Prediction Model Based on User Profile and Partial Course Progress for a Digital Media Learning Environment

Arturo Fernandez Espinosa Meaghen Regts Jayshiro Tashiro Miguel Vargas Martin
University of Ontario Institute of Technology
{arturo.fernandez, meaghen.regts, jay.tashiro, miguel.vargasmartin}@uoit.ca
Oshawa, Canada

***Abstract*—This work in progress reports on the implementation of a data mining system aimed at predicting the final GPA of healthcare students within the context of our learning environment called IPSims. The predictions are based on the user profile, and their choices and preferences within IPSims. The intended predictions will use various well-known and widely used IT technologies and data mining techniques combined in such a way to overcome the complexity of IPSims and its associated disaggregation variables. The present work shows the application of well-known data mining techniques within the environment called IPSims.**

Keywords—Databases; Data Mining; Education, Knowledge Discovery; Digital Media Learning Environment.

I. INTRODUCTION

Knowledge discovery is an iterative process that involves different stages of information processing. Fu [1] and Luo [2] divide the knowledge discovery process into five stages: selection, preprocessing, transformation, data mining, and evaluation. The final result of applying these stages should be the acquisition of knowledge.

The purpose of this work is to test the following hypotheses:

1) There is an identifiable set of variables in the student profiles which have the highest impact in students' success, 2) It is possible to predict, with high level of accuracy, the performance of a student at the beginning of a course, based on their profile, and 3) It is possible to predict student performance based on their profile and their history of choices and preferences in the system.

The prediction of the student performances (usually reflected by their grades) may help teaching environments of all levels. Since elementary school to postsecondary education, the detection of students with poor performance is not possible until the final grade is known and it is already too late to take a preventative action.

The main problem of this research is to apply well-known data mining techniques in an efficient way in order to obtain useful information from IPSims [3], which is a digital media-learning environment for health sciences students.

In order to apply data mining techniques, there are different issues to address such as the normalization of a non-normalized database, inconsistency in some student records (registers of the database), the transformation of multiple qualitative variables into quantitative values, lost information, data integration from multiple data sources, outliers for some variable values, the efficiency of the prediction algorithms and cluster techniques, and a reduced number of real cases to train and test the system.

The data integration, data reduction and data transformation steps have been already accomplished. The statistical analysis of the variables in order to detect outliers and understand the behaviour of the information is the next step on this work as well the implementation of the neural network for prediction purposes and cluster techniques for pattern recognition that will surface the variables with greatest impact in the final grade of the students.

The rest of the paper is organized as follows. Section II is the related literature. A description of IPSims is given in Section III. In Section IV, the techniques applied in order to retrieve knowledge from IPSims are reviewed. Section V presents the concluding remarks, final notes and future work.

II. RELATED WORK

The idea of using neural networks as predictive models has been proposed by multiple authors. Bargallo [4] makes a study of comparison for different neural network architectures for prediction of radio frequency propagation loss this methodology. The architectures proposed are: Multilayer perceptron networks and radial basis networks. In Wang and Yu [5], once again, neural networks are being use to predict. On this specific case the prediction is over software quality. Vilovic and Burum [6] present a comparison study of different neural network models and the accuracy of them on the prediction of indoor field strength.

The previous works lacks a description of one characteristic that is being analyzed in this paper; this is the measure of accuracy and model optimization over time periods. The previous group of papers applied the model that fits best in general for a given phenomena.

The current work describes the specific study case in function of discretized units of time. The phenomena are divided and the model is adequate accordingly in function of this discretize time units. A more detailed description will be given later in the paper.

The related literature shows a common denominator: Neural networks are black boxes. This makes it difficult to know if the results are going to be good enough to support our hypotheses.

The way these papers evaluate the efficiency of each neural network is not known until the testing in real cases is finished.

 ISBN: 978-1-61208-185-4

III. IPSIMS DESCRIPTION

A. General Description

IPSims system is designed to help students enhance their learning experience through the effective use of a simulation environment supported by multimedia materials.

