Exploring the Usage of Data Mining and Knowledge Discovery Methodology in e-Learning

Abdel Rahman Alzoubaidi *, Reem Al-Kharouf** and Abdel-Badeeh M. Salem***

*University College of Bahrain, Manama, Kingdom of Bahrain
**Yarmouk University, Yarmouk, Jordan
***Faculty of Computer and Information Sciences, Ain Shams University, Cairo, Egypt

alzoubaidi@gmail.com  alkharoufreem@yahoo.com  abmsalem@yahoo.com

Abstract: - Intelligent data mining and knowledge discovery methodology aims to extract useful information and discover some hidden patterns from huge amount of databases, which statistical approaches cannot discover. It is a multidisciplinary field of research includes: machine learning, databases, statistics, visualization, high performance computing, and knowledge engineering. Intelligent data mining techniques are very efficient tools towards the analysis of the data of student activities and behaviour which accumulated by learning management systems. This paper discusses some of the data mining techniques from the e-learning prospective. The benefits of usage clustering, classification, association rules mining, sequential pattern mining, and text mining techniques in e-Learning tasks are discussed as well.

Key-Words: - Intelligent e-Learning, Data mining, Knowledge discovery, Association rule-based techniques, Clustering, Classification, Machine learning.

1 Introduction

In recent years, the research has extended into usage of machine learning and computational intelligence in developing intelligent e-Learning technology [1, 2, 3, 4, 5]. Researchers have begun to investigate the application of data mining and knowledge discovery methodology to develop a new generation of intelligent e-learning systems [6, 7, 8, 9]. The area of data mining and knowledge discovery is inherently associated with databases. In this sense knowledge discovery significantly augments databases by making them more user friendly and helping people to feel more comfortable dealing with the vast amounts of data and making use of them. Almost at the very inception, databases were found extremely useful. One could easily store records of important transactions, retrieve them without any difficulty and make some decisions based on the facts found in the databases.

Knowledge discovery in databases (KDD) process (see Fig 1) involves the following processes; (a) using the database along with any required selection, preprocessing, sub-sampling, and transformations of it, (b) applying data mining methods (algorithms) to enumerate patterns from it, and (c) evaluating the products of data mining to identify the subset of the enumerated patterns deemed knowledge. The data mining phase of the KDD process is concerned with the algorithmic means by which patterns are extracted and enumerated from data. The overall KDD process includes the evaluation and possible interpretation of the mined patterns to determine which patterns can be considered new knowledge. KDD process is interactive and iterative, involving numerous steps with many decisions made by the user. For more details we refer to the books [10, 11].

Data mining is supported by a host that captures the character of data in several different ways. In what follows, a brief description of the main data mining tasks.

1. **Clustering:** The key objective is to find natural groupings (clusters) in highly dimensional data. Clustering is an example of unsupervised learning, and it is a part of pattern recognition.

2. **Regression Models:** These originate from standard regression analysis and its applied part known as system identification. The underlying idea is to construct a linear or nonlinear function

3. **Classification:** This concerns learning that
classifies data into the predetermined categories. The term originates from pattern recognition, in which a vast number of classifiers have been developed.

4. **Summarization**: This is an approach towards characterizing data via small number of features/attributes. In the simplest scenario one can think of a mean and standard deviations as two extremely compact descriptors of the data. This technique is often applied in an interactive exploratory data analysis and automated report generation.

5. **Link analysis**: It is concerned with determination of relationships (dependencies) between fields in a database. In a particular case we may be interested in the determination of the correlation between the variables.

6. **Sequence Analysis**: This type of analysis is geared toward problems of modeling sequential data. Pertinent models embrace time series analysis, time series models, and temporal neural networks.

Based on our research [3,4,5, 9, 12 ] and the comprehensive analysis of the published papers during the last years in the intelligent e-Learning domain [6,7,8 ], table (1) shows the data mining tasks and the corresponding techniques which fit these tasks.

![Fig.1. Phases of knowledge discovery in databases process](image)

**Table 1: Data Mining Tasks and Techniques**

<table>
<thead>
<tr>
<th>Data Mining Task</th>
<th>Data Mining Algorithm &amp; Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>Classification</td>
<td>Neural Networks</td>
</tr>
<tr>
<td></td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td></td>
<td>Decision Trees</td>
</tr>
<tr>
<td></td>
<td>Genetic Algorithms</td>
</tr>
<tr>
<td></td>
<td>Rule induction</td>
</tr>
<tr>
<td>Clustering</td>
<td>K-Means</td>
</tr>
<tr>
<td>Regression and prediction</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td></td>
<td>Decision Trees</td>
</tr>
<tr>
<td></td>
<td>Rule induction, NN</td>
</tr>
<tr>
<td>Association and Link Analysis (finding correlation between items in a dataset)</td>
<td>Association Rule Mining</td>
</tr>
<tr>
<td>Summarization</td>
<td>Multivariate Visualization</td>
</tr>
</tbody>
</table>

