2013 IEEE Conference on e-Learning, e-Management and e-Services, December 2 - 4, 2013, Sarawak, Malaysia

Educational Courseware Evaluation Using Machine Learning Techniques

Shaveen Singh
School of Computing, Information & Mathematical Science
University of the South Pacific
Suva, Fiji Islands
e-mail: singh_sv@usp.ac.fj

Sunil Pranit Lal
School of Computing, Information & Mathematical Science
University of the South Pacific
Suva, Fiji Islands
e-mail: lal.sunil@ieee.org

Abstract— **With the introduction of massive open online courses (MOOCs) and other web-based learning management systems (LMS), there is a greater need to develop methods for exploring the unique types of data that come from the educational context. This paper highlights the advantage of using Machine Learning (ML) as an e-planning tool to enhance learning and improve courseware development. Researchers generally consider student evaluation survey on courses to be highly reliable and at least moderately valid on courseware evaluation. However, low response rate, retaliation, grades and comparison with past instructors sometimes affects the reliability of the result. ML algorithms has been deployed in this paper to intelligently examine the interaction log data from the LMS to obtain a predictive map that permits mapping the online interaction behaviour of students with their course outcome. These predictive relationships are then investigated and ranked using various ML algorithms to evaluate and validate the various learning tools and activities, and their effectiveness within the course.**

Keywords— Artificial Intelligence; e-learning; attribute ranking; machine learning; online development.

I. INTRODUCTION

Machine Learning (ML) is a branch of artificial intelligence that is invariably concerned with the design and development of algorithms that allow computers to evolve behaviors and generate rules based on empirical data [1]. The major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on past and/or present data. Therefore, the accomplishments of the ML techniques are primarily based on its ability to derive predictive patterns from a set of features [2] and guide people in decision making.

In the recent past, there has been a significant growth in the use of the ML techniques in the area of higher education. Functionalities such as: student performances [3-5, 7], student modeling [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.

MOOCs are on the rise with universities exploiting this platform to provide open access to their courses. Since the e-learning systems involve acquisition and storage of large volumes of data, which most of the time need to be handled simultaneously, the deployment of ML tools are seen as a more effective tool for the analysis of these immense educational data repositories. What MLs also brings to seismic interpretation, is the ability to study multiple attributes and their relationships at the same time, a problem that can be too complicated for the human brain.

The objective of this paper is to use the knowledge of student interaction log from a LMS to identify courseware activities that greatly aid in fulfilling the learning outcome of a course. The results are validated using ML techniques. Instructional designers can use this finding when designing new courses, improving the structure of existing courses, appropriately weighting assessment tasks and also use the data to provide guidance to students to improve their learning experience.

A. 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 of teaching and learning [12].

While, the blended and fully online modes introduced in USP are seen as being cost-effective, scalable, innovative, student centered, flexible, and attracting huge student enrolments, a couple of problems they inherently face are the lack of self-motivation and inconsistent participation. This results in an excessive number of unfinished tasks late in the semester and as a consequence, many students fail the online course. This has called for an immediate need to evaluate and validate the curriculum and assessment design to lay the groundwork for a more effective learning process at USP. 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 throughout the semester as they add-value to the course while the not so significant activities can be re-examined for its suitability for the context and effectiveness for the specific discipline or course.

II. Feature Selection Approach to Determine Significant Activities

Feature selection is a fundamental problem in many different domains, ranging from bioinformatics, forecasting, document classification, object recognition and in modeling of complex technological processes [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 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]) ranging from quizzes, lessons, forums, blogs, assignments, glossaries, surveys, wiki, workshops and so on. The challenge for any instructor designing a new course or improving an existing course is to identify the correct mix of learning activities that can assist in attaining the expected learning outcome of the course. A simple rule of thumb (on the choice of LMS activities and assessments) may not be suitable for all online courses or courses within a specific discipline. Another important factor to note is that, suitability of courseware activities and assessment types may vary for courses at different levels within a programme.

In the context of our university, another common problem that has to be considered is the slower dial-up speeds at home for those rural students enrolled in online courses from different regional campuses in different countries. 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 an environment. The feature selection approach can also drastically reduce the spectrum of activities available in the LMS and help identify redundant, irrelevant, or insignificant courseware which may be adding a lot to student workload but may not necessary contribute much towards the learning process and experience of the students.

