International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:10, No:1, 2016

Questions Categorization in E-Learning Environment Using Data Mining Technique

Vilas P. Mahatme, K. K. Bhoyar

***Abstract*—**Nowadays, education cannot be imagined without digital technologies. It broadens the horizons of teaching learning processes. Several universities are offering online courses. For evaluation purpose, e-examination systems are being widely adopted in academic environments. Multiple-choice tests are extremely popular. Moving away from traditional examinations to e-examination, Moodle as Learning Management Systems (LMS) is being used. Moodle logs every click that students make for attempting and navigational purposes in e-examination. Data mining has been applied in various domains including retail sales, bioinformatics. In recent years, there has been increasing interest in the use of data mining in e-learning environment. It has been applied to discover, extract, and evaluate parameters related to student's learning performance. The combination of data mining and e-learning is still in its babyhood. Log data generated by the students during online examination can be used to discover knowledge with the help of data mining techniques. In web based applications, number of right and wrong answers of the test result is not sufficient to assess and evaluate the student's performance. So, assessment techniques must be intelligent enough. If student cannot answer the question asked by the instructor then some easier question can be asked. Otherwise, more difficult question can be post on similar topic. To do so, it is necessary to identify difficulty level of the questions. Proposed work concentrate on the same issue. Data mining techniques in specific clustering is used in this work. This method decide difficulty levels of the question and categories them as tough, easy or moderate and later this will be served to the desire students based on their performance. Proposed experiment categories the question set and also group the students based on their performance in examination. This will help the instructor to guide the students more specifically. In short mined knowledge helps to support, guide, facilitate and enhance learning as a whole.

***Keywords*—**Data mining, e-examination, e-learning, moodle.

I. Introduction

IN present world, e-learning gained a great deal of importance in education. The world is going fast towards online learning. Universities are providing online courses. Student can study the course, write the examination and get certified. Instructors and students are very important stake holders of education system. It consists of instruction and learning. Instructors are trying to reveal new approaches and learning methods to enhance learning. Each student has its own learning style, and as a result that each student's performance in learning cannot be assessed and evaluated in a unique way. Learning occurs in student's mind with cognitive steps [10]. In this regards instructional technologies must be developed. Online examinations systems are being widely adopted in academic environments. Among these, multiple-choice tests are very much popular. Student's performance cannot be evaluated and assessed in a web based application only by considering the test results depending on the number of right and wrong answers. Hence, testing must be intelligent enough. It should behave as if an instructor asks questions to a student in real class environment. If student cannot answer the question, instructor must ask easier question about similar topic and if student answer this, then, again more difficult question can be asked. In this regard, the question levels must be determined. To address this issue the use of clustering technique is suggested. The proposed method decides the level of difficulty of questions and to categorize and group the questions as tough, easy or moderate.

Vilas P. Mahatme is with the Department of Computer Technology, Kavikulguru Institute of Technology & Science, Ramtek, Dist-Nagpur, (MS) India (e-mail: mahatme.vilas@gmail.com).

K. K. Bhoyar is with the Department of Information Technology, YeshwantraoChavanCollege of Engineering, Nagpur, (MS) India.

II. Related Work

The research interest in using data mining in e-learning is constantly rising. According to F. J. Liu, B.J. Shih, the discovered knowledge can be used to better understand students' behavior to access student's learning style [3]. J. Mamcenko et al. proposed solution to improve the examination system. They showed that behavior of each student is different. Each test retaking decreases the mean examination time duration and increases mean of number of correct answer [4]. Gennaro Costagliola presented an approach where instructors monitor learner behavior and test quality. They focused on the discovery of behavioral patterns of learners and conceptual relationships among test items. They provided a tool that reviews the whole assessment process and evaluates possible improvements [5]. Bernardete Ribeiro and Alberto Cardoso emphasized the use of neural networks and support vector machines to build prediction models able to track student's behavior. This will is able to successfully predict students' final outcome while bringing useful feedback during course learning [6]. Yair Levy, Michelle M. Ramim used data analytics techniques to explore the problem of procrastination. The trends uncovered in the study indicate that students who perform their online exams during the morning hours, appear to do better than those during the afternoon hours. It is also concluded that the more the procrastination, lower the online exams score [7]. Application of text mining methodology and algorithms for academic dishonesty detection and evaluation on open-ended college examinations, based on document classification techniques studied in [8]. They proposed classification models for cheating detection by using a decision tree supervised

