech/apple_1.xml</td>
<td>29/11/2012</td>
<td>19:25:21</td>
<td>39</td>
<td>abandono</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>transcript_apple_1_1</td>
<td>speech/apple_1.xml</td>
<td>29/11/2012</td>
<td>19:27:34</td>
<td>24</td>
<td>abandono</td>
<td>0</td>
</tr>
<tr>
<td>3</td>
<td>transcript_apple_1_1</td>
<td>speech/apple_1.xml</td>
<td>29/11/2012</td>
<td>19:28:13</td>
<td>68</td>
<td>completado</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>transcript_apple_1_2</td>
<td>speech/apple_1.xml</td>
<td>29/11/2012</td>
<td>19:29:43</td>
<td>65</td>
<td>completado</td>
<td>75</td>
</tr>
</tbody>
</table>

![Fig. 2. General data table about the level.](image)

- **Detailed data about the player actions of each level.**

On each level, the player must choose the word that he or she believes is right. For every empty word, the game records the action the player performed, namely: the correct word and the text the player answered, instant of time when the empty word appeared on the screen and instant when the player
selected an answer for it, if it was a hit or a mistake and if power-ups were used. Figure 3 provides an example of these data.

<table>
<thead>
<tr>
<th>note_id</th>
<th>partida</th>
<th>speech</th>
<th>tiempo_aparicion</th>
<th>tiempoRespuesta</th>
<th>estado</th>
<th>metodo</th>
<th>textoRespuesta</th>
<th>powerup</th>
</tr>
</thead>
<tbody>
<tr>
<td>iPad</td>
<td>1</td>
<td>speech/apple_1.xml</td>
<td>97</td>
<td>17.385</td>
<td>falla</td>
<td>raton</td>
<td>iPod</td>
<td>nil</td>
</tr>
<tr>
<td>aprendido</td>
<td>1</td>
<td>speech/apple_1.xml</td>
<td>19.063</td>
<td>37.483</td>
<td>falla</td>
<td>raton</td>
<td>retocado</td>
<td>nil</td>
</tr>
<tr>
<td>iPad</td>
<td>2</td>
<td>speech/apple_1.xml</td>
<td>95</td>
<td>11.003</td>
<td>falla</td>
<td>raton</td>
<td>iPod</td>
<td>nil</td>
</tr>
<tr>
<td>segunda</td>
<td>3</td>
<td>speech/apple_1.xml</td>
<td>10.02</td>
<td>12.86</td>
<td>acierta</td>
<td>raton</td>
<td>segunda</td>
<td>nil</td>
</tr>
</tbody>
</table>

Fig.3. Detailed data table about the actions of each level.

### 3 DATA ANALYSIS METHODOLOGY

For this experiment we fed 5 Neural Networks [13], which were used to predict future marks of students in 5 final exercises (written, oral, blog, quiz and average). Each Neural Network needed to receive a constant number of inputs (usually called features) describing each student being processed. Therefore, starting from the raw data obtained from the game, we selected and structured data in an appropriate way to feed the Neural Networks.

We considered the data about each gap being correctly filled or not and how, to be more relevant for predicting final marks, as it resembles more to the actual exercises of the final exam. So, for each student we created a record with all the gaps they found while playing and how did they answer to these gaps. It is important to notice that, in CoIn, gaps can only appear in the place of predetermined words that the designer of the level established previously. In our experiment, the set of different predetermined words contains 556 units. This implies that, for each player and each possible word to appear in gaps, we feed the Neural Network with the information shown in figure 4.

Moreover, we also added a similar record to figure 4, but with global frequency and success/failure information of all the actual gaps the user found while playing. This means that, for each student we set up \(557 \times 7 = 3899\) information fields about the possible gaps they could have found while playing plus the actual data about the total number of found gaps and success/failure frequencies and time lapses.

