CLASSIFICATION OF LEARNING STYLES IN VIRTUAL LEARNING ENVIRONMENT USING J48 DECISION TREE

Renato R. Maaliw III[1] and Melvin A. Ballera[2]
[1]*Southern Luzon State University - Philippines*
[2]*AMA University - Philippines*

ABSTRACT

The usage of data mining has dramatically increased over the past few years and the education sector is leveraging this field in order to analyze and gain intuitive knowledge in terms of the vast accumulated data within its confines. The primary objective of this study is to compare the results of different classification techniques such as Naïve Bayes, Logistic Regression, Conjunctive Rule and J48 Decision Tree in detection and identification of student's learning styles in a Virtual Learning Environment to provide adaptation strategy according to identified learning styles of the students. The data sets were collected from 507 students of Computer Programming 1 course with a total of 52,815 rows of data extracted from their interaction logs and navigational patterns in a virtual learning environment. A mapping of student's learning style according to the selected learning style model had been accomplished. The performance of each classification techniques and its classification quality were measured in terms of correctly classified instances, kappa statistics, receiver operating characteristics, and area under the curve plots. Based from the analysis of the comparative results, the classification technique that has produced the highest collective average accuracy is the J48 Decision Tree with correctly predicted instances of 87.42%. The classification technique could be used to identify student's learning styles in a virtual learning environment.

KEYWORDS

Educational data mining, learning styles, learning style model, virtual learning environment, J48 decision tree

1. INTRODUCTION

The field of educational data mining (EDM) is an "emerging and evolving discipline that is particularly concern in creating and developing methods for exploring the different and unique types of data that comes from the educational settings". Higher Education Institutions (HEIs) have been investing thoroughly in providing educational infrastructure such as Virtual Learning Environment (VLE). This type of technology enables educators to represent knowledge by delivering different types of content and, monitor student participation that can enhance the learning process. VLE provides the creation of knowledge representation and contains set of tools for communications, assessments, and various features. Most universities employs traditional face-to-face interaction as the main approach in teaching in the sector of education but HEIs are investing on the infrastructure to supplement traditional methods of teaching with learning technologies in order to enhanced and improve the quality of learning (Dumcience et al., 2010). But despite the astronomical increase in practice of using VLEs in HEIs, the application of these technologies are still failing miserably in terms of inciting student motivation to learn, and "it fails to serve the ultimate goal of having on-line learning" (Ballera et al., 2013).

VLEs are rich sources of learner's data that the education sectors are completely overlooking and sometimes completely neglecting. It stores valuable source of information that should be considered in order to understand how student learns. The amount of data that can be extracted that are residing inside these VLEs can be exploited to better understand learner's behavior, improve the pedagogical process that leads to an efficient and streamlined teaching and learning progress. It can be used in the detection and identification of student's learning styles (LS) to provide adaptation strategies to the course design and contents of VLEs in order to match the student's preference in terms of their learning styles. There are substantial research studies that considered learning style as a vital factor that directly affects the student's learning process. It is one of

the responsibilities of the teacher to understand how different learner learns in accordance to their learning styles. It cannot be denied that learners can encounter difficulties in an environment where their learning styles are not supported. This study focuses in the detection and identification of student's learning style in the context of a virtual learning environment using data mining techniques based from their extracted interaction logs and navigational patterns.

2. REVIEW OF RELATED LITERATURE

2.1 Learning Styles

Keefe (1979) defines learning styles as the "collective characteristics of affective, cognitive, and physiological factors serving as stable indicators of how a particular learner interacts with, perceives, and responds to a learning environment" while Stewart et al. (1992) defines learning styles as the "educational conditions under which a student is most likely to learn".

Learning style's primary concern is how an individual prefers to learns and not what the learner learns. When an individual tries to learn something new they vary regarding their preferences. Some learn by communicating to someone, some by reflecting about what they have learned, or prefers textual and graphical representation of the learning material. It can also be describe as a set of behaviors, and attitudes that enhance and streamline learning in any situation. Each individual has their own way of preferences when it comes to learning; it is an innate characteristic that are influenced by the environment, experiences, and developments. There are many learning style models that have existed in the literature but notable educational theorist are in agreement that each individual has their own preferred ways of learning that must be met by the learning environment. These learning styles are clear indicators of how learners perceive, react, interact with, and respond to the learning environments. There are "overwhelming evidences that learning styles are very diverse and educators should consider these differences in students to be taught in a method that are well-suited to their learning styles" (Pashler et al., 2008).