algorithm. An Intelligent Examination Framework (I-EXAM) is presented by author who helps the teacher to understand students' behavior during multiple choice online examinations by monitoring the interactions data logs [9]. Survey of various clustering algorithms for different data sets appearing in statistics, machine learning and computer science are presented in [15].

III. E-Learning Environments & Learning Management Systems

Moving away from conventional class room learning environments to the e-learning, Learning Management Systems (LMS) enable the instructor and the students to handle the varied study materials in a much easier manner. Moodle is an open-source learning management system to help instructors to create effective online learning communities [14]. Moodle's modular design helps the instructor to create new courses, adding content that will engage students. Logging is record keeping, which keep track of what materials students have accessed. Moodle logs every click that students make for navigational purposes and also has a log viewing system. Log files can be view by course, participant, day and activity. The instructor can use these logs to find out who has been active in the course, what they did, and when they did it. For activities such as quizzes, a detailed analysis of each student's responses item analysis, score and elapsed time are available. Instructors can easily get full report of the activities of individual students, or of all students for a specific activity. Feedback on performance is a critical part and one of the most important activities in a learning environment. Instructor cannot tell what is going on in the minds of students, so it is require a method for them to demonstrate what they understand and what they don't. A well-planned test, like a multiple-choice test, can give critical information about student performance [13].

IV. Data Mining

Data Mining is the process of analyzing data from different perspectives and summarizing the results as useful information. It has been defined as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data [1]. Data mining encompasses different predictive and descriptive algorithms that are varied in their methods and aims.

- Classification - Segregate data into several predefined classes.
- Regression - Predict a value of a given continuously valued variable based on the values of other variables, considering either a linear or nonlinear model of dependency.
- Deviation detection - Discover the most significant changes in data from earlier measured or normative values.
- Clustering - Identify a set of categories that describe the data.
- Summarization - Find a concise description for a subset of data. Tabulating the mean and standard deviations for all fields is a simple example of summarization.
- Dependency modeling - Find a model that describes significant dependencies between variables.

There are some tasks in data mining that have both descriptive and predictive aspects.

- Association Rule Discovery - Produce dependency rules which will guess the occurrence of an item based on patterns found in the data.
- Sequential Pattern Discovery - Find rules that predict strong sequential dependencies among different events.

V. Use of Data Mining in E-Learning

The combination of data mining and e-learning is still in its formative years. Data mining can be used to mine knowledge from e-learning systems through the information available in the form of data generated by their students [11], [12]. It helps to support instructor to improve the e-learning environment. It is an iterative cycle in which the mined knowledge enters the loop of the system and guide, facilitate and enhance learning as a whole. It turns data into knowledge and also filter mined knowledge for decision making [2].

- The e-learning data mining process consists of the same four steps like in the general data mining process
- Data collection - The LMS system is used by students. The usage and interaction information is stored in the database. In this work student's examination data of the moodle system has been used.
- Preprocessing of the data - The data is cleaned and transformed into an appropriate format.
- Application of data mining algorithm - The data mining algorithms applied to build and execute the model to find out and review the knowledge of interest which will be helpful for the user.
- Interpretation - The obtained results are interpreted and used by the instructor for further proceedings.

VI. Research Data and Method

Following steps are involved to carry out the experimentation.

A. Data Gathering and Preprocessing

In this research, online examination data of Artificial Intelligence subject is analyzed. For online examination 48 questions were set. There were 95 students appeared for examination. Maximum time allotted for examination was 60 minutes and maximum marks were 100. Questions are composed of text formatted. They are multiple-choice questions. These types of questions are commonly used in electronic testing environments. If question responses were correct, then they were graded by 1. If question response were wrong, it was graded by 0. If there is no response existed, then it was graded by 0as well. All necessary data for this research work is collected from moodle log file through local moodle server. Collected log data was used for preprocessing. The purpose of the preprocessing stage is to cleanse the data as much as possible and to place it into a form that is suitable for

use in later stages. Knowledge Discovery in Databases (KDD) process used for preprocessing. The sources of selected data are analyzable examination records such as question given to students, chosen answers, total marks secured in examination and so on.