IPSims is able to track the activity of users. The traces include time spent on each module or learning activity as well as the path followed within IPSims.

First-time users get registered in the system as they enter a number of disaggregation variables, which include: gender, course, faculty, term, age, undergraduate academic year, undergraduate major, number of hours/week surfing the web, number of hours/week playing video games, how they rate their computer literacy, likelihood to choose a career in computer sciences, interest in course material, experience with computer-based simulations, their perceived educational value of computer-based simulations, expected grade in the course, and current GPA.

These variables are stored in a secure database that is accessed at a later time and used in the knowledge discovery process.

Returning users logon with a user name and password. Upon successfully logging in, IPSims presents a menu of six simulation environments, each consisting of a different healthcare scenario and involving multiple learning activities.

Each of these six simulations contains three scenarios, a library with web links and scholarly journals, scopes of practice, interprofessional competencies, interprofessional perspectives, case records, case encounter (video), main menu, logout, and bookmark links.

Figure 1. General architecture of IPSims and data recording into the database.

B. Architecture

The user interface of IPSims is constructed using FLEX technology [7]. FLEX offers an elegant and attractive design environment for user interfaces.

In general, the structure of the IPSims system is as follows, LSPL is a set of PHP scripts that are responsible for the interaction between the XML files and the database as these scripts record activities and perform data retrieval.

Figure 1 describes IPSims architecture in a simplified form. LSPL counts with a query module where we can consult and extract the information related to time stamps and path tracking (but by now and for our research purposes this paper will focus just in the recording functionality of the system).

IV. KNOWLEDGE DISCOVERY IN THE IPSIMS DATABASE

To fully exploit the potential of the IPSims database it is necessary to create appropriate mechanisms that allow for information retrieval.

The database contains data from three sources: 1) the user profile, 2) users' choices and preferences in IPSims, and 3) student surveys collected after completion of all the learning activities performed in IPSims. The student surveys include nine different sets of questions, namely 1) learning assessment based on assigned learning activities, 2) demographic information, 3) preferences for learning resources and educational scaffolding, 4) rating of web-based course work, 5) rating of IPSims learning environment, 6) satisfaction with educational simulations and serious games, 7) disposition to engage in effortful cognitive endeavor (need for cognition and ambiguity tolerance), 8) expectancy-value questionnaire, and 9) performance evaluation in interprofessional learning activities.

Before the system can exploit these three sources of data, algorithms that perform preprocessing and transformation steps are required; as well as a neural network model will be applied in order to predict the final GPA of the students.

A. Data Preprocessing and Transformation

It is convenient to firstly analyze the user profile. The user profile is automatically recorded in our database, and the validation of the web registration form is carefully designed to avoid inconsistencies in the data provided by the students.

The qualitative variables in the user profiles require a transformation into quantitative values. For our approach it is desirable to normalize our input set of values into the range [0-1].

Referring back to the question of how to treat the qualitative variables, Driscol et al. [8] propose multiple methods that could be used. One of the most common methods is to count the number of times a qualitative code occurs for each user. It will be interesting to see how the two-three different techniques mentioned in [8] like substitute the values for descendent or ascendant values, or frequency of each value repetition will impact the model accuracy.

The next step to achieve a high-quality input set is the treatment of the timestamps and path tracking for the users. The timestamp for each section has to have a special treatment so that every time the user starts a session the system starts recording time stamps for each web document visited for the user. For example, in a situation where a given user u_x that visits the web document d_a n times, then we will have n time stamps for that given section of the system. The obvious solution in this situation is to merge these times into one total time T for the document d_a given a user u_x , denoted by $T(u_x, d_a)$, the relative times to each session for a given user u_x at a document d_a is given by the expression $t_i^i \mid i \in Z^+$,where i denotes the number of session.

$$T(u_x, d_a) = t_0(u_x, d_a) + t_1(u_x, d_a) + t_2(u_x, d_a) + \cdots + t_n(u_x, d_a); \quad (1)$$

Therefore, at the end, there will be just one variable for the time stamp for the web document d_a that will prevent the fast growing of information and will fix the input vector in a specific dimension.