2.2 Dimensions of the Felder-Silverman Learning Style Model

The Felder-Silverman learning style model (FSLSM) is a learning style model based on the notion that students have preferences in terms of the way they receive and process information. The model presents different dimensions that are indicative of learning preferences. Four dimensions are described in the learning style model of FSLSM namely processing, perception, input, and understanding. Learner's learning styles are specifically defined for each of these dimensions. All dimensions are independent from one another and are based from major dimensions of learning styles. They describe learner's preferences to process information (active/reflective), perceive information (sensing/intuitive), receive information (visual/verbal), and organize (sequential/global) information.

Active learners are characterized as learners who prefer to process information by doing something with the learned material. The most obvious pattern of behavior of an active learner is that they have a strong tendency to discuss and interact with other learners. On the other hand, reflective learners prefer to think about the learning material and they work alone. In a VLE, active learners are expected to post more often in a forum while reflective learner's tendencies are to participate passively by visiting forum but rarely posting. Active learners also tends to attempt more self-assessment tests, a type of test where the result is not graded but it is important in order for the learner to assess their knowledge on a particular topic. The characteristic of reflective learner to think about the material leads to more visits in learning objects that are textual-based in context.

Sensing learners tends to repeatedly visit concrete learning materials that contains facts, data, and when a learning object is being linked to real life context whereas intuitive learners prefer to visit abstract learning materials that contains theories and their underlying meanings, histories, glossaries, syntax, and concepts. The number of visits on these kinds of learning objects serves as a pattern. On the other hand, intuitive learners supposed tendencies are to visit learning objects such as an example as a supplementary material. The number of visits on these kinds of learning objects tends to be higher for intuitive learners. Intuitive learners are also characterized as a careful worker. With regards to this, intuitive learners tend to be careful

and tend to review their answers more when performing a test especially when it is being graded. The pattern can be conceived by the number of attempted answer reviews they made in an exercise before attempting to submit their answers.

Accordingly, verbal learner's preference on a learning object is composed chiefly of words or texts, they tend to like communication with others and discussions. Therefore, verbal learner's tendencies are to use the forum, thus a high number of forum postings can indicate a verbal learning style. On the opposite end, visual learners learn best from what they actually see. Therefore, they tend to view more learning objects that usually contains graphics such as diagrams, charts, and pictures. Video presentations also are highly preferred by visual learners while verbal learners are expected to visit a learning object of textual-based types.

Sequential learners are more comfortable with details, whereas global learners tend to be good in seeing the overall picture and connections to other fields. Because of this kind of behavior, global learners are interested in getting the "big picture" and an overview. Course outlines are especially important to them whereas sequential learners tend to skip these kinds of learning objects. A high number of visits spent on chapter outlines, course overview or chapter overview page indicate a global learning style. The navigational patterns of learners when using a VLE can be used also to differentiate sequential or global learning style as well. While sequential learners tend to go through the course step by step in a linear way, global learners tend to learn in large gaps by skipping learning objects and jumping to more complex and advanced learning objects. Therefore, the navigational patterns can be seen as an indicator to differentiate the two styles.

3. METHODOLOGY

3.1 Data Source and Methods of Collection

The data sets are based from the log files of Introduction to C++ Moodle Course (Computer Programming 1) in Southern Luzon State University (SLSU) in the Philippines. A blended learning environment has been implemented with this specific course by supplementing traditional face-to-face setup by a Moodle (Moodle Learning Management System) course that includes several and carefully designed learning objects such as chapter overviews, visual material, video presentation, textual material, audio material, examples, glossary, concrete, and abstract materials. Different types of assessment such as self-assessment and exercises are provided to allow students to practice their programming skills. Forum usage is also encouraged to increase interaction among students and to solve group related problems during the duration of the course. Although there are several courses in the virtual learning environment, this specific course was selected for it is found to have a large number of students enrolled thus it contains large amount of interaction and log data.

Five hundred seven (507) students have participated and answered the Index of Learning Style (ILS) questionnaires based from FSLSM. Students have been given ample amount of time and each question in the questionnaires is carefully explained to them to avoid contaminating the data. Table 1 shows the learning styles' distribution for all dimensions of the FSLSM without considering the degree of learning style preference.

