A Framework for Flexible Educational Data Mining

Kyle DeFreitas and Margaret Bernard,
Department of Computing and Information Technology
University of the West Indies, St Augustine Trinidad and Tobago

Abstract — Educational Data Mining (EDM) focuses on tools and techniques for discovering previously unknown patterns in the data generated by the educational process. Despite the increased awareness of data mining, the adoption among educators has been slow because the data mining tools are complex and the use of the tools requires a detail understanding of the parameters and requirements of the algorithms used. In this paper, we present a framework for developing an educational data mining environment that is flexible in that it caters for educators who have little technical skills as well as for more advanced users with some data mining expertise.

Keywords: Educational Data Mining, Learning Management Systems

1 Introduction

THE increased use of technology in education has brought about a fundamental change in the way educational institutions operate. The increased usage of electronic based systems has amplified the amount of data available for making better decisions, and improvements in the data mining algorithms make analysis of this volume of data easier and more accessible. In recent years, there has been tremendous interest and research in the field of Educational Data Mining [1].

The use of Learning Management Systems (LMS) are increasingly popular within e-learning environments [2]. The offerings vary from open source solutions such as the Modular Objective Oriented Development Learning Environment (Moodle) [3], commercial solutions such Blackboard [4] and software as a service systems such as Edmodo [5]. However despite their popularity and the universal desire for useful information, LMS are not designed to facilitate analysis using data mining techniques [6]. This compounded with the unique challenges of applying data mining techniques within the educational context has seen a low adoption of data mining among stakeholders [1],[7]. This paper proposes a framework that allows developers to build a data mining environment that is flexible to the level of knowledge and skill of the educator.

This paper is structured as follows. Section 2 briefly explains the challenge of adoption of data mining within educational environments and describes two approaches for user interface tools catering to technical and non-technical educators. In Section 3, the proposed framework is presented and clarified. Section 4 discusses an example of an implementation of the proposed framework while section 5 concludes the paper and provides suggestions for future works.

2 Flexibility in Educational Data Mining

Data mining is the extraction of relevant and useful information from a relatively large dataset [8]. Educational Data Mining (EDM) is the application of different techniques and algorithms with the focus on discovering unknown patterns in data generated by the educational process [9]. Learning Management Systems log all interaction of users (learners and educators) therefore every click and action performed by users is stored in the LMS database. This vast quantity of data becomes a 'gold mine' for extracting interesting and useful information, beyond the simple statistical reports that are provided by popular LMS. The process is not straightforward as EDM techniques must consider the unique pedagogical and semantic characteristics of the data that is extracted and stored [10].

Educators in particular are often required to have expert knowledge of Data Mining to set up parameters and configurations that can be used by the Data Mining process. They need to select the tables that they are interested in, apply the right filters and configurations for modeling the Data Mining tasks, understand enough of the DM techniques to be able to make a choice amongst methods as to which method is the most suitable in the context, and even how to interpret the results.

A number of researchers have tackled educational data mining from different directions. In [6], the authors propose a data model to structure and export usage data stored by the LMS. All EDM techniques must do some preprocessing of the raw log data from the LMS so this is an important step in simplifying the whole process. In [11], the authors apply a classification techniques to predict student performance. EPRules [12] provides a visual means of identifying associative rules within an Adaptive Hypermedia Architecture Course. TADA-ED [13] combines visualization and data mining to analyze web logs from we based courses. MultiStar [14] utilizes data warehousing and data mining resources for supporting distance learning courses within universities. GISMO [15] visualizes students usage data

extracted from Moodle which allows educators to assess where issues may exists and what topics may be potentially problematic. Moodle Mining Tool [16] allows researchers to apply data mining analysis such as clustering and associative rule mining with the usage data contained in the logs of the LMS and eLAT [17] which allows users to choose from a set of indicators related to the hypothesis they would like to test.

Each of the tools previously highlighted makes assumptions about the user's ability and competence with analysis. Therefore we considered the question, can the process at the user end by simplified, discoverable and explanatory?