B. Choosing the Right Data Mining Technique

Data mining techniques broadly classified as descriptive and predictive modeling. A descriptive model, clustering is used in this work to identifies patterns or relationship in data. Basically, the goal of data clustering is to discover the natural grouping of a set of unlabeled data. Data clustering has been used for the three main purposes: underlying structure, natural classification and compression. The basic clustering steps are Preprocessing and Feature Selection – It involves choosing an appropriate feature, preprocessing and feature extraction on data items. It is often desirable to choose a subset of all the features available which reduces the dimensionality of the problem space.

- Similarity Measure - A set of objects are grouped into several clusters, so that similar objects will be in the same cluster and dissimilar ones in different cluster.
- Clustering Algorithm - This uses particular similarity measures as subroutines. A clustering algorithm attempts to find natural groups of data based on some similarity and also finds the centroid of a group of data sets.
- Result Validation - If a result does not make sense then iterate back to some prior stage.
- Result Interpretation and Application-Typical applications of clustering include data compression, hypothesis generation, hypothesis testing and prediction.

C. K-Means Algorithm

The K-means algorithm requires three user-specified parameters. These are number of clusters, cluster initialization and distance metric. K-means is typically used with the Euclidean metric for calculating the distance between points and cluster centers. K-means starts with an initial partition with K clusters and assign patterns to clusters to reduce the squared error. Squared errors always decrease with an increase in the number of clusters K. Typically, K-means is run independently for different values of K and the partitions that shows the most significant to the domain expert is selected. Different initializations can lead to different final clustering.

Algorithmic steps for k-means clustering;

Let $X = \{x_1, x_2, x_3, \ldots\ldots, x_n\}$ be the set of data points and $V = \{v_1, v_2, \ldots\ldots, v_c\}$ be the set of centers.

1) Randomly select '*c*' cluster centers.
2) Calculate the distance between each data point and cluster centers.
3) Assign the data point to the cluster center whose distance from the cluster center is minimum of all the cluster centers.
4) Recalculate the new cluster center using

$$V_i = \left(1/C_i\right) \sum_{j=1}^{C_i} X_i \quad (1)$$

where, 'c_i' represents the number of data points in i^{th} cluster.

5) Recalculate the distance between each data point and new obtained cluster centers.
6) If no data point was reassigned then stop, otherwise repeat from step (3).

An aim of this algorithm is at minimizing an objective function, in this case a squared error function. The objective function:

$$J = \sum_{j=1}^{k} \sum_{i=1}^{n} \left\| X_i^{(j)} - C_j \right\|^2 \quad (2)$$

where, $\left\| x_i^{(j)} - c_j \right\|^2$ is a chosen distance measure between a data point $x_i^{(j)}$ and the cluster centre c_j is an indicator of the distance of the n data points from their respective cluster centres.

D. Methodology

An E-Learning database system has thousands of questions in a question pool but the difficulty levels of questions from student's perspective are not determined. In the proposed experiments an attempt has been made to develop a descriptive model of evaluation of online examination. This intelligent question categorization process consists of data pre-processing, question difficulty and clustering. In order to categorize questions into different clusters, difficulty level of each question has been found one by one. Difficulty level is inversely proportional to the number of correct answers of each question. This means that if any question has the smallest amount of correct answer is the hardest question in test. From this point of view, a formula has been generated. That is:

$$\text{Weight} = \text{SCA} / \text{N} \quad (3)$$

where, Weight: Difficulty level of a question, SCA: Sum of marks of correct answers of a questions, N: Total no of students.

Later, questions are categorized based on their difficulty weight parameter. During clustering three clusters were generated. Analyzing cluster statistical data, it is noticed that 48 questions are categorize as 19 easy, 15 moderate and 14 tough questions. Further, by increasing number of k, they are categorized as very easy, easy, moderate, tough and very tough questions.