Imagine a situation where a user X visits seven web documents and user Y visits just five, then there will be a dimensionality inconsistency on the vectors, even worse if the documents they visited are partially or completely different. The first solution that comes to mind is to set up the input vector to a defined dimensionality, and fill these n-elements with zeros, and for every available variable fill the element in the vector with the corresponding variable.

Supposing that u_x and u_y define different users, and $d_i; i = 0,1, \ldots n;$ define a visited document. Then the vector for each user will be defined by V(u_x) and V(u_y) and will be composed as follows:

$$V(u_x) = \begin{bmatrix} User\ profile\ var1 \\ User\ profile\ var\ 2 \\ \vdots \\ User\ profile\ varn \\ T(u_x, d_0) \\ T(u_x, d_1) \\ T(u_x, d_2) \\ \vdots \\ T(u_x, d_n) \end{bmatrix}$$

$$V(u_y) = \begin{bmatrix} User\ profile\ var1 \\ User\ profile\ var2 \\ \vdots \\ User\ profile\ varn \\ T(u_y, d_0) = 0; \\ T(u_y, d_1) \\ T(u_y, d_2) \\ \vdots \\ T(u_y, d_n) \end{bmatrix}$$

$T(u_y, d_0) = 0$ indicates that the user u_y has not visited that document yet.

The path tracking functionality of IPSims is being wasted in this approach, but by the moment and for prediction purposes the model will be based just in the time stamps, user profile and paper survey information.

The last step in our input data preprocessing requires the capture of all the variables in the paper survey; the algorithms applied for preprocessing will be the same algorithms mentioned before for the qualitative, and the quantitative variables. This step will require the addition of new columns or tables to our database.

Figure 2 generalizes the process described before. We can observe how the integration of the three different data sources is accomplished into a one single data source.

B. Neural Network Model

The next step in this framework is the model construction. Different neural network architectures will be used in order to compare different results and determine which architecture is most suitable for the data set.

An optimistic approach is being adopted when the idea that a neural network with *n* inputs and *m* outputs will be enough to solve the linear separability problem for the different groups of students.

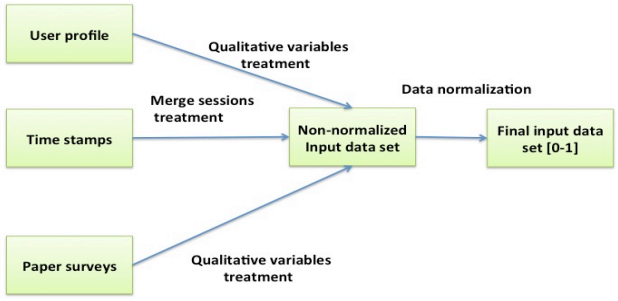

Figure 2. Data preprocessing and transformation processes before reaching the final input data set for the predictive model.

The justification for the selection of the model will be determined by the observations in the cluster analysis, the statistical analysis and distribution of the variables as well as the testing results. However, if the linear separability problem cannot be solved for the dataset with neural networks then the use of a support vector machine (SVM) will be required. This has been left for future work.

If the results with the neural network model are acceptable, then a comparison with a SVM model will be interesting in order to determine the tool with the best performance for the given study case.

Our IPSims already has around 70 real student registers with the variables and the final learning outcome. This set of cases will populate the training and testing set with data that can validate the accuracy of the results.

From 70 real cases available, around 75% are going to be used for training purposes and 25% for the testing stage of the model. This is usually the proportion for training and testing proposed in most of the literature.

The last stage in the framework is the refinement of the model as presented in Figure 3. This step will be accomplished with the comparison of results of the different proposed architectures (feed-forward propagation and backtracking) with different numbers of neurons in the input layer, and different activation functions between the neurons. In the experiments, a table with the results will be the one that allow determining, which one is the model with the highest accuracy.

