ROLE OF DATA MINING TECHNIQUES IN EDUCATIONAL AND e-LEARNING SYSTEM

Dr. N. Venkatesan
Associate Professor, Dept. of IT, Bharathiyar College of Engg. & Tech, Karaikal
envenki@gmail.com

ABSTRACT

The aim of this research is to provide an up-to-date snapshot of the current state of research and applications of Data Mining methods in education and e-learning process. Educational data mining concerned with developing methods for discovering knowledge from educational domain. Use of data mining algorithms can help discovering pedagogically relevant knowledge contained in databases obtained from Web-based educational systems. Students’ classification based on their learning performance; detection of irregular learning behaviors; e-learning system navigation and interaction optimization; clustering according to similar e-learning system usage and systems’ adaptability to students’ requirements and capacities. This paper shows the types of modeling technique used which includes neural networks, Genetic Algorithms, Clustering and Visualization Methods, Fuzzy Logic, Intelligent agents, Association rule analysis and Inductive Reasoning. From the same point of view, the information is organized according to the type of Data Mining problem dealt with: clustering, classification, prediction, etc.

Keywords: Data mining techniques, e-learning, fuzzy logic, clustering, classification.

1 INTRODUCTION

There are increasing research interests in using data mining in education. These new emerging fields, called educational data mining, concerned with developing methods that extract knowledge from data come from the educational context [109]. The data can be collected from historical and operational data reside in the databases of educational institutes. The student data can be personal or academic. Also it can be collected from e-learning systems which have a large amount of information used by most institutes [108][109].

The main objective of higher education institutes is to provide quality education to its students and to improve the quality of managerial decisions. One way to achieve highest level of quality in higher education system is by discovering knowledge from educational data to study the main attributes that may affect the students’ performance. The discovered knowledge can be used to offer a helpful and constructive recommendations to the academic planners in higher education institutes to enhance their decision making process, to improve students’ academic performance and trim down failure rate, to better understand students’ behavior, to assist instructors, to improve teaching and many other benefits.
Within a decade, the Internet has become a pervasive medium that has changed completely, and perhaps irreversibly, the way information and knowledge are transmitted and shared throughout the world. The education community has not limited itself to the role of passive actor in this unfolding story, but it has been at the forefront of most of the changes. Indeed, the Internet and the advance of telecommunication technologies allow us to share and manipulate information in nearly real time. This reality is determining the next generation of distance education tools. Distance education arose from traditional education in order to cover the necessities of remote students and/or help the teaching-learning process, reinforcing or replacing traditional education.

E-learning (also referred to as web-based education and e-teaching), a new context for education where large amounts of information describing the continuum of the teaching-learning interactions are endlessly generated and ubiquitously available. This could be seen as a blessing: plenty of information readily available just a click away. But it could equally be seen as an exponentially growing nightmare, in which unstructured information chokes the educational system without providing any articulate knowledge to its actors. Data Mining was born to tackle problems like this. As a field of research, it is almost contemporary to e-learning. It is, though, rather difficult to define. Not because of its intrinsic complexity, but because it has most of its roots in the ever-shifting world of business. At its most detailed, it can be understood not just as a collection of data analysis methods, but as a data analysis process that encompasses anything from data understanding, pre-processing and modeling to process evaluation and implementation. It is nevertheless usual to pay preferential attention to the Data Mining methods themselves. These commonly bridge the fields of traditional statistics, pattern recognition and machine learning to provide analytical solutions to problems in areas as diverse as biomedicine, engineering, and business, to name just a few. An aspect that perhaps makes Data Mining unique is that it pays special attention to the compatibility of the modeling techniques with new Information Technologies (IT) and database technologies, usually focusing on large, heterogeneous and complex databases. E-learning databases often fit this description. Therefore, Data Mining can be used to extract knowledge from e-learning systems through the analysis of the information available in the form of data generated by their users. In this case, the main objective becomes finding the patterns of system age by teachers and students and, perhaps most importantly, discovering the students’ learning behavior patterns. This work aims to provide an as complete as possible review of the many applications of Data Mining to e-learning; that is, a survey of the literature in this area up to date. We must acknowledge that this is not the first time a similar venture has been undertaken: a collection of papers that cover most of the important topics in the field was concurrently presented in [71]. The findings of the survey are organized from different points of view that might in turn match the different interests of its potential readers: The surveyed research can be seen as being displayed along two axes: Data Mining problems and methods, and e-learning applications.