TABLE I
QUESTIONS CATEGORIZATION

Attribute	Cluster1 Moderate	Cluster2 Tough	Cluster3 Easy
Weight	0.781	0.383	1.419
No. of Questions	19	15	14
Percentage of questions	39.5	31.25	29.16

Based on the secured marks in online examination, students are classified into four groups as grade A+, A, B and C. Further, student performance in different clusters in attempting tough questions and/or moderate and easy

questions is identified. From this, it is noticed that students who secured grade A, B were not done well in cluster 2. There is a need to groom those students and to provide additional input on those particular topics of the subject.

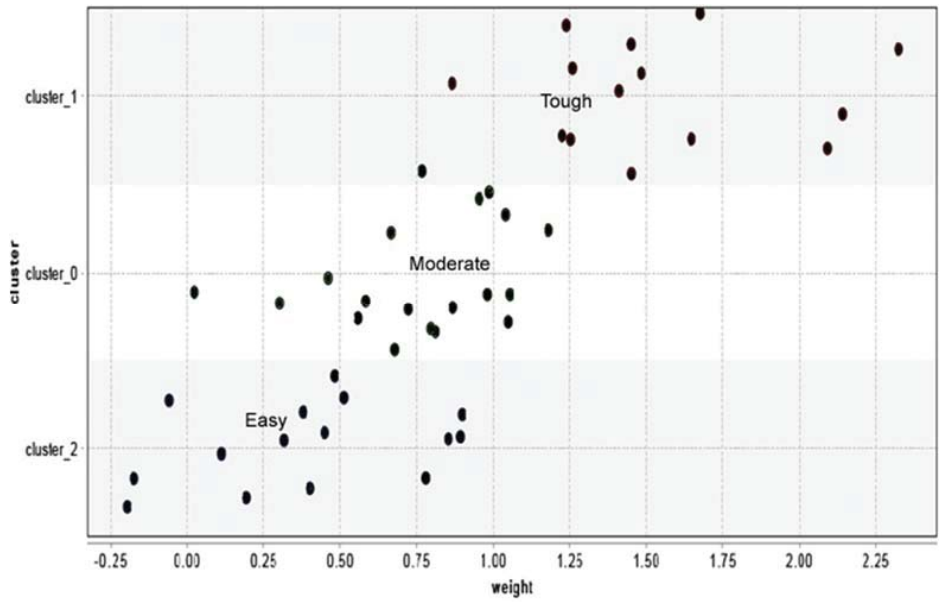

Fig. 1 Clustering

TABLE II
CLUSTERWISE STUDENTS PERFORMANCE

Grade	No of Students	Performance		
		Cluster1 Moderate	*Cluster2 Tough*	*Cluster3 Easy*
A+	28	20.28	6.46	22.50
A	45	14.83	4.89	19.64
B	21	10.49	3.96	17.13
C	01	4.16	6.24	8.32

VII. CONCLUSION

The application of data mining in e-learning systems has specific requirements which are not present in other domains. Data Mining can be used to improve the understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students. Online examinations are being widely used in academic environments to evaluate the student's performance. This introduces an assessment mechanism. This experimental study revealed some facts that "How many students could answer question correctly?". "What students know?", "How difficult was the question?" and further tell about "Which topics, instructor should revise to the students". By this intelligent question categorization, instructors will be able to give different questions to different students. This is the kind of instructional methodology which can develop education quality and efficiency. In this work, the KDD process is followed and its application to the domain of online examination is highlighted. The clustering is used as an aid in the "discovery" of interesting facts of the data. Evaluation of examination found intelligent. If student cannot answer the question, instructor ask easier question about similar topic and if student answer this, then again more difficult question can be asked. In this regard, the question's difficulty level has been determined and categorized the questions set into different clusters as easy, moderate and tough questions. Students are grouped based on their performance in examination and also, contribution of different class of students in various clusters has been shown. Thus, some detail of current state of the research in data mining applied to e-learning has been tried and highlighted its future perspectives and opportunities. If e-learning education system is adopted in full-fledged manner then it is certain that data mining techniques mange it in a better way.