 ISBN: 978-1-61208-185-4

TABLE I. A SAMPLE OF THE MODEL REFINEMENT.

No. of Neurons	Activation Function	Feed-forward propagation	Backtracking
W	Sigmoid	67.5%	78%
X	Sigmoid	70%	82%
Y	Step	45%	29%
Z	Signum	38%	34%

C. *Model application*

Once the selection of the final model is done, the system will attempt to predict the final learning outcome of IPSims users at any time during the course. The specific time in which the model is applied to the dataset will show a prediction for that specific time; therefore, still there is the question how accurate the system is when is applied at different given times? This is probably the most challenging question to solve in this work. Right now the idea to approach this problem is to divide the course into time periods, and according to these periods modify the weighted edges that go from the input vector to the first hidden layer in the neural network.

In the density function given by Eq. (2), an example of how to generate a dynamic weighted model for the neural network model is observable. This density function will depend on the time the model is applied; this time will be determined for discrete divisions of the given course, in the example shown the course is divided in three sections. Depending on which period of time the model is applied is the weight for every specific edge.

$$W_{xy}(t) = \begin{cases} 0 \leq W_{xy} \leq A \\ A < W_{xy} \leq B; \\ B < W_{xy} \leq \mathbf{1} \end{cases} \qquad (2)$$
$$\forall x, y \ni W \text{and } 0 \leq A, B \leq 1$$

V. CONCLUDING REMARKS

This work aimed at exploiting all the capabilities of the IPSims system. This continuing work is part of a larger project using the IPSims system for healthcare education. High impact variables identification (cluster analysis), predictive models (neural networks and support vector machine), behaviour prediction (neural networks and support vector machines), and uncertain data are part of the projects that will involve the use of IPSims. We expect to get acceptable results when measuring the accuracy of the model against real cases; however, we have as future work the use of a support vector machine. An ambitious approach on this work will be the generation of a framework that could be applied to virtual learning environments of similar nature to that of IPSims. In the future, we expect not only be able to predict outcomes, but to provide feedback and advice to users as to how they can improve their learning performance to maximize their results in a particular course. We are currently redesigning the IPSims database architecture to integrate the three data sources discussed in this paper, as well as the scripts required for the experiments related to data preprocessing and transformation.

AKNOWLEDGEMENTS

This work was supported in part by the Social Sciences and Humanities Research Council of Canada.

REFERENCES

[1] Y. Fu, “Data mining Tasks, techniques and applications,” 1997 19th International Conference on Data Engineering. IEEE Potentials, 18-20. Doi:10.1109/ICDE.2003.1260873

[2] Q. Luo, “Advancing Knowledge Discovery and Data Mining,” *2008* International Conference on Computer Science and Software Engineering, 7-9. doi:10.1109/WKDD.2008.153

[3] IPSims:http://199.212.33.78/LPSL_V2_040610/Main.html [Last accesed: December 17 2011].

[4] J. Bargallo, “A comparison of neural network models for prediction of RF propagation loss,” 48th IEEE Vehicular technology conference. Pathway to global wireless revolution, 445-449.

[5] Q. Wang and B. Yu, “Extract rules from software quality prediction model based on neural network. 16th IEEE International conference on tools with artificial intelligence. 191-195

[6] I. Vilovic and N. Burum, A comparison of neural network models for indoor field strength prediction. Elmar 2007. 251-254.

[7] Adobe Developer Connection. Flex Developer Center. Available: http://www.adobe.com/devnet/flex.html [Last accesed: December 12 2011].

[8] D. L. Driscoll, A. Appiah-Yeboah, P. Salib, and D. J. Rupert, “Merging qualitative and quantitative data in mixed methods research: How to and why not. Ecological and Environmental Anthropology,” (University of Georgia*)*, *3*(1), 18.

 ISBN: 978-1-61208-185-4