Using Feature Selection and Association Rule Mining to Evaluate Digital Courseware

Shaveen Singh

School of Computing, Information & Mathematical Science
University of the South Pacific
Suva, Fiji Islands
singh_sv@usp.ac.fj

Sunil Pranit Lal

School of Computing, Information & Mathematical Science
University of the South Pacific
Suva, Fiji Islands
lal.sunil@ieee.org

Abstract— **Effective digital courseware should be easy to implement and integrate into instructional plans, saving teachers time and helping them support their students' learning needs. It should also not only enable students to achieve explicit learning objectives but also accelerate the pace at which they do so. This paper highlights the advantage of using Feature Selection techniques and Associative rule mining to get insightful knowledge from the log data from the Learning Management System (Moodle). The Machine Learning approach can be objectively deployed to obtain a predictive relationship and behavioral aspects that permits mapping the interaction behaviour of students with their course outcome. The knowledge discovered could immensely assist in evaluating and validating the various learning tools and activities within the course, thus, laying the groundwork for a more effective learning process. It is hoped that such knowledge would result in more effective courseware that provides for a rich, compelling, and interactive experience that will encourage repeated, prolonged, and self-motivated use.**

Keywords— *e-learning; attribute ranking; machine learning; online development,association rule mining.*

I. INTRODUCTION

Technology can help support effective teachers, by giving them new ways to design and provide personalized instruction to their students. More than 90% of all teachers in a recent Harris Interactive survey indicated that "they would like to use more education technology in the classroom than they do now" (Harris Interactive, 2013);

In the recent past, there has been a significant growth in the use of the ML techniques to seek knowledge in the area of higher education. Functionalities such as: student performances prediction [3-5, 7], student behaviour modelling [6], mediation of student e-discussions [8], student retention [9, 10] and other data mining applications have been established and researched into, with reference to the e-learning systems.

E-learning systems provide multiple ways of learning (self-paced, collaborative, synchronous & asynchronous, tutorial-based, homework, etc.) and incorporate numerous interactive online activities such as forums, quizzes, lessons, blogs, assignments, surveys, glossaries, wiki and workshops [11]. Most of these LMS maintain a comprehensive log of user interaction and assessment data. The use of this student

learning and engagement data can greatly drive decision-making and continuous improvement of courses in the digital landscape. It can provide improved selection of online activities and better organization of the courseware to support student learning.

The seminal objectives of this paper is to highlight the usage of student interaction log from the LMS to identify effectiveness of courseware activities in fulfilling the learning outcome of a course and garnering optimized students' performances. The paper discusses the use of Feature Selection techniques with the combination of Association Rule Mining within the context of Education Data Mining to rank significant course activities and identify more noteworthy relationships amongst these attributes. Engagement data from a generic course from the University of the South Pacific is used to demonstrate the concept. Instructional designers can use this knowledge when designing courses, improve the structure of course content, appropriately weight assessment tasks and use the data to provide prompts for students to improve their learning experience.

A. ELearning at The University of the South Pacific

The University of the South Pacific (USP) was set up in the South Pacific region in 1968 by its 12 member countries - Cook Islands, Fiji Islands, Kiribati, Marshall Islands, Nauru, Niue, Samoa, Solomon Islands, Tokelau, Tonga, Tuvalu and Vanuatu. A total of 14 campuses are spread over an area of 30 million square kilometers of the Pacific Ocean. Due to this geographical separation, the university is expected to take its products and services to the doorstep of each and every household in the USP region [13]. This has resulted in a pedagogical shift from traditional face-to-face or chalk-and-talk to more flexible learning modes of delivery. Even though the print-based has been the preferred mode of delivery, the low pass rates have prompted the educational practitioners to shift emphasis to blended and fully online modes [12].

While, the blended and fully online courses are seen as being cost-effective, scalable, innovative, student and learning centered, flexible scheduling of student activities, attracting huge student enrolments and providing students an equal platform, a couple of problems they inherently face are the lack of self-motivation and irregular participation resulting in excessive number of unfinished tasks late in the course. As a consequence, many students fail the courses offered through

these delivery modes. Therefore, there is a genuine need to evaluate and validate the curriculum and assessment design to lay the groundwork for a more effective learning process [12]. The knowledge acquired from this research can be used to highlight activities that significantly improve the students' ability to pass the course. Such activities could be emphasized as they add-value to the course while the insignificant activities can be re-examined for its suitability for the context and courseware designers can re-visit its integration methodology.