2 Applications of Grouping –Based Techniques in e-Learning

2.1 Clustering Technique

Clustering techniques apply when the instances of data are to be divided into natural groups. The classical clustering technique is k-means where clusters are specified in advance prior to application of the algorithm. This corresponds to parameter k. Then k points are chosen at random as clusters centers. All instances are assigned to their closest cluster center according to the Euclidian distance metric. Next the centroid, or mean, of each cluster center is calculated. These centroids are taken to be the new cluster centers for their respective clusters. The whole process is repeated with the new cluster centers. Iteration continues until the same points are assigned to each cluster in consecutive runs. At this point the cluster centers have stabilized and will remain the same. There are many variants of clustering even for the kmeans algorithm depending upon the method of choosing the initial centers [11, 12].

2.2 Clustering Technique in e-Learning

Clustering is a process of grouping objects into classes of similar objects. It is an unsupervised classification or partitioning of patterns (observations, data items, or feature vectors) into groups or subsets (clusters) based on their locality and connectivity within an n-dimensional space. In e-learning, clustering has been used for:

a) Finding clusters of students with similar learning characteristics and to promote group-based collaborative learning as well as to provide incremental learner diagnosis.

b) Discovering patterns reflecting user behaviours and for collaboration management to characterize similar behaviour groups in unstructured collaboration spaces.

c) Grouping students and personalized itineraries for courses based on learning objects.

d) Grouping students in order to give them differentiated guiding according to their skills and other characteristics.

e) Grouping tests and questions into related groups based on the data in the score matrix.

f) Grouping users based on the time-framed navigation sessions.

2.3 Classification Technique in e-Learning

A classifier is a mapping from a (discrete or continuous) feature space X to a discrete set of
labels Y [13]. Classification or discriminant analysis predicts class labels. This is supervised classification which provides a collection of labeled (preclassified) patterns, the problem being to label a newly encountered, still unlabeled, pattern. In e-learning, classification has been used for:

a) Discovering potential student groups with similar characteristics and reactions to a specific pedagogical strategy.
b) Predicting students’ performance and their final grade.
c) Detecting students’ misuse or students playing around.
d) Predicting the students’ performance as well as to assess the relevance of the attributes involved.
e) Grouping students as hint-driven or failure-driven and finding students’ common misconceptions.
f) Identifying learners with little motivation and finding remedial actions in to lower drop-out rates.
g) Predicting course success

3- Applications of Association Rules-Based Techniques in e-Learning

3.1 Association Rules Mining (ARM) Approach

ARM is one of the most well studied data mining tasks. It discovers relationships among attributes in databases, producing if-then statements concerning attribute-values [14]. 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 respectively the antecedent and consequent of the rule. The strength of such a rule is measured by its support and confidence. The confidence of the rule is the percentage of transactions with X in the database that contain the consequent Y also. The support of the rule is the percentage of transactions in the database that contain both the antecedent and the consequent.

ARM has been applied to e-learning systems for traditionally association analysis (finding correlations between items in a dataset). An efficient algorithm to discover these association rules was first introduced in [14]. The algorithm constructs a candidate set of frequent item sets of length k, counts the number of occurrences, keeps only the frequent ones, then constructs a candidate set of item sets of length k+1 from the frequent item sets of smaller length. It continues iteratively until no candidate item set can be constructed. In other words, every subset of a frequent item set must also be frequent. The rules are then generated from the frequent item sets with probabilities attached to them indicating the likelihood (called support) that the association occurs. We use this idea of association rules to train our recommender agent to build a model representing the web page access behavior or associations between on-line learning activities.

3.2 ARM in Web-Based Education Systems

ARM has been applied to web-based education systems for the following tasks:

a) Building recommender agents that could recommend on-line learning activities or shortcuts.
b) Diagnosing student learning problems and offer students advice.
c) Guiding the learner’s activities automatically and recommending learning materials.
d) Determining which learning materials are the most suitable to be recommended to the user.
e) Identifying attributes characterizing patterns of performance disparity between various groups of students.
f) Discovering interesting relationships from student’s usage information in order to provide feedback to course author.
g) Finding out relationships in learners’ behaviour patterns.
h) Finding students’ mistakes that often accompany each other.
i) Guiding the search for best fitting transfer models of student learning.
j) Optimizing the content of the e-learning portal by determining what most interests the user.