This paper describes an EDM system, called FLexEDM that was developed as a plug-in to Moodle LMS and, based on that system, an EDM framework that provides flexibility for educators with different levels of technical expertise is proposed. The Educator is provided with options that allows the utilization of Data Mining techniques through either the use of predefined questions or through a guided Data Mining process.

Fig 1. User interaction flow for non-technical user

In FlexEDM the non-technical educators are presented with pre-defined questions on student performance such as:

1. Is there a relationship between the grade on a given assignment and the final grade that students achieve in a course?
2. Is there a relationship between assignment performance and Resource usage?
3. If a student does well on a particular question on a quiz, are they more likely to pass the course?
4. Can students be classified based on assignment performance?

These questions are developed by educators knowledgeable in Data Mining. The system presents these questions to the non-technical educator who can simply select a question of interest for the given course. The system then selects the appropriate Data Mining technique, performs the configure procedures using the stored settings and presents the results to the user. However the selection of the techniques and the configuration details are hidden from the user. The steps involved in performing this type of analysis is highlighted in figure 1.

This approach provides a place for the educator to start as often educators are not sure what questions to ask as they are unsure what the system can provide. Frequently re-used questions developed by other educators are added to the list so it becomes a growing repository of questions that gives educator a 'pot of gold' for analysis of student performance.

Alternatively, for educators who want to perform more exploratory analysis and who have some knowledge of Data Mining, a guided Data Mining Discovery option is presented to them. This gives the educator the ability to explore different configurations to determine any value that may exist within their data. The steps involved in performing this type of analysis is highlighted in figure 2. Educators can first select from different DM techniques supported by the Data Mining engine, including various Classification, Clustering and Association techniques. They can select what dimensions/activities to include in the analysis; this includes assignment, forum, questions, quiz, wiki, attendance and any other resource that are supported by the LMS. Educators are also guided through the process of modeling the Data Mining problem and applying filters such as discretize to convert numeric attributes in the dataset to nominal attributes. This provides a wealth of exploration for educators and the accompanying descriptions on their choices allows educators to learn while performing data mining tasks.

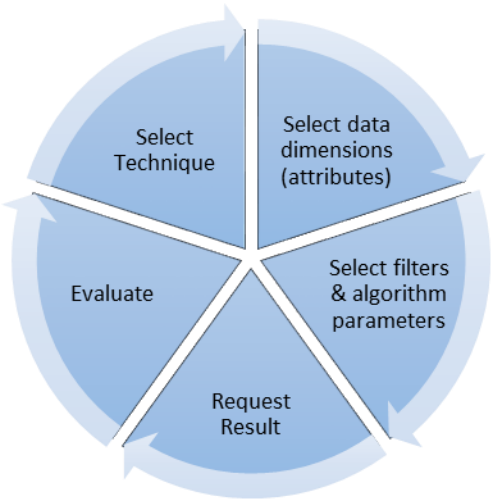

Fig 2. User interaction flow for expert user

3 Framework for Flexible EDM

The literature shows a growing number of EDM systems being developed by various researchers [1]. While many of them are external to the LMS there is an increased focus on developing tools that are integrated and provide useful insights within the learning environments. Utilizing the experience from FlexEDM, we present a framework for the development of an EDM environment. The framework was developed as a representation of the set of tasks that are performed among the various educational data mining systems. Its contribution is to reduce the level of effort among developers allowing them to focus on the specific section of interest rather than requiring the redevelopment of the entire application architecture for every research question within educational data mining.

The flexibility of the framework is based on its ability to

cater for multiple types of users; here we focus on educators who are expert users as well as educators who are non-technical users. The system facilitates this by decoupling the interaction of the user from the other components of the system. These interfaces can then be modified and improved to increase the user-friendliness of the interface for the specific types of users.

Each component of the framework as illustrated in Figure 3, is pluggable and interact with each other via a standardized interface. This allows each component to be swapped and changed without adversely affecting the functionality of the other components. The following describes each of the components by highlighting the functionality of each.