II. FEATURE SELECTION TECHNIQUE

Feature selection is an active field in computing science. It has been a fertile field of research and development since 1970s in statistical pattern recognition [15], machine learning and data mining [13, 16, and 17]. Feature selection is a fundamental problem in many different areas, especially in forecasting, bioinformatics, document classification, and object recognition or in modeling of complex technological processes and behaviour [18]. Datasets with thousands of features are not uncommon in such applications. All features may be important for some problems, but for some target concepts, only a small subset of features is usually relevant.

A similar dilemma is faced by the instructional designers and course coordinators in this ever-changing e-learning landscape. The e-learning system provides as spectrum of activities and tools that can be integrated within the LMS platform (about 25 different types of activities are available in Moodle [11]). When designing online courses, instructors have to evaluate their courses and choose the most appropriate technical and pedagogical tools/activities for their course offering. The challenge is to create a correct mix of learning activities that employ different strategies for attaining the expected learning outcomes of the course. In doing so, they should be able to appeal to the different learning styles that are present in the student population. By just having the same types of activities in all courses in the same mode of offering may not be effective within and across disciplines. Another very important factor to determine is which online courseware is most suitable for courses placed at different levels in a programme.

In the context of our university, another common problem in offering online courses is slower dial-up speeds at home, especially for those rural students enrolled in online distance-learning courses. These students cannot take courses effectively unless instructors scale back course activities accordingly. Thus, the correct exploitation of feature selection techniques of ML can be used to highlight most successful courseware activities in such environment. The feature selection approach can also drastically reduce the variety of activities available in the LMS and help identify and eliminate redundant, irrelevant, or insignificant courseware.

A. Feature Selection Process

It is possible to derive the general architecture from most of the feature selection algorithms. It consists of four basic steps (refer to Fig. 1); subset generation, subset evaluation, stopping criterion, and result validation [14]. The feature selection algorithms create a subset, evaluate it, and loop until

an ending criterion is satisfied. Finally, the subset found is validated by the classifier algorithm.
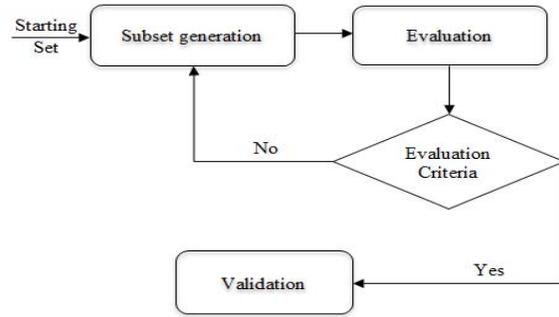
Figure 1. Overview of Feature Selection Process.

A. Subset Generation

Subset generation is a search procedure that evaluates a subset of features as a group for suitability. Evaluation of the subsets requires a scoring metric that grades a subset of features. The total number of candidate subsets is 2^N, where N is the number of features in the original data set. Having to exhaustively search through the feature space with even moderate N is infeasible. Non-deterministic search like evolutionary search or the use heuristic search methods can be used to build the subsets. There are two main families of these methods: forward addition (starting with an empty subset, we add features after features by local search) or backward elimination (which is based on elimination of features).

B. Subset Evaluation

In Subset evaluation, each candidate subset generated needs to be evaluated and compared with the previous best one according to a certain evaluation criterion. If a new subset turns out to be better, it replaces the previous best subset. The process evaluation is repeated until a given stopping criterion is satisfied.

The evaluation procedure used can be divided into *wrapper* and *filter* methods. In the wrapper method, classifier algorithm usually wrapped in a loop, whereas, in filter methods does not rely on the classifier algorithm, but uses other criteria based on correlation notions.

C. Stopping Criteria

Unless there is a suitable stopping criterion, the feature selection process will run exhaustively. The process may be tuned to stop under one of the following reasonable criteria: (1) once a predefined number of features are identified, (2) a predefined number of iterations have been reached, (3) in case addition (or elimination) of feature fails to produce a better subset, (4) an optimal subset as per the evaluation criterion is obtained.

D. Validation

The selected best feature subset needs to be validated by carrying out different tests on both the selected subset and the original set either by using artificial and/or real-world datasets.

III. Feature Ranking and Selection

Diverse feature ranking and feature selection techniques have been proposed in the machine learning literature. The purpose of these techniques is to discard irrelevant or redundant features from a given feature vector. For the purpose of this experiment, we used feature ranking and selection methods with two basic steps of general architecture: subset generation and subset evaluation for ranking the feature in the dataset.