A. Feature Selection Process in Machine Learning

The process of feature selection [14] in ML consists of the following four steps:

- subset generation,
- subset evaluation,
- stopping criterion and
- result validation.

The aim is to create a subset of features, evaluate it using a given classification or correlation notion, and loop until an ending criterion is satisfied. Finally, the subset found is validated by the classifier algorithm using some unseen data set.

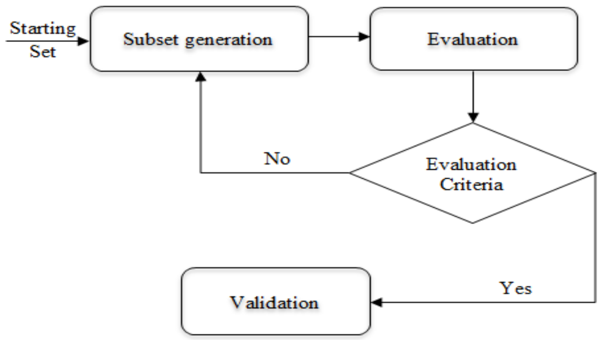

Figure 1. The Feature Selection Process

III. Feature Ranking and Selection

A number of feature selection techniques have been highlighted in the machine learning literature [13, 16 and 17]. The primary purpose of feature selection is to discard any irrelevant or redundant features from a given feature vector.

For the purpose of this experiment, all the four steps of the feature selection process in Fig. 1 was followed. To evaluate, each subset, the following commonly used statistical and entropy-based methods were used.

- Information Gain (IG),
- Gain Ratio (IGR),
- Symmetrical Uncertainty (SU),
- Relief-F (RF),
- Chi-Squared (CS).

All the above mentioned methods have been selected based on their good performance in various domains.

Entropy based methods are commonly used in the information theory [19] and is a measure of how "mixed up" an attribute is. It is sometimes equated to the purity or impurity of a variable. It is the foundation of the IG, IGR, and SU ranking methods. The entropy of Y is computed as:

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

where *p(y)* is the marginal probability density function for the random variable *Y*. Features or variables are not mutually exclusive in all situations. If a relationship exists between features *Y* and *X*, the entropy of *Y* after observing *X* is then calculated by:

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

where *p(y|x)* is the conditional probability of *y* given *x*.

A. Information Gain

Information Gain (IG) is a measure the uncertainty associated with a random feature *X* and is computed as:

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

where H is the information entropy. Although IG is commonly used, it is biased towards tests with many outcomes (attributes having a large number of values).

B. *Gain Ratio*

The Gain Ration (GR) overcomes the bias of IG by measuring the IG with respect to the class.

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

C. *Symmetrical Uncertainty*

Symmetrical Uncertainty (SU) evaluates the worth of an attribute with respect to the class and is good if the attribute highly correlates to the class but not with the other attributes. It is computed as:

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

D. *Chi-Squared*

Chi-Squared statistic is an overall measure of how close the observed frequencies are to the expected frequencies and is computed with the following formula:

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

where O_{ij} is the observed frequency and E_{ij} is the expected (theoretical) frequency.

E. *Relief-F*

Relief-F is a simple measure to estimate the quality of attributes in problems with strong dependencies between attributes. It 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. *Identifying the Final Ranks*

Different ranking methods incorporate different internal computation measures are thus are suited to different data types and data distribution. Hence, each ranking method mentioned above may generate a different rank for each feature within the feature vector. In order to resolve the ranking, we employ a "panel of judges" approach in our experiment. The "panel of judges" approaches uses a majority voting (MV) scheme to ascertain the final list of ranked features using the rank commonality of each algorithm for each ranked attribute. The MV scheme is demonstrated the Fig. 2.