Table 1. Distribution of learning styles based from ILS

Dimension	Processing	No. of Students	Percentage
Processing	Active	244	48.13%
	Reflective	263	51.87%
Perception	Sensing	348	68.64%
	Intuitive	159	31.365
Input	Visual	388	76.53%
	Verbal	119	23.47%
Understanding	Sequential	250	49.31%
	Global	257	50.69%

From the analysis of ILS questionnaires collected from 507 students, it reveals that at the processing dimension there are 263 (51.87%) students that have a reflective learning style while 244 (48.13%) students have an active learning style. This manifest that students learning styles in the processing dimension is fairly balanced. The perception dimension reveals that 348 (68.64%) students have a sensing learning style. This means that most students prefer learning objects that are based from real facts and data, and a learning object that is linked to real life situations. At the input dimension, the data collected reveals that 388 (76.53%) of students prefers learning objects that mostly contains graphics such as pictures, charts, diagrams, and video presentation materials. Finally, the data reveals that student's learning style in the understanding dimension is fairly balanced with 257 (50.69%) students with global learning styles and 250 (49.31%) students with sequential learning styles.

3.2 Mapping of Students Learning Styles in Virtual Learning Environment

Table 2 provides the lists of the learning style mapping of relevant student's behavior in a VLE. These sets of relevant behaviors were extracted from the VLE database to construct the data sets for data mining.

Table 2. Learning style mapping of relevant students' behavior in VLEs

Learning Style	Relevant Behavior	Attribute Name	Attribute Value
Active	Post more often in discussion forum	forum_posts	no. of posting in forum
	Perform more self-assessment tests	self_assesment	no. of viewed post in forum
Reflective	Reading post but rarely posting by themselves	forum_view	no. of viewed post in forum
	Prefers learning material in textual form	text_materials	no. of visits
Sensing	Prefers concrete learning materials (facts, data)	concrete_materials	no. of visits
	Prefers examples	examples	no. of visits
Intuitive	Prefers abstract learning material (definition, theories, syntax, flowcharts)	abstract_materials	no. of visits
	Prefers to review answers in graded exercise tests	exercises_rev	no. of attempted answer reviews
Visual	Prefers learning materials supplemented with pictures, diagrams, graphs	visual_materials	no. of visits
	Prefers learning materials presented in a video format	video_materials	no. of visits
Verbal	Prefers learning material presented in text or audio	text_materials	no. of visits
	Post more often in discussion forum	forum_post	no. of posting in discussion
Sequential	Prefers to go through the course step by step (linear way)	nav_pattern_dist	sequence of navigational pattern
Global	Prefers overviews, outlines	course_overviews	no. of visits
	Prefers to learn in large leaps by skipping learning material & jumping to more complex materials (non-linear way)	nav_pattern_dist	sequence of navigational pattern

3.3 Navigational Pattern Sequence Data Collection

Navigational patterns refers to how learners navigate through the course and in which order they visit certain types of learning objects. In a study by Imra et al. (2016) on personalized learning recommendation system, they have proposed identification of navigational sequence pattern using the formula for Euclidean distance to compute the similarity and difference between learners on their navigational characteristics. With this formula, navigational patterns of students in a Virtual Learning Environment can be inferred by analyzing and computing their navigational sequence distance values. In a similar research by Benlarmi (2003) on dynamic learning modeler, they have used approximately thirty (30) navigational sequences to cluster similarities between students navigational behavior in a hypermedia courseware. In accordance to previous studies, the navigational distance values in order to distinguish a user that are navigating sequentially or navigating globally can be tracked when accessing the online course.

3.4 Data Preprocessing, Transformation and Attribute Value Extraction

Every logs and activities of all students are recorded in the VLE database. Primarily, VLEs such as Moodle provides a module for extraction of user logs and activities for a specific course by exporting to varieties of file formats such as Microsoft Excel (*.xlsx) file. The table is comprised of data labels such as 'Time', 'User Full Name', 'Affected User', 'Event Context', 'Event Name', 'Description', 'Origin' and 'IP Address'. The initial data sets were subjected to data preprocessing by removing all unnecessary data. Interaction logs of each target students were extracted to produce a reduced log file that contains the data labels of 'User Full Name', 'Event Context', 'Event Name' and 'Type'.