In this paper, we consider the filter methods to evaluate the ranking. The following practical and commonly used statistical and entropy-based measures have been selected based on its good performance in various domains:

- Information Gain (IG),
- Gain Ratio (IGR),
- Symmetrical Uncertainty (SU) ,
- Relief-F (RF) attribute evaluation,
- One-R (OR) attribute evaluation,
- Chi-Squared (CS) attribute evaluation.

Entropy is commonly used in the information theory measure [19], which characterizes the purity of an arbitrary collection of examples. It is in the foundation of the IG, IGR, and SU attribute ranking methods. The entropy measure is considered a measure of the system's unpredictability. The entropy of Y is calculated as:

$$H(Y) = -\sum_{y \in Y} p(y) \log_2(p(y)) \qquad (1)$$

where *p(y)* is considered the marginal probability density function for the random variable *Y*. If the observed values of *Y* in the training data set *S* are partitioned according to the values of a second feature *X*, the entropy of *Y* after observing *X* is then computed as:

$$H(Y \mid X) = -\sum_{x \in X} p(x) \sum_{y \in Y} p(y \mid x) \log_2(p(y \mid x)) \quad (2)$$

where *p(y|x)* is the conditional probability of *y* given feature *x*. If the entropy of *Y* with respect to the partitions induced by *X* is less than the entropy of Y prior to partitioning, then we can conclude that there is a relationship between features *Y* and *X*.

A. Information Gain

Information Gain (IG) evaluates attributes by measuring their information gain with respect to the class. It discretises numeric attributes first using the MDL-based discretisation method.

$$IG(Y, X) = H(Y) - H(Y \mid X) \qquad (3)$$

where *H* is the information entropy. IG is a widely used standard feature selection method but it does not take into account feature interaction.

B. Gain Ratio

The Gain Ration (GR) evaluates attributes by measuring their gain ratio with respect to the class. It is calculated by dividing the original information gain by the information value of the attribute.

$$IGR(Y, X) = \frac{IG}{H(X)} \qquad (4)$$

In opposition to IG, IGR favours variables with fewer variables.

C. Symmetrical Uncertainty

Symmetrical Uncertainty (SU) compensates for the IG's bias toward features with more values and normalizes its values within the range [0, 1] with the value 1 indicating that knowledge of either one of the values completely predicts the value of the other and the value 0 indicating that X and Y are independent.

$$SU(Y, X) = 2 \frac{IG}{H(Y) + H(X)} \qquad (5)$$

D. Chi-Squared

Feature Selection via chi-squared (χ^2) test is another, very commonly used method [20]. The χ^2 method evaluates features individually by measuring their chi-squared statistic with respect to the class. The initial hypothesis H_0 is the assumption that the two features are unrelated, and it is tested by chi-squared formula:

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \qquad (6)$$

where O_{ij} is the observed frequency and E_{ij} is the expected (theoretical) frequency, asserted by the null hypothesis. The greater the value of χ^2, the greater the evidence against the hypothesis H_0 is.

E. One-R

One-R (OR) is a simple algorithm proposed by Holte [21]. It builds one rule for each attribute in the training data and then selects the rule with the smallest error. It works with continuous values and thus uses a straightforward method to divide the range of values into several disjoint intervals.

This is one of the most primitive schemes. It produces simple rules based on one feature only. Although it is a minimal form of classifier, it can be useful for determining a baseline performance as a benchmark for other learning schemes.

F. Relief-F

Relief-F (RF) attribute evaluation a simple yet efficient procedure to estimate the quality of attributes in problems with strong dependencies between attributes. The key idea of the Relief-F is to estimate the quality of attributes according to how well their values distinguish between instances that are near to each other.

The weight computation is based on the probability of the nearest neighbors from two different classes having different values for a feature and the probability of two nearest neighbors of the same class having the same value of the feature. The higher the difference between these two probabilities, the more significant is the feature.

A. *Validation of Ranking Technique*

In order to validate the results for each of the ranking algorithm, four commonly used supervised learning algorithms are adopted. These are, namely, IB1, Naive Bayes, C4.5 decision tree and the radial basis function (RBF) network.