3.1 Non-Technical Educator

The non-technical educator user interface component provides the elements for educators to attempt to extract meaningful trends from within the data. This component consists of pre-configured parameters and algorithms which are displayed to the user via questions. These questions provide the user with the ability of performing the analysis without knowledge of the underlying algorithms, data transformation or parameters for configurations. The non-technical educator component will return the results of the algorithm as well as additional information that indicates the performance of the algorithm with the data supplied. The additional data will help educators determine the extent to which they can rely on the accuracy of the result produced by the analysis. The choices of techniques were based on work covered by [18] and [19].

3.2 Expert User

The expert user component provides granular control of the data mining algorithm. The interface provides the user with the data mining task (association rule mining, classifications and clustering) and the options and filters that correspond to each of the respective tasks. The data mining

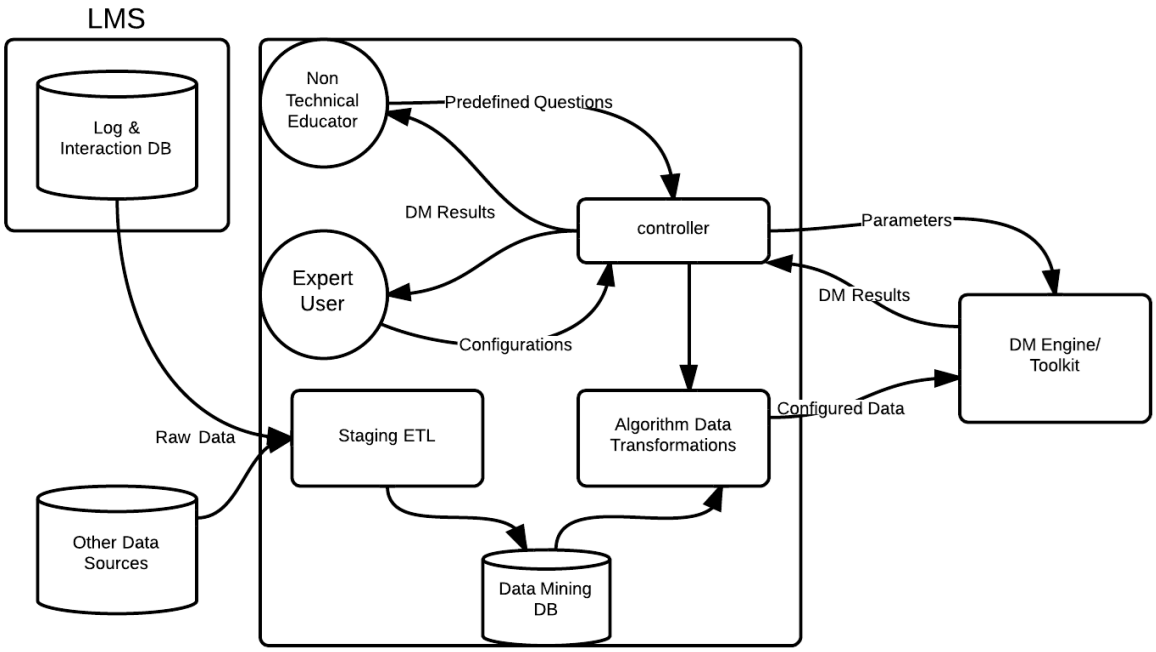

Fig 3. Framework for development of a Flexible Education Data Mining Application

results as well as the performance data is returned which allows the user to determine how to continuous modify the parameters of the algorithms in order to discover new insights.

3.3 Controller

The Controller acts as the intermediary between the various components within the framework. It coordinates and controls the passing for data and results within the system to facilitate the data mining analysis. The controller will receive the commands from the user-based components and invoke the appropriate transformation and send the data to the data mining engine based on the technique that the user selects. The controller will receive the results and pass the results to the user interface elements which will translate the results into the appropriate representation for the user.