The reduced log file extracted contains 52,815 rows of student logs and activities for the course that were used in classification of individual learning styles. With reference to Figure 1, a 'Learning Object Type' field has been created in order to identify as to what type of learning object each particular students interacts with. Identification of learning object type in the course was also mapped in order to identify as to what kind of learning object each truly represents based on the learning object literature types whether textual learning materials, visual learning materials, abstract learning materials, concrete learning materials and examples. These learning object type identification are evaluated by educational experts. It is necessary to distinguish these learning object types in order to effectively create data sets in inferring student's learning styles.

User full name	Event context	Event name	Learning Object Type
Beneluz Abustan	Page: Chapter 1 Overview	Course module viewed	course_overviews
Beneluz Abustan	Page: C++ Terminology	Course module viewed	abstract_materials
Beneluz Abustan	Page: Assignment Statements	Course module viewed	text_materials
Beneluz Abustan	Page: Assignment Statements	Course module viewed	text_materials
Beneluz Abustan	Page: Literals	Course module viewed	text_materials
Beneluz Abustan	Page: Literals	Course module viewed	text_materials

Figure 1. Excerpt of reduced log file of student's interaction and activities in a VLE

The next phase in the construction of the data sets is data transformation. The reduced log data in Microsoft Excel format is transformed and converted to a Microsoft Access file (*.accdb) format in order to easily aggregate the total number of interaction a particular learner to each learning objects. An aggregate SQL statement commands was used to extract the needed values for data mining and analysis. Derived variables was extracted by aggregating variables such as number of visits to a specific learning object, the number of submitted self-assessment tests, number of forum postings and views, and more.

The final phase in the construction of data sets is to arrange the derived values of number of interactions of the students to a specific learning object. It was arranged and categorized based from the learning behavior pattern mapping in Table 2. An excerpt of final data set construction is depicted in Figure 2. An additional data field was created to accommodate the reported learning styles of each student based from their answers from the ILS questionnaire. The results of the questionnaire served as the class labels of each student in terms of their learning styles for each learning style dimensions. The final data sets were used in data mining.

No.	Student	forum_post	forum_view	self_assesment	text_materials	PROCESSING
1	ABUSTAN, BENELUZ	1	2	8	27	ACTIVE
2	ABUAN, SIDNEY JANE	11	6	3	0	REFLECTIVE
3	ABULAR, MA. FREDA MAE	21	10	7	29	ACTIVE
4	ABULENCIA, APPLE GEM	9	2	8	1	ACTIVE
5	ABUSTAN, AUDREY CASSIE	4	3	8	17	ACTIVE
6	ACERO, CHANTREA FELICHE	15	11	0	14	ACTIVE

Figure 2. Excerpt of final data set construction (processing dimension)

3.5 Feature Selection

To determine the best features or attributes for determining the learning styles of the students in each dimension, attribute selection was used. Filtering method using Information Gain attribute evaluation was selected. The objective of feature selection technique testing is to empirically confirm and improve the classification performance of the predictors, providing faster and more cost-effective predictors, and providing a better understanding of the underlying process that generated the data. By applying attribute selection, significant predictors were extracted from each mapped attributes for each learning style dimensions. Summary of feature selection results can be seen in Table 3 and based from the results, eleven (11) attributes have been found to be significant in inferring learning styles in VLEs.

Table 3. Results of feature selection for each FSLSM Attributes

Information Gain Attribute Evaluation		
Processing Dimension Attributes	**Rank Value**	**Significant? (yes/no)**
forum_view	0.449	yes
self_assessment	0.338	yes
forum_posts	0.267	yes
textual_materials	0	no
Perception Dimension Attributes	**Rank Value**	**Significant? (yes/no)**
concrete_materials	0.353	yes
exercises_rev	0.241	yes
examples	0.107	yes
abstract_materials	0.093	yes
Input Dimension Attributes	**Rank Value**	**Significant? (yes/no)**
video_materials	0.382	yes
visual_materials	0.269	yes
forum_posts	0	no
textual_materials	0	no
Understanding Dimension Attributes	**Rank Value**	**Significant? (yes/no)**
course_overview	0.285	yes
nav_pattern_dist	0.039	yes

3.6 Pattern Discovery

In this phase, different classification data mining techniques was applied to the derived data sets. For the reason of its open-source implementation, Waikato Environment for Knowledge Analysis (WEKA) was used in order to analyze the performances of the selected classification methods to discover the most appropriate and best model that can be used for the future development of an adaptive VLE for different learning styles.