IB1 is a nearest neighbour classifier. It uses normalized Euclidean distance to find the training instance closest to the given test instance, and predicts the same class as the training distance. The advantage of IB1 is that they are able to learn quickly from a very small dataset. Naïve Bayes is a simple probabilistic classifier based on the elementary Bayes, Theorem. The advantage of Naive Bayes classifier is that it requires a small amount of training data to estimate the parameters (means and variances of the variables) necessary for classification. C4.5 is an algorithm used to generate a decision tree using the concept of information entropy. C4.5 decision tree has various advantages. It is simple to understand and interpret, requires little data preparation, is robust, and performs well with large data in a short time. Radial Basis Function (RBF) network is an artificial neural network that uses radial basis functions as activation functions. It has many uses, including function approximation, time series prediction, classification, and system control [22]. RBF network offers a number of advantages, including requiring less formal statistical training, ability to implicitly detect complex nonlinear relationships between dependent and independent variables, ability to detect all possible interactions between predictor variables, and the availability of multiple training algorithms.

IV. ASSOCIATION RULE MINING

Association mining is one of the most well studied methods in data mining [24] [25]. It has served as a useful tool for discovering correlated items in a large transactional database. It produces if-then statements concerning attribute-values. An association rule $X \Rightarrow Y$ expresses that in those transactions in the database where X occurs; there is a high probability of having Y as well. X and Y are called the antecedent and consequent of the rule. The strength of such a rule is measured by its support and confidence. The support of the rule is the percentage of transactions in the database that contain both the antecedent and the consequent. The confidence of the rule is the percentage of transactions with X in the database that contain the consequent Y also.

Association rule mining has been applied to e-learning systems for traditional association analysis such as building recommender agents for on-line learning activities, automatically guiding the learner's activities and intelligently generate and recommend learning materials, finding out the relationships between each pattern of learner's behavior, identifying attributes characterizing patterns of performance disparity between various groups of students.

A problem experienced with association rule mining algorithms traditionally was that it normally discovers a huge quantity of rules and thus do not guarantee that all the rules found are relevant. With the incorporation of the aforementioned feature selection technique, the rules generated are more specific and demonstrate improved confidence.

V. EXPERIMENTAL RESULTS AND DISCUSSIONS

To validate the ML approach to courseware evaluation, real dataset from a large university-wide course, UU100, was selected for experimentation. Six (6) different feature ranking and feature selection methods were used on the specified dataset to obtain the ranking of the courseware activities. Being able to correctly rank the activities is central to our research objective. Thus, to validate the ranking, four (4) widely used ML algorithms were used to confirm the rationalization of the selected courseware activities and how it significantly contributed towards student success in the course.

A. *UU100 Dataset*

UU100, *Communication and Information Literacy*, was delivered in Blended Mode (20% face-to-face and 80% Online) via Moodle LMS platform in Semester 2, 2012. The course had 2,172 students enrolled and incorporated many online activities such as weekly discussion forums, glossary, resources, quizzes, chat, blog and surveys, etc. Moodle, the LMS used by the university, logged every click that students make for navigational purposes and has a modest log viewing system built into it (Fig. 2).

Figure 2. A screenshot of the Moodle log report for UU100.

The log report within the LMS show summarized student statistics; however, there is still provision for further evaluation of these data to ascertain undiscovered knowledge. Activity reports for all students are available and details about each module (last access, number of times read, etc.) as well as the detailed involvement are clearly highlighted. However, the mere raw interaction counts do not suggest the effectiveness of the courseware within the course.

To derive usefulness from the user log and to extract the behavioral data embedded within it, some of the attribute values were aggregated. Table I lists the 26 different attributes extracted/aggregated for each student from the log table of the course.

TABLE I
SELECTED/AGGREGATED ATTRIBUTES OF UU100

Attribute	Description
LoginFreq	Login Frequency
AssignSubmitted	Num. of Assignment Submitted
DistAssignSubmittd	Num. of Distinct Assignments submitted
ForumViews	Num. of Forums Read
DistForumViews	Num. of Distinct Forums Read
ForumPosts	Num. of Forum Posting
DistForumPosts	Num. of Distinct Forum Postings
QuizStarted	Num. of Standard Quizzes Attempted
QuizCompleted	Num. of Standard Quizzes Completed
DistQuizAttempt	Num. of Distinct Standard Quizzes
NumRQuizStarted	Num. of Review Quizzes Attempted
NumRQuizPassed	Num. of Review Quizzes Passed
NumRQuizFailed	Num. of Review Quizzes Failed
DistRQuizPassed	Num. of Review Quizzes Passed
AvgRQuizScore	Average Review Quiz Score
ResourceViews	Num. of Resources Viewed
DistResourceViews	Num. of Distinct Resources Viewed
BlogViews	Num. of Blog/Wiki Participation
BookViews	Num. of Book Views

The course employed around 11 different online activities. Participation in some of these activities contributed towards the grade in the course. The pre-processed dataset contained 2172 instances, 19 features, and there were no missing values.