3.4 Staging ETL

The staging ETL contains the necessarily logic and procedures for extracting the information from the LMS or e-learning system that is being analyzed and storing that preprocessed data in the Data Mining DB. This component will handle some of the preprocessing by transforming the data that may be distributed across multiple tables within the LMS into more logical groupings that make the final extraction for analysis much easier. This component in

addition to the Data Mining Database is based on an extension of a proposed model that attempts to make the data mining analysis of educational data more consistent across the various LMS platforms [6].

3.5 Algorithm Data Transformations

The Algorithm Data Transformation is responsible for converting the data into the required form that it needs for processing. Different algorithms have different requirements such as whether data can accept text, nominal or only numerical data. This component will handle the transformation in a way that is easily extendable. This allows future developers to add new algorithms and their corresponding transformations without changing the overall structure of the application.

4 Implementation

This section highlights how the component of the proposed framework is represented in the application developed.

FlexEDM uses the WEKA toolkit [20] as the data mining engine. Two techniques from the classification, clustering and association rule mining categories of analysis were selected. They were selected based on common usage [18] [19] and availability within the WEKA toolkit. The C4.5 decision tree and the Simple Bayes Network performed classification analysis, the Apriori and the FP-Growth algorithms were selected for association analysis, while the Simple K Means and the Simple Expectation Maximization (EM) algorithms were used for clustering analysis. LMS are varied in the environment in which they are developed, however the majority run over the web. Therefore a REST-based wrapper was developed to allow external entities to access the analytical capabilities of the toolkit. The system upon receiving the data for analysis will perform the following steps using the Java API for WEKA:

1. Set the class index of the data,
2. Apply the selected filters for the desired data transformation before analysis,
3. Develop the model based on the technique specified using 10-fold cross validation to reduce data over-fitting,
4. Evaluate the results acquired from the model and
5. Send the results including the report of evaluator to the requesting controller from the framework.

Though the data format and the required attributes or fields for each of the analysis types and the specific technique is different the process or sequence of activities for generating the model and performing the analysis is constant.

The Staging ETL was developed for the Moodle LMS and extracted the information from the multiple set of Moodle tables and store the results in the Data Mining Database.

For the non-technical educators the questions discussed in section 2 were presented for selection. The system performed an association rule mining using the Apriori algorithm with the usage logs for question 1, 2 and 3 while it performed the clustering analysis using the simple k-means algorithm for question 4. For the selection of the associative rule mining, the system used statistics to determine the correlation between the dimensions in the question and presented a list of generated rules with their respective confidences. For each question an explanation of the analysis was given to help the user better understand what decisions can be made from the question asked and how best to interpret the results received.

For example the question "*Is there a relationship between the grade on a given assignment and the final grade that students achieve in a course*" with use the Apriori algorithm to determine if any relationships exists. The algorithm is configured with a support threshold of 60% and a confidence threshold of 85%. Performing the rule analysis on the raw grades due to their continuous nature will not produce useful results, therefore all of the grades, including the final grade were discretized and the analysis performed on the nominal groups created from this process. The rules created and their respective support and confidence were then displayed to the user for evaluation.

For the expert user the system provides the user with a series of step by step instructions to allow the selection and configuration of the analysis desired. The first step was the selection of the techniques to be used. The user was presented with the choice between performing classification, clustering or association rule mining.

During the second step the user selected the dimensions for analysis (forums, quizzes and courses) and the filters and parameters to be passed along with the data to the data mining engine.

In using the guided data mining discovery, one educator selected Decision tree Classifier (C4.5 algorithm) to classify students on the assignment dimension splitting the users on the basis of grades on a scale of 0 – 5; another used an association rule mining technique (Apriori algorithm) to produce a list of associations to find possible relationships that may exist between dimensions.

5 Conclusion

The paper proposes a framework that allows developers and researchers to build data mining applications that are flexible to the various user types providing layers of abstractions that speak to a common interface of the application. It further demonstrates the application of this framework utilizing tools that are commonly used for data mining analysis within a LMS that is popular with both institutions and researchers alike. It builds and extends previous initiatives to decouple the data mining process from specific e-learning tools and attempts to provide more intuitive user interfaces that facilitate self-discovery of rules

and trends.