3.7 Evaluation

The model evaluation is an integral part of the model development process. The quality of classification was evaluated using the Receiver Operating Characteristics (ROC) and Area under the Curve (AUC) plots to measure the accuracy of the model. A random classifier has an area under the curve of 0.5, while AUC for a perfect classifier is equal to 1. In practice, most of the accurate classification models have an AUC between 0.5 and 1. A rough guide for classifying the quality of classification of the area under the curve is based on the traditional academic point system as shown in Table 4.

Table 4. Traditional Academic Point System

Range	Description
.90 – 1.00	Excellent
.80 - .90	Good
0.70 – 0.80	Fair
.60 - .70	Poor
.50 - .60	Fail

Another evaluation measure of classification quality in data mining is the Cohen's Kappa Equivalent. It is a coefficient which measures the inter-rater agreement for qualitative categorical items. The measurement was applied also to measure the classification accuracy when performing classification in data mining. Kappa statistics (Cohen, 1960) was used to assess the accuracy of any particular measuring cases. Cohen's kappa equivalent values are shown in Table 5.

Table 5. Cohen's Kappa equivalent values

Kappa Score	Equivalent
0.81 – 1.00	Perfect
0.61 – 0.80	Substantial
0.41 – 0.60	Moderate
0.21 – 0.40	Slight
<= 0	None

4. RESULTS AND DISCUSSIONS

A total of eight (8) classification techniques have been utilized to test the performance of the classifiers based on the final data sets but the top four performing algorithms such as Logistic Regression, Naïve Bayes, Conjunctive Rule and J48 Decision Tree were selected. Classification performances are tested on all four dimensions of the FSLSM. Comparative tests results can be seen in Table 6 and it is summarized based on classification accuracy and the kappa statistics values.

Table 6. Comparative performance results of classification using different techniques

Processing Dimension (Active/Reflective)				
	Simple Logistic	**Naïve Bayes**	**Conjunctive Rule**	**J48**
Correctly Classified Instances	85.99%	89.34%	75.14%	**92.50%**
Incorrectly Classified Instances	14.01%	10.65%	24.85%	7.49%
Kappa Statistics	0.719	0.786	0.497	0.849
Perception Dimension (Sensing/Intuitive)				
	Simple Logistic	**Naïve Bayes**	**Conjunctive Rule**	**J48**
Correctly Classified Instances	81.65%	82.24%	68.63%	**88.16%**
Incorrectly Classified Instances	18.34%	17.75%	31.36%	11.83%
Kappa Statistics	0.550	0.586	0	0.699
Input Dimension (Visual/Verbal)				
	Simple Logistic	**Naïve Bayes**	**Conjunctive Rule**	**J48**
Correctly Classified Instances	85.79%	85.99%	76.52%	**86.58%**
Incorrectly Classified Instances	14.20%	14.00%	23.47%	13.41%
Kappa Statistics	0.582	0.634	0	0.677
Understanding Dimension (Sequential/Global)				
	Simple Logistic	**Naïve Bayes**	**Conjunctive Rule**	**J48**
Correctly Classified Instances	80.27%	74.95%	81.26%	**82.44%**
Incorrectly Classified Instances	19.72%	25.04%	18.73%	17.55%
Kappa Statistics	0.605	0.500	0.624	0.647

Based on the comparative classification performance of different classification techniques, J48 decision tree classifier obtained the highest accuracy with an average of 87.42% collectively for all dimension. The Logistic Regression, Naïve Bayes and Conjunctive Rule yield a collective average accuracy across all learning dimension of 83.42%, 83.13% and 75.38% respectively. Kappa scores from the J48 method of 0.849, 0.699, 0.677, and 0.647 for the respective dimensions reflected that there is a 'Substantial' accuracy in classification.

To further test the classification quality of the J48 classification model, a total of four (4) ROC curve plots have been generated for each learning style dimension using KnowledgeFlow in WEKA.