<table>
<thead>
<tr>
<th>Frequency</th>
<th>Total number of times that a gap appeared instead of this word</th>
</tr>
</thead>
<tbody>
<tr>
<td>Success Pct</td>
<td>Percentage (0-1) of times the student correctly filled in the gap for this word</td>
</tr>
<tr>
<td>Ignoring Pct</td>
<td>Percentage (0-1) of times the student ignored the gap for this word</td>
</tr>
<tr>
<td>Failure Pct</td>
<td>Percentage (0-1) of times the student wrongly filled in the gap for this word</td>
</tr>
<tr>
<td>Avg. time lapse (ATL)</td>
<td>Average time lapse (secs) that the user took to fill in the gap for this word</td>
</tr>
<tr>
<td>ATL Success</td>
<td>ATL for correct answers only</td>
</tr>
<tr>
<td>ATL Failure</td>
<td>ATL for wrong answers only</td>
</tr>
</tbody>
</table>

Fig.4. Data to feed Neural Network for each student and each possible word to appear in a gap.

For this data, we designed 5 Feedforward Neural Networks [14], each one consisting in 3899 input neurons, 40 hidden neurons in 1 layer and 101 output neurons. Each layer of neurons is fully
connected with the next one. This design represents an hypothesis function $h_{\theta}(x)$ that accepts a 3899-feature-values vector $x^{(i)}$ for each student $i$ as input, and produces a 101-values output vector $h_{\theta}(x^{(i)}) \in [0, 1]^K$, $K=101$. Each component of this output vector $(h_{\theta}(x^{(i)}))_k$, $k \in \{0, ..., 100\}$ represents the predicted probability of the student being considered to get a mark of $k\%$ in the final exercise being predicted. With this design, we use a standard cost function $J(\theta)$ to minimize (figure 5), based on the Sigmoid Activation Function [15][16] for modelling neuron behaviour.

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{k=1}^{K} y^{(i)}_k \log(h_{\theta}(x^{(i)}))_k + (1 - y^{(i)}_k) \log(1 - (h_{\theta}(x^{(i)}))_k) + \frac{\lambda}{2m} \sum_{l=1}^{L-1} \sum_{i=1}^{n_l} \sum_{j=1}^{s_{l+1}} (\theta^{(l)}_j)^2$$

Fig. 5. Cost function $J(\theta)$ to be minimized in order to train our neural networks.

The function $J(\theta)$ measures the error (or cost) that the Neural Network incurs in when trying to predict the final marks $y$ of the $m$ students being used to train the neural network, given a matrix of parameters $\theta$ and a regularization parameter $\lambda$, and taking into account that the neural net has $L$ layers with $s_l$ neurons each one.

To get a very approximated predictive model, what we had to do is to find a set of parameters $\theta$ and a regularization parameter $\lambda$ that minimized the prediction error $J(\theta)$. With appropriate values for these $\theta$ and $\lambda$, our hypothesis function $h_{\theta}(x^{(i)})$ becomes a function that outputs a very approximated probability distribution for a student $i$ over the possible marks they will get in the final exercise being studied.

In the field of Neural Networks, the process finding a good set of parameters $\theta$ and $\lambda$ is known as **learning** and running the algorithm for iteratively approximating these values is known as **training** the neural network. Therefore, in order to train our neural networks, we split our sample set of 57 students into two mutually exclusive subsets called Training Set ($Tr$) and Test Set ($Te$). We assigned 47 random student samples to $Tr$ and the rest 10 students samples to $Te$.

We used the Training Set $Tr$ to iteratively train our 5 neural networks during 10000 iterations of the standard Backpropagation [17] algorithm, using Conjugate Gradient Method [18][19] for minimization. After this iterative training, we found 5 different matrices $\theta$ which converted our 5 neural networks into very accurate predictors of the final grades of our 47 students in the $Tr$ set (see results section).