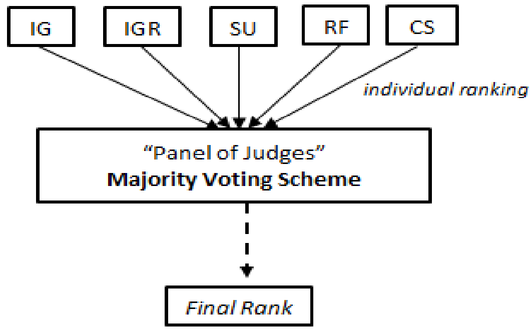

Figure 2. Majority Voting (MV) scheme to resolve ranking

B. *Validation of Ranked Features*

In order to validate the results 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 that uses normalized Euclidean distance to find the training instance closest to the given test instance, and predicts the same class as the training distance. Naïve Bayes, on the other hand, 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 classification tree using the concept of information entropy. 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.

All the above mentioned classification algorithms have been selected based on their good performance in various domains.

IV. EXPERIMENTAL RESULTS AND DISCUSSIONS

In this experimentation, real student dataset from the offering of a university-wide course at USP, UU100, was selected. To validate the ML approach to courseware evaluation, five (5) different feature ranking methods were employed on the stated dataset. The majority voting (MV) scheme was used to list the resolved ranking from all the algorithms.

The ranking and the significance of the activities was finally validated using the four (4) widely used ML algorithms to confirm the rationalization of the selected courseware activities towards student success in the course.

A. UU100 Dataset

UU100, *Communication and Information Literacy*, was offered 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 logged every click that students make for navigational purposes and has a modest log viewing system built into it (Fig.2).

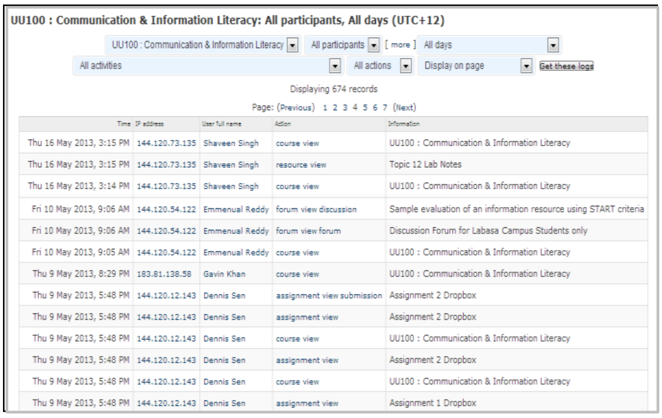

Figure 3. Screenshot of Moodle log report for UU100

The log report within the LMS shows the entry for each user activity; however, this mere raw interaction summary does not provide much meaning for decision making.

To derive usefulness from the user log and to extract the behavioral data embedded within it, the feature values were aggregated. Table I lists the 26 different attributes extracted/aggregated for each student from the log table of the UU100 course.

TABLE I
AGGREGATED ATTRIBUTES - 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 pre-processed dataset consisted of 2172 instances, 19 features, and there were no missing values. The course employed around 11 different online activities and participation in some of these online activities contributed towards the final grade in the course.

Feature ranking provides meaningful insight to course designers on how interaction with specific courseware activities correlated with the students' performance in the course (pass or fail). Evaluating this context would mean identifying the significant and/or insignificant courseware activities. Table II shows the result of feature ranking to rank the19 different features.

TABLE II
RESULTS OF RANKING METHOD ON UU100 DATASET

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

The results show that different ranking methods yield slightly varied ranks based on their internal evaluation measures. ML algorithms were then employed to validate the results measuring the value-added by each feature towards the student's ability to succeed in the course.

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 of the test to eliminate the biasness of any imbalance in the dataset. F_1 score for a given set of features 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 random successes.

The results from Fig. 4 show that eliminating the insignificant features significantly improves the F_1 Score. It is notable in the case of UU100 that removing the bottom 9 ranked features contributed significantly towards improved prediction performance. This suggests that Login Frequency, Assignments, standard Quizzes and Resources were effective and significant courseware items in this 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 and may need to be re-examined for its effectiveness within the course.