The future work of this framework is the inclusion of feedback. The system should progressively increase the options available to the non-technical users based on the feedback and successes of the expert users. Further work can be extended within the guided steps for the expert user by giving insightful suggestions based on the structure and quality of the data available. Further work can also include the exploration and measurement of different interfaces to determine which encourages the utilization of data mining analysis adoption among non-technical users. Finally, we intend to extend the framework to other classes of users such as students and administrators.

6 References

[1] C. Romero, and S. Ventura, "Educational data mining: a review of the state of the art," *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on,* vol. 40, no. 6, pp. 601-618, 2010.

[2] C. Romero, P. G. Espejo, A. Zafra, J. R. Romero, and S. Ventura, "Web usage mining for predicting final marks of students that use Moodle courses," *Computer Applications in Engineering Education,* vol. 21, no. 1, pp. 135-146, 2013.

[3] "Moodle," 12 Apirl 2014, 2014; https://moodle.org/.

[4] "Blackboard," 12 April 2014, 2014; http://www.blackboard.com.

[5] "Edmodo," 12 April 2014, 2014; https://www.edmodo.com/.

[6] A. Krüger, A. Merceron, and B. Wolf, "A Data Model to Ease Analysis and Mining of Educational Data." pp. 131-140.

[7] T. C.-K. Huang, C.-C. Liu, and D.-C. Chang, "An empirical investigation of factors influencing the adoption of data mining tools," *International Journal of Information Management,* vol. 32, no. 3, pp. 257-270, 2012.

[8] J. Han, M. Kamber, and J. Pei, *Data mining: concepts and techniques*: Morgan kaufmann, 2006.

[9] R. Baker, "Data mining for education," *International encyclopedia of education,* vol. 7, pp. 112-118, 2010.

[10] F. Castro, A. Vellido, À. Nebot, and F. Mugica, "Applying data mining techniques to e-learning problems," *Evolution of teaching and learning paradigms in intelligent environment*, pp. 183-221: Springer, 2007.

[11] A. Zafra, C. Romero, and S. Ventura, "Multiple instance learning for classifying students in learning management systems," *Expert Systems with Applications,* vol. 38, no. 12, pp. 15020-15031, 2011.

[12] C. Romero, S. Ventura, P. De Bra, and C. De Castro, "Discovering prediction rules in AHA! courses," *User Modeling 2003*, pp. 25-34: Springer, 2003.

[13] A. Merceron, and K. Yacef, "Tada-ed for educational data mining," *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning,* vol. 7, no. 1, pp. 267-287, 2005.

[14] D. R. Silva, "MTP: Using Data Warehouse And Data Mining Resources For Ongoing Assessment Of Distance Learning."

[15] R. Mazza, and C. Milani, "Gismo: a graphical interactive student monitoring tool for course management systems." pp. 18-19.

[16] O. R. Zaïane, and J. Luo, "Towards evaluating learners' behaviour in a web-based distance learning environment." pp. 357-360.

[17] A. L. Dyckhoff, D. Zielke, M. A. Chatti, and U. Schroeder, "eLAT: An Exploratory Learning Analytics Tool for Reflection and Iterative Improvement of Technology Enhanced Learning." pp. 355-356.

[18] V. Kumar, and A. Chadha, "An Empirical Study of the Applications of Data Mining Techniques in Higher Education," *International Journal of Advanced Computer Science and Applications,* vol. 2, no. 3, pp. 80-84, 2011.

[19] X. Wu, V. Kumar, J. R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G. J. McLachlan, A. Ng, B. Liu, and S. Y. Philip, "Top 10 algorithms in data mining," *Knowledge and Information Systems,* vol. 14, no. 1, pp. 1-37, 2008.

[20] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, and I. H. Witten, "The WEKA data mining software: an update," *SIGKDD Explorations,* vol. 11, no. 1, pp. 10-18, 2009.