3.3 Sequential Pattern Mining (SPM) in e-Learning

SPM is a more restrictive form of association rule mining in which the accessed items’ order is taken into account. It tries to discover if the presence of a set of items is followed by another item in a time-ordered set of sessions or episodes [15]. The applications of sequential patterns in e-learning can be summarized in the following:

a) Evaluating learners’ activities and can be used in adapting and customizing resource delivery.
b) Discovering and comparison with expected behavioral patterns specified by the instructor that describe an ideal learning path.
c) Giving an indication of how to best organize the educational web space and be able to make
suggestions to learners who share similar characteristics.

d) Generating personalized activities to different groups of learners.

e) Supporting the evaluation and validation of learning site designs.

f) Identifying interaction sequences indicative of problems and patterns that are markers of success.

4. Application of Text Mining (TM) Approach in e-Learning

TM can be viewed as an extension of data mining to text data and it is closely related to web content mining. Its methods include text mining that can work with unstructured or semi-structured data sets such as full-text documents, HTML files and emails [16]. The specific application of text mining techniques in e-learning can be used for the following:

a) Grouping documents according to their topics and similarities and providing summaries.

b) Finding and organizing material using semantic information.

c) Supporting editors when gathering and preparing the materials.

d) Evaluating the progress of the thread discussion to see what the contribution to the topic is.

e) Collaborative learning and a discussion board with evaluation between peers.

f) Identifying the main blocks of multimedia presentations.

g) Selecting articles and automatically constructing e-textbooks and personalized courseware.

h) Detecting the conversation focus of threaded discussions, classifying topics and estimating the technical depth of contribution.

5. Discussion

Table (2) summarizes benefits of the five techniques discussed in the previous sections in the context of e-Learning. From this table, it can be seen that data mining methods are able to discover interesting and useful knowledge based on students’ usage data. Some of the main e-learning problems or subjects to which data mining techniques have been applied are dealing with: (a) the assessment of student’s learning performance, (b) provide course adaptation and learning recommendations based on the students’ learning behaviour, (c) dealing with the evaluation of learning material and educational web-based courses, (c) provide feedback to both teachers and students of e-learning courses, and (d) detection of a typical student’s learning behaviour.

6. Conclusions

This paper discusses the application of data mining techniques in e-learning tasks and domains. The following techniques: clustering, classification, sequential pattern mining, and text mining are discussed from e-learning prospective. Data mining approach provides an effective computational intelligence techniques and robust environment for e-Learning domain. These methods are able to improve the efficiency and performance of the e-learning systems. Data mining techniques can enhance on-line education for the educators as well as the learners. While some tools using data mining techniques to help educators and learners are being developed, the research is still in its infancy. Most of the current data mining tools are too complex for educators to use their features go well beyond the scope of what an educator might require.

References


[5] Sarma Cakula and Abdel-Badeeh M. Salem, "ONTOLOGY-BASED COLLABORATIVE MODEL FOR E-Learning, Proc.of the Annual Int Conf on”Virtual and Augmented Reality in Education” (combined with EEA and Norwegien Finansial Instruments project practical


---

Table 2. Benefits of the Data Mining Techniques: Clustering, Classification, Association Rule Mining (ARM), Sequential Pattern Mining (SPM), and Text Mining (TM)

<table>
<thead>
<tr>
<th>Learning Benefits</th>
<th>Clustering</th>
<th>Classification</th>
<th>ARM</th>
<th>SPM</th>
<th>TM</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Finding clusters of students with similar learning characteristics</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2. Promote group-based collaborative learning</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>3. Discovering patterns reflecting student behaviors</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>4. Grouping students for courses based on learning objects</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>5. Grouping students to give them differentiated guidance according to their skills</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>6. Grouping tests and questions based on the data in the score matrix</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>7. Grouping students based on the time-framed navigation sessions</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>8. Predicting students’ performance and their final grade</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>9. Identifying learners with little motivation</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>10. Predicting course success</td>
<td>+</td>
<td>+</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

11. Recommending on-line learning activities                                         + +

12. Diagnosing student learning problems                                             + +

13. Guiding and evaluating the learner’s activities and recommending learning materials. + +

14. Discovering interesting relationships from student’s usage information           +

15. Finding our relationships in learner’s behavior patterns                          +

16. Finding students’ mistakes that often accompany each other                        +

17. Guiding the search for best fitting transfer models of student learning           +

18. Optimizing the content of the e-learning portal                                   +

19. Giving an indication of how best organize the educational web space               +

20. Generating personalized activities to different groups of learners               +

21. Supporting the evaluation and validation of learning site designs                 +

22. Identifying interaction sequences indicative of problems and patterns             +

23. Grouping documents according to their topics and similarities                     +

24. Finding and organizing material using semantic information                        +

25. Supporting editors when gathering and preparing the materials                     +

26. Evaluating the progress of the thread discussion                                  +

27. Collaborative learning and a discussion board with evaluation between peers       +

28. Identifying the main blocks of multimedia presentations                           +

29. Constructing e-textbooks and personalized coursework                              +

30. Classifying topics and estimating the technical depth of contribution             +