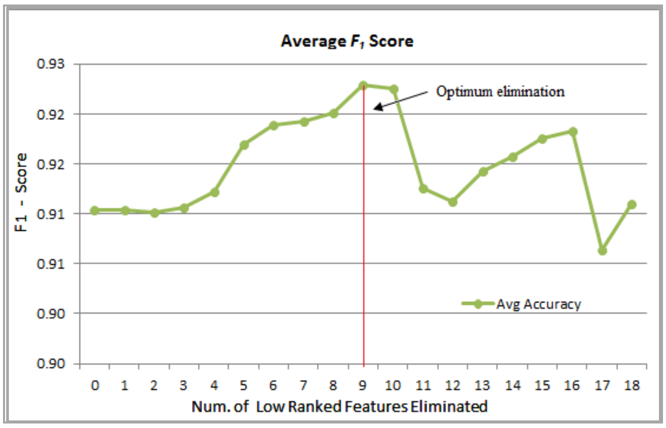

Figure 4. Ranking method MV and F_1 Score using Average F_1 Score

TABLE III
RESULTS OF F_1 SCORE ON TEST DATA SET

Algorithm	**Average F_1 Score**
IB1	0.9198
Naïve Bayes	0.9227
C4.5	0.9215
RBF	0.9274
Average F_1 Score	**0.9228**

Table III validates this result by showing that high F_1 *score* (prediction accuracy) is attained using only the top 10 significant activities in the course. Such information not only serves to inform course coordinators on the effective courseware items but also provides a means to develop an early warning system for identifying at-risk students during the semester. Students who do not seem to perform at a satisfactory level in the significant activities (top 10 ranked) can be identified as at-risk in the course. These students can be provided remedial actions early in the semester before it becomes too late.

Detailed analysis can also be done on courseware by doing a further analysis within each category of interactions to identify specific effective and ineffective assessments tasks and resources items.

This is demonstrated using two courseware modules, namely *Discussion Forums* and *Review Quizzes* from the same course. The course consisted of 5 assessed review quizzes throughout the semester.

TABLE IV
RANKING INDIVIDUAL REVIEW QUIZZES

Activity	Ranking (based *on Attempts*)	Ranking *(based on Score)*
Review Quiz 1	5	5
Review Quiz 2	4	4
Review Quiz 3	3	3
Review Quiz 4	1	1
Review Quiz 5	2	2

Review Quiz in Table II was identified an ineffective courseware activity despite being used by the coordinator as a critical coursework item. Results from Table IV show that of the five review quizzes, Review Quiz 1 & 2 ranked poorly. This means that regardless of whether the student has attempted the quiz or not and regardless of the score attained, it did not contribute as highly towards the students' performance in the course compared to the other four quizzes. On the contrary, Review Quiz 4 seemed to be the most effective Review Quiz assessment.

Similarly, Table IV shows the ranking of the 21 different discussion forums existing within the course. Discussion forum participation was identified as an effective courseware item in Table II. Table IV shows that of all the discussion forums, participation in "assessment specific" forums has been more effective in guiding students to succeed in the course amongst the listed forums. This knowledge identified during the process can be emphasized in consequent offerings of the course as recommendation for students or to identify at-risk students.

TABLE V
RANKING INDIVIDUAL DISCUSSION FORUMS

Forum Name	Rank (*based on participation*)
Assignment 2 Help! Forum	1
Assessed Task: Netiquette	2
Assessed Task: Digital Divide	3
Week 4 Experiences	4
EPortfolio Discussion Forum	5
News & Announcement Forum	6
Week 2 Experiences	7
Week 6 Experiences	8
Week 3 Experiences	9
Week 5 Experiences	10
Week 9 Experiences	11
Assignment 1 Help! Forum	12
Week 10 Experiences	13
Week 8 Experiences	14
Week 13 Experiences	15
Week 1 Experiences	16
Week 11 Experiences	17
Week 7 Experiences	18
Week 12 Experiences	19
Social Cafe for Students	20
Week 14 Experiences	21

The research supports that this e-planning approach of using ML techniques can be particularly useful in developing new and improving existing online courseware. The approach is able to provide valuable insight about the effective practices within the course with realistic date from the students. 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. Identifying of significant courseware activities can also aid in identifying and supporting at-risk students in the course.

V. 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.

As demonstrated in the paper, the use of user generated data can not only aid in identifying the significant courseware activities but also provide a means of improving and validating the courseware and supporting personalized student learning in the evolving e-learning landscape. This data-driven approach ensures decision making on courseware is not left to chance or philosophical beliefs but rather based on proven performances.

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.