The accumulated data particularly aimed to provide course designers an insight on how interaction with specific courseware activities affected the students' performance in the course (pass or fail). Evaluating this context would mean identifying the significant and/or insignificant courseware activities towards this classification. We employed the feature ranking and feature selection techniques to rank the 19 different features available to us.

TABLE II
RESULTS OF RANKING METHOD ON UU100 DATASET

UU100 Data - Ref	IG	IGR	SU	CS	OR	RF	MV
1- LoginFreq	1	4	3	1	3	3	3
2- AssignSubmitted	2	2	2	2	1	16	2
3- DistAssignSubmittd	3	1	1	3	2	1	1
4- ForumViews	13	15	14	14	9	17	14
5- DistForumViews	9	13	13	9	10	9	13
6- ForumPosts	15	14	15	15	11	8	15
7- DistForumPosts	16	16	16	16	14	4	16
8- QuizStarted	8	6	6	8	8	13	8
9- QuizCompleted	7	7	8	7	6	12	7
10- DistQuizAttempt	5	8	7	5	5	2	5
11- NumRQuizStarted	11	11	11	12	17	14	11
12- NumRQuizPassed	10	12	12	13	16	15	12
13- NumRQuizFailed	17	17	17	17	15	10	17
14- DistRQuizPassed	12	10	10	10	13	5	10
15- AvgRQuizScore	14	9	9	11	12	6	9
16- ResourceViews	6	5	5	6	7	11	6
17- DistResourceViews	4	3	4	4	4	7	4
18- BlogViews	18	18	18	18	18	18	18
19- BookViews	19	19	19	19	19	19	19

Table II displays the results of six different ranking methods. The different ranking methods yielded slightly varied ranks based on their evaluation measures. To reach a consensus, Majority Voting scheme (MV) was applied to the ranking provided by the six methods. This produced a final ranking which was more so like a resolution reached by a "panel of judges". ML algorithms were then employed to validate the results of MV, measuring the value-added by each feature towards the student's ability to succeed in the course. The value-added is synonymous to increase in prediction accuracy in this context. The entire process uses the concept of backward elimination. Fig. 3-6 outlines the accuracy of the feature selection technique as insignificant features were eliminated.

The accuracy is measured as a F_1 score (also known as *F-score* or *F-measure*). It considers both the precision *p* and the recall *r* to eliminate the biasness of imbalanced data. F_1 score can be interpreted as a weighted average of the precision and recall, where an F_1 score reaches its best value at 1 and worst score at 0. It gives the harmonic mean and is computed as:

$$F_1 = 2 \times \frac{precision \times recall}{precision + recall} \qquad (7)$$

The tests were conducted with 10 fold cross-validation to reduce variability and randomness. The results from Fig. 3-6 show the improvement in prediction accuracy as the lowly ranked features are eliminated as per the ranking of MV. This is evident from the relatively good prediction accuracy (F_1 score) achieved by three of the four algorithms (namely, Naïve Bayes, C4.5, RBF Network). For all these three classifiers, dropping off the lowly ranked features improves the F_1-Score to a point where the knowledge in the features left was insufficient to predict results accurately (deteriorates performance). Results from IB1 classifier on the other hand was slightly contrasting but very much due to its internal classification algorithm.