However, in general, there are infinite hypothesis functions able to accurately predict a finite set of samples like this. We were interested in finding one of them able not only to accurately predict our samples in $Tr$, but also to be general enough to accurately predict our samples in $Te$, which are samples that the Neural Network had never been presented with. This means that samples in $Te$ were completely new to our Neural Network, just as if they were students from a new year and we were trying to predict their final results before knowing them (as it is our final purpose). Therefore, the prediction error of each neural network under the $Te$ set is a good starting measure of how well the predictions of our neural networks could generalize to new students.

Finally, with the aim of getting a set of parameters $\theta$ that made the predictions of our neural networks generalize well to new, previously-unseen students, we launched a battery of 1000 training experiments with different values of $\lambda$ and different number of neurons in the hidden layer, thus obtaining the final results we present in the next section.

## 4 RESULTS

This section presents the results got from the experiments undertaken. For available space reasons, only 2 samples of the results are shown in graphs. However, these two graphs are highly representative of the results got and we serve them with the final numeric results.

Figure 6 is shown to prove that our training algorithms were working as expected and the neural networks were indeed learning to accurately predict samples in $Tr$. The graph show the error of the neural network learning to predict the average final results of the students, as a function of the number of iterations. The graph presents only the first 1000 iterations for clarity, but the error continues to drop at the expected easy-to-visualy-infer rate. This behaviour is the standard expected for a well designed neural network and confirms that our neural network were appropriately learning the relation between
the inputs (performance of students under CoIn) and the outputs (prediction of the final mark students are expected to get in the final exercises on average).

Once it was stated that our neural networks were performing well at the task of learning to predict accurate outputs for inputs from the Tr set, the next step was to verify how well they performed in the task of predicting marks for student samples never seen during training. Figure 7 shows the results of our best trained neural network for this task. Each pair of bars in the figure represents an actual grade (from 0 to 10) a sample student got as average of the final exercises they did (left bar), versus the prediction our neural network thinks they should have got (right bar). The graph shows 10 pairs of bars, each one for a sample student in the Te set.

Finally, figure 8 presents the numeric results of the 5 most relevant results of the experiments undertaken for our 5 neural networks. The Tr prediction accuracy percent represents the percentage of accurate predictions each neural network did after 10000 of training, with samples coming from the Tr set. It is important to notice that getting 100% on all of them is possible (and actually some experiments did get 100%) but that does not guarantee a good ability to generalize and a good performance under the Te set, which is what we really want. Te prediction mean error is the result of calculating the mean of the prediction errors the neural network performed when presented with all the elements in the Te set. This means that, for instance, the neural network that predicted blog grades outputted prediction grades that were, in mean, the actual grade the student finally got ± 4%.

The global accuracy percent represents the percentage of accurately predicted student grades for all students, either they were in Tr or Te.
5 CONCLUSIONS

In this paper we have presented our Serious Game CoIn for learning foreign languages. Along with CoIn we have presented an experiment we have undertaken to assess the viability of using students results playing CoIn as a source to predict their expected final grades in an English subject.

Overall, the results of the experiment are quite encouraging and interesting. From the results we infer that our approach is very promising and following this research line can lead to develop powerful tools to help teachers greatly improve their lessons. The goal would be to develop CoIn and a neural network powered prediction system to let teachers early predict student deficiencies and appropriately guide them to surpass these deficiencies and success in foreign language subjects. By means of knowing accurate prediction of the expected grades a student is to obtain in future final exercises, the teacher could better understand these student deficiencies and point them towards useful ways to overcome them and efficiently improve their skills.
It is important to state that we consider these results to be promising but by no means definitive. Taking into account that these results have been obtained with a sample set of just 57 students, and that we have been tweaking the learning system to get a good generalization for a test set of just 10 students, we are conscious that our results could probably be biased. Our future experiments will concentrate mainly on getting a sample set big enough for stating deeper conclusions. Moreover, we will try different learning algorithms and configurations that we expect can lead us to get really useful and directly applicable results.

REFERENCES