The ROC curves for all learning style dimensions did not fall below the random guessing line threshold of 0.5 which suggests that the quality of qualification is well beyond random guessing and the classification did not happen by chance only. Furthermore, the values obtained from the AUC plots are 0.9176 (Excellent) for processing dimension, 0.8318 (Good) for perception dimension, 0.9016 (Excellent) for input dimension, and 0.8124 (Good) for the understanding dimension. Average AUC scores across all four dimensions are 0.86585 that signifies that the model provides a good classification quality. Having determined that the J48 classifier was the best fit for the classification purposes of learning styles, the rules generated from the J48 classification technique can be used in the development of a prototype of an adaptive VLE for different learning styles.

5. CONCLUSION

The interaction logs and navigational patterns of the participating students were extracted from the VLE's database. A total of 52,815 rows of data were extracted from the five hundred seven (507) participating students in the study. This study was limited only on the obtained Moodle log data for the specified course and did not consider the analysis of student's respective homework, recitation and discussion performances. The classification model was implemented on the data sets for the Introduction to C++ Moodle Course of SLSU in the Philippines. Based from the empirical results and comparative performance results of different classification algorithms, the technique by means of J48 decision tree classifier attained the highest accuracy with 87.52%.

Implementation of VLEs has been successful in most HEIs but in most cases this piece of technology fails to meet the unique needs of each students particularly their learning styles. Furthermore, there is no mechanism or a feature in most VLEs that automatically classifies student's based from their learning styles. The knowledge divulged from the empirical results can be a basis for HEIs, educators, and course developers that each student is unique in characteristics in terms of their learning styles thus they have different needs.

The next phase of the study is to develop a prototype software of an adaptive Virtual Learning Environment system in order to dynamically adjust its course design and contents to respond immediately to each student's needs based from the derived classification model from this study.

REFERENCES

Ballera, Melvin A., and Elssaedi, Mosbah Mohamed, 2013. New E-learning Strategy Paradigm: A Multi-Disciplinary Approach to Enhance Learning Delivery, *Proceedings of the 2nd E-learning Regional Conference*, State of Kuwait, pp. 25-27.

Benlarmi, S.K. 2003. A Dynamic Modeler for Navigational Sequence Patterns. *Journal of Research on Technology and Innovations*. vol 3., no. 2, pp. 121-127.

Carver, C.A., Howard, R.A., and Lane, W.D., Addressing Different Learning Styles through Course Hypermedia. IEEE Transactions on Education, vol. 42, no.1, pp. 33-38, 1999

Cohen J. A. 1960. Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20, pp. 37-46.

Dumciene, A., Lapeniene, D., 2010. Possibilities of Developing Study Motivation in E-Learning Products. Electronics and Engineering. – Kaunas: Technolojia, No. (102) pp. 43-46.

Felder, R.M., and B.A. Soloman, 2016. Index of Learning Styles, http://www.ncsu.edu/felder-public/ILSpage.html, 2004, accessed May 2.

Felder, R.M., and Silverman, L.K., 1987. Learning Styles and Teaching Styles in Engineering Education. *Presented at the 1987 Annual Meeting of the American Institute of Chemical Engineers*, New York.

Garcia P., Amandi A., Schiaffin S., Campo M., 2007. Evaluating Bayesian Networks Precision for Detecting Student's Learning Styles. *Computers and Education*, 49, pp.794-808, Elsevier.

Imra J. and Zawili S., 2016. Personalized Learning Recommendation System for E-learning. *International Journal of Science and Technology*. vol.2, no. 1, pp. 85-90.

Keefe, J.W., 1979. Learning Style: An Overview. *In National Association of Secondary School Principlas (Ed.)*, Student Learning Styles: Diagnosing and Prescribing Programs, pp. 1-17

Moodle Learning Management System. [Online]. Available: http://www.moodle.org.

Pashler, H., McDaniel, M. Rohrer, D, Bjork, R., 2008. Learning Styles: Concepts and Evidence. *Psychological Science in the Public Interest*, vol. 9, pp. 105-119.

Stewart, K.L., Felicettie, L.A., 1992. Learning Styles of Marketing Majors. *Educational Research Quarterly*, 15(2), pp. 15-23.

WEKA at http://www.cs.waikato.ac.nz/~ml/weka. Retrieved April 29, 2016.