To identify the ultimate elimination point for the insignificant features, we compute the average F_1 Score of the four classifiers. This resolves the variability amongst the classifiers (Fig. 7). Another benefit of averaging the score is to ensure that our majority voting scheme (MV) is not biased towards a particular classifier, but rather produces acceptable results for all given classifiers.
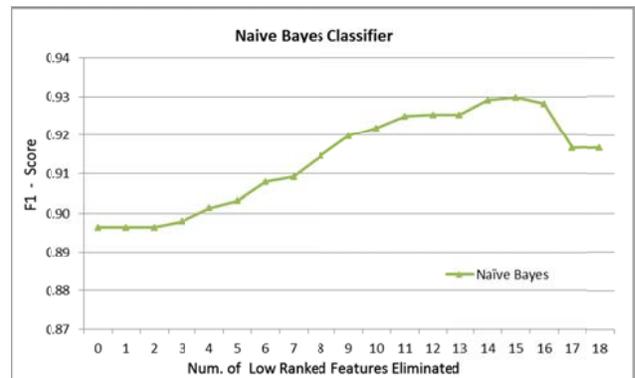
Figure 3. Ranking method MV and F_1 Score using Naïve Bayes Classifier.
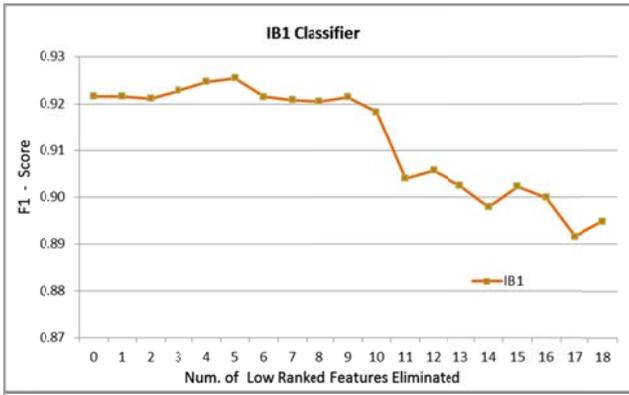
Figure 4. Ranking method VT and F_1 Score using IB1 Classifier.
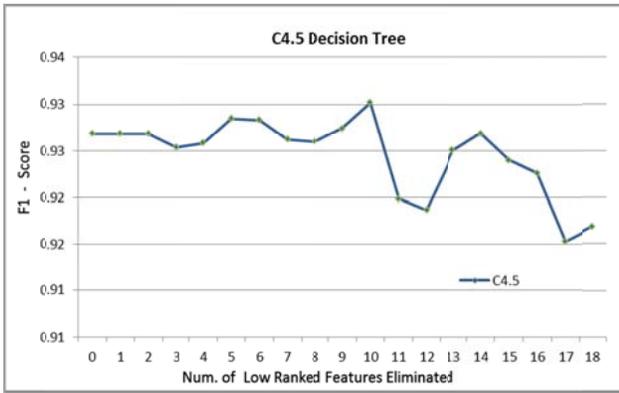
Figure 5. Ranking method MV and F_1 Score using C4.5 Decision Tree.
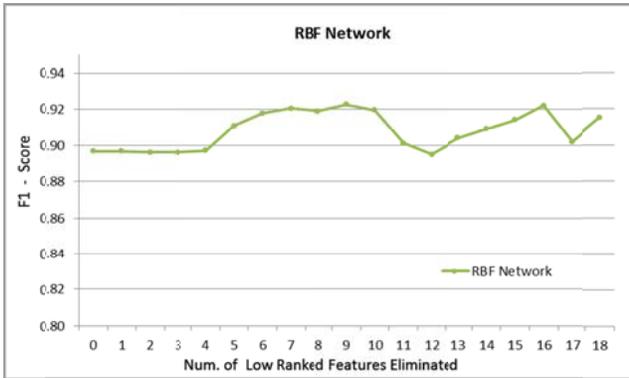
Figure 6. Ranking method MV and F_1 Score using RBF Network.

From Fig. 7, it can be noted that the first 10 ranked features contribute significantly towards improving the prediction performance (that is, after eliminating the 9 lower ranked attributes). The instructional designer can thus infer that Login Frequency, Assignments, Standard Quizzes and Resources were effective courseware items in the course. On the contrary, interaction with activities such as Review Quizzes, Discussion Forums, Blogs and Book modules within the LMS was not as successful in improving the overall student performance in the course. In future, these courseware activities may require re-examination of its effectiveness within the course.
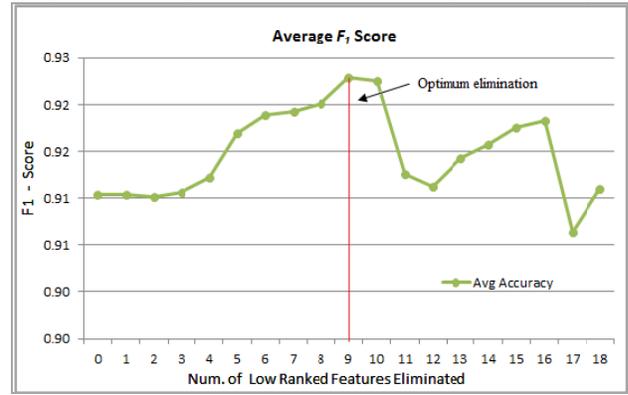
Figure 7. Ranking method MV and F_1 Score using Average F_1 Score.