Section 2 presents the literature review introduction of the research. Section 3 presents the survey of data mining techniques in education field for the decision making analysis. Section 4 deals with Data Mining and e-learning process. Classification problems of e-learning are
described in the section 5 with the use of data mining techniques. Clustering based problems are discussed in the section 6. Section 7 describes the more data mining related problems. Section 8 is concluded along with future work.

2. A SURVEY ON DATA MINING IN EDUCATION AND e-LEARNING

As stated in the introduction, it is aim to organize the findings of the survey in different ways that might correspond to the diverse readers’ academic or professional backgrounds. This section presents the surveyed research according to the Data Mining problems (classification, clustering, etc.), techniques and methods (e.g., Neural Networks, Genetic Algorithms, Decision Trees, or Fuzzy Logic).

In fact, most of the existing research addresses problems of classification and clustering. For this reason, specific subsections will be devoted to them. But first, let us try to find a place for Data Mining in the world of e-learning. The following sections are described the role of data mining in education field and e-learning process.

3. DATA MINING IN EDUCATION

Data mining is used in higher education is a recent research field; there are many works in this area. That is because of its potentials to educational institutes.

Romero and Ventura [108], have a survey on educational data mining between 1995 and 2005. They concluded that educational data mining is a promising area of research and it has a specific requirements not presented in other domains. Thus, work should be oriented towards educational domain of data mining.

El-Halees [104], gave a case study that used educational data mining to analyze students’ learning behavior. The goal of this study is to show how data mining can be used in higher education to improve students’ performance. Students’ data from database course and collected all available data including personal records and academic records of students, course records and data came from e-learning system. Data mining techniques are applied to discover many kinds of knowledge such as association rules and classification rules using decision tree.

Baradwaj and Pal [102], applied the classification as data mining technique to evaluate students’ performance, they used decision tree method for classification. The goal of their study is to extract knowledge that describes students’ performance in end semester examination. They used students’ data from the students’ previous database including Attendance, Class test, Seminar and Assignment marks. This study helps earlier in identifying the dropouts and students who need special attention and allow the teacher to provide appropriate advising.
Shannaq et al. [110], applied the classification as data mining technique to predict the numbers of enrolled students by evaluating academic data from enrolled students to study the main attributes that may affect the students’ loyalty. The extracted classification rules are based on the decision tree as a classification method, the extracted classification rules are studied and evaluated using different evaluation methods. It allows the University management to prepare necessary resources for the new enrolled students and indicates at an early stage which type of students will potentially be enrolled and what areas to concentrate upon in higher education systems for support.

Al-Radaideh et al. [100], applied the data mining techniques, particularly classification to help in improving the quality of the higher educational system by evaluating student data to study the main attributes that may affect the student performance in courses. The extracted classification rules are based on the decision tree as a classification method, the extracted classification rules are studied and evaluated. It allows students to predict the final grade in a course under study.

Chandra and Nandhini [103], applied the association rule mining analysis based on students’ failed courses to identify students’ failure patterns. The goal of their study is to identify hidden relationship between the failed courses and suggests relevant causes of the failure to improve the low capacity students’ performances. The extracted association rules reveal some hidden patterns of students’ failed courses which could serve as a foundation stone for academic planners in making academic decisions and an aid in the curriculum re-structuring and modification with a view to improving students’ performance and reducing failure rate.

Ayesha et al. [101], used k-means clustering algorithm as a data mining technique to predict students’ learning activities in a students’ database including class quizzes, mid and final exam and assignments. These correlated information will be conveyed to the class teacher before the conduction of final exam. This study helps the teachers to reduce the failing ratio by taking appropriate steps at right time and improve the performance of students.

Data mining encompasses different algorithms that are diverse in their methods and aims. It also comprises data exploration and visualization to present results in a convenient way to users. A data element will be called an individual. It is characterized by a set of variables. In this context, most of the time an individual is a learner and variables can be exercises attempted by the learner, marks obtained, scores, mistakes made, time spent, number of successfully completed exercises and so on. New variables may be calculated and used in algorithms, such as the average number of mistakes made per attempted exercise.

**Data exploration and visualization**

Raw data and algorithm results can be visualized through tables and graphics such as graphs and histograms as well as through more specific techniques such as symbolic data analysis. The aim
is to display data along certain attributes and make extreme points, trends and clusters obvious to human eye.

**Clustering** algorithms aim at finding homogeneous groups in data. We used k-means clustering and its combination with hierarchic clustering [10]. Both methods rest on a distance concept between individuals.

**Classification** is used to predict values for some variable. For example, given all the work done by a student, one may want to predict whether the student will perform well in the final exam