REFERENCES

[1] Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P. and Uthurusamy, "Advances in knowledge discovery and data mining", AAAI Press, 1996.
[2] C. Romero, S. Ventura and E. García, "Data mining in course management systems: moodle case study and tutorial", Computer Education 51(1),2008, pp. 368–384.
[3] F.-J. Liu, B.-J. Shih, "Learning activity based elearning material recommendation system", Ninth, IEEE International Symposium on Multimedia, 2007, pp. 343–348.
[4] J .Mamcenko, I. Sileikiene, J. Lieponiene, R. kulvietiene, "Analysis of e-exam data using data mining techniques", accessed on 19th Aug 2015, http://www.isd.ktu.lt/it2011/material/Proceedings/6_ITTL_3.pdf
[5] Gennaro Costagliola, Vittorio Fuccella, Massimiliano Giordano, and Giuseppe Polese, "Monitoring online tests through data visualization", IEEE Transactions on Knowledge and Data Engineering", vol. 21, no. 6, June 2009.
[6] Bernardete Ribeiro and Alberto Cardoso, "Behavior pattern mining during the evaluation phase in an e-learning course", International Conference on Engineering Education – ICEE 2007", Coimbra, Portugal September 3 – 7, 2007.

[7] Yair Levy, Michelle M. Ramim "A study of online exams procrastination using data analytics techniques" Interdisciplinary Journal of E-Learning and Learning Objects,vol 8, 2012.
[8] Elmano Ramalho, Cavalcanti, Carlos Eduardo Pires, "Detection and evaluation of cheating on college exams using supervised classification", Informatics in Education, vol. 11, no. 2, pp.169–190, 2012.
[9] Essam Kosba, Osama Badawy, Passant Sabri, "Intelligent examination system to support teacher's reflection measurement of student's guided feedback", International Conference on the Future of Education.Egypt.
[10] J. Mamčenko, I. Šileikienė, J. Lieponienė and R. Kulvietienė," Evaluating the data of an e-examination system using a descriptive model in order to identify hidden patterns in students answers", The Online Journal on Computer Science and Information Technology, vol. 1, no.2
[11] C. Romero and S. Ventura, "Educational data mining: a review of the state-of-the-art", IEEE Trans. Syst. Man Cybernet", C Appl. Rev., 40(6),pp. 601–618, 2010.
[12] A. C. Romero and A. S. Ventura, "Educational data mining: A survey from 1995 to 2005", Journal of Expert Systems Applications, 33(1), pp. 135-146, 2007.
[13] J. Gamulin, O. Gamulin, "Enhancing laboratory teaching in higher education environment using web-based formative colloquiums", MIPRO 2011-34thnternational Convention on Information and Communication Technology, Electronics and Microelectronics-Proceedings, art. no. 5967237, pp.1189-1194, 2011.
[14] Moodle, 2013. https://moodle.org/
[15] R. Xu and D. Wunsch, "Survey of clustering algorithms", IEEE Transaction Neural Networks, vol. 16, no. 3, 2005, pp. 645– 678.

Vilas P. Mahatmeis presently working as Associate Professor in the Department of Computer Technology, Kavikulguru Institute of Technology and Science, Ramtek, Dt.Nagpur, (MS) India. He has completed M.Tech. in Computer Engineering from Indian Institute of Technology, Khargpur, India. He is currently pursuing Ph.D degree in Computer Science and Engineering from Rashtrasant Tukadoji Maharaj Nagpur University, Nagpur, India. He is member of IETE, CSI, IE(I) and ISTE. His research interest includes data mining, E-learning and soft computing.

K.K. Bhoyar is presently working as Professor in the Department of Information Technology, Yeshwantrao Chavan College of Engineering, Nagpur, (MS) India. He has completed Ph.D. in Computer Science and Engineering from Visvesvaraya National Institute of Technology, Nagpur, India. He is member of ACM, CSI and ISTE. His research interest includes data mining, soft computing and image processing.