Next, we apply the association rule mining algorithm to discover valuable patterns and relationship amongst the features. Only the top 10 ranked features are considered for this purpose to eliminate the common problem of association rule mining where the algorithm discovers a huge quantity of rules and rules that are found to be irrelevant.

The data was analyzed using WEKA. WEKA, the Waikato Environment for Knowledge Analysis, is a collection of machine learning algorithm to analyze the data set for data mining tasks [23]. Association rule mining is very useful for educational objectives because it presents significant relationships between the activities within the course [16]. Apriori, algorithm has been used for this research and evaluated with a minimum support of 0.2 and a minimum confidence of 0.9. All the continuous attributes have been transformed into discrete attributes that can be treated as categorical attributes. Discretization allows the numerical data to be divided into categorical classes that are easier for the instructor to understand.

Table III shows some examples of rules that were mined using the Association Rule Mining (ARM) algorithm. For each rule, we show the antecedent and the consequent constructed, as well as the confidence measures.

TABLE III
RULES EXTRACTED USING APRIORI-FREQUENT
ALGORITHM

Rule	Antecedent	Consequent	Conf
1	DistAssignSubmitted=MEDIUM, DistResourceViews=HIGH, DistQuizAttempt=HIGH, AvgRQuizScore=HIGH	Class =Pass	1
2	ResourceViews=HIGH, LoginFreq =MEDIUM	Class=Pass	0.92
3	DistAssignSubmitted=MEDIUM	Class=Pass	0.97
4	DistQuizAttempt=HIGH	Class=Pass	0.96
5	AssignSubmitted=LOW DistQuizAttempt=HIGH ResourceViews=LOW QuizCompleted=LOW	Class=Pass	0.96

As it can be noticed in Table III, the rules are very precise and very easy to interpret. The confidences of the rules are

also quite high. Rule 4, for example, states that having a high value for *DistQuizAttempts* is a behaviour that could significantly contribute towards achieving a pass in the course. Similar proposition is also demonstrated by Rule 5. It also unveils that not all the highly significant activities have to be completed in order to achieve success in the course. For instance, if a student is lacking in certain activities, being able to identify the correct supplementing activity can assist in achieving success. Recommendation system platforms can be based on this notion. Comparison of rules and their confidence reveals interesting knowledge. For instance, comparing Rule 3 and 4 suggests that distinct Assignment activities are more effective in the course compared to the Quiz activities. This inference is also supported in Table I by the MV scheme.

The author heralds the fact that the use of Machine Learning tools can be particularly useful in developing new and revising existing online courses. The approach is able to provide valuable insight in the effective practices within the course. It also means that the online course development and revision will no longer be left to chance and philosophical beliefs but rather based on proven performances. Instructional Designers during the planning process in course development can be much better informed when developing similar courses for optimized delivery. Ranking of courseware items can also serve a platform for future recommender system development.

VI. CONCLUSION

In this paper, the effectiveness of ML technique in evaluating and improving curriculum design is discussed. With the abundance of data now available at our disposal and the emerging shift towards MOOCs and mobile learning means employing such approaches will become critical in terms of conceptualizing and delivering the right type of instructions and activities to the students.

The choice of activities and assessments will differ from course to course and discipline to discipline. However, the proposed method of using user generated data (as used in other matured domains) will play a pivotal role in educational institutions to validate the curriculum and assessment design and lay the groundwork for a more effective learning process for its diverse audience.

REFERENCES

[1] M. Negnevitshy, Artificial Intelligence: A Guide to Intelligent Systems, 2nd ed., Addison Wesley, 2005.

[2] J. Holland, Adaptation in Natural Artificial Systems, University of Michigan Press, 2nd ed., MIT Press, 1992.

[3] P.S. Saini, D. Sona, S. Veeramachaneni, M. Ronchetti, "Making E-Learning Better Through Machine Learning", *Int. Conf. on Methods and Technologies for Learning (ICMTL)*, 2005.

[4] B. Minae-bidgoli, D.A. Kashy, G. Kortmeyer, W.F. Punch, "Predicting student performance: an application of data mining methods with an educational Web-based system", *In International Conference on Frontiers in Education*, 2003, pp.13-18.

[5] Z. Ibrahim, D. Rusli, "Predicting students' academic performance: comparing artificial neural network, decision tree and linear regression", *In Annual SAS Malaysia Forum*, Kuala Lumpur, 2007, pp. 1-6.

[6] T. Y. Tang, G. Mccalla, "Student modeling for a web-based learning environment: a data mining approach" *In Conference on Artificial intelligence*, Edmonton, Canada, 2002, pp. 967-968.

[7] V.O. Oladokun, A.T. Adebanjo, O.E. Charles-owaba, "Predicting student's academic performance using artificial neural network: A case study of an engineering course", In Pacific Journal of Science and Technology, vol. 9, no. 1, pp. 72-79, 2008.

[8] F. Abel, I. I. Bittencourt, N. Henze, D. Krause, J. Vassileva, "A Rule-Based Recommender System for Online Discussion Forums.", *In International Conference on Adaptive Hypermedia and Adaptive Web-Based Systems*, Hannover, Germany, 2008, pp. 12-21.

[9] G.W. Dekker, M. Pechenizkiy, J.M. Vleeshouwers, "Predicting Students Drop Out: A Case Study.", *In International Conference on Educational Data Mining, Cordoba,* Spain, 2009, pp. 41-50.

[10] S. Kotsiantis, C. Pierrakeas, P. Pintelas, "Preventing student dropout in distance learning systems using machine learning techniques", *In International Conference on Knowledge-Based Intelligent Information & Engineering Systems*, Oxford, 2003, pp. 3-5.

[11] C. Romero, S. Ventura, E. Salcines, "Data mining in course management systems: Moodle case study and tutorial.", Computer &Education, vol. 51, no. 1, pp.368-384, 2008.

[12] S. Watanabe, A. Jokhan and B. Sharma, "Use of ICTs in Online Distance Education." In S. Barton et al. (Eds.), Proceedings of Global Learn 2011, AACE., pp. 949-954, 2011.

[13] S. Singh, A. Jokhan, B. Sharma and S . Lal "An innovative approach of progressive feedback via artificial neural networks", Journal of Machine Learning Technologies, ISSN: 2229-3981 & ISSN: 2229-399X, vol. 2, no. 2, pp-64-71, 2011.

[14] J. Novaković, P. ŠTRBAC, and D. Bulatović. "Toward optimal feature selection using ranking methods and classification algorithms." The Yugoslav Journal of Operations Research ISSN: 0354-0243 EISSN: 2334-6043 21, no. 1, pp. 119-135, 2011.

[15] M. Ben-Bassat, *Pattern recognition and reduction of dimensionality*, in: P. R. Krishnaiah and L. N. Kanal, (eds), Handbook of Statistics-II, North Holland, 1982, pp. 773–791.

[16] P. Mitra, C. A. Murthy, and S. K. Pal, "Unsupervised feature selection using feature similarity", IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 3, pp. 301–312, 2002.

[17] J.G. Dy and C.E. Brodley, "Feature subset selection and order identification for unsupervised learning", *in: Proceedings of the Seventeenth International Conference on Machine Learning*, 2000, pp. 247–254.

[18] L. Talavera,"Feature selection as a preprocessing step for hierarchical clustering", *in: Proceedings of Internationl Conference on Machine Learning (ICML'99)*, 1999.

[19] N. Abe and M. Kudo, "Entropy criterion for classifier-independent feature selection", Lecture Notes in Computer Science, vol. 3684 , pp. 689-695 , 2005.

[20] H. Liu, and R. Setiono, "Chi2: Feature selection and discretization of numeric attributes", *Proc. IEEE 7th International Conference on Tools with Artificial Intelligence*, 1995, pp. 338-391.

[21] R.C. Holte, "Very simple classification rules perform well on most commonly used datasets", Machine Learning, vol. 11, pp. 63-91, 1993.

[22] M. J. L. Orr, "Recent advances in radial basis function networks." , Technical Report for Institute for Adaptive and Neural Computation Division for Informatics, Edinburgh University, Edinburgh UK, 1999.

[23] M. Zdravko and I. Russell, "An introduction to the WEKA data mining system." In ACM SIGCSE Bulletin, vol. 38, no. 3, pp. 367-368, 2006.

[24] R. Agrawal, T. Imielinski, A.N. Swami, "Mining Association Rules between Sets of Items in Large Databases". *In: Proc. of SIGMOD*, 1993, pp. 207-216.

[25] S. Kotsiantis, D. Kanellopoulus, "Association rules mining: A recent overview", International Transactions on Computer Science and Engineering Journal, vol. 32, no. 1, pp. 71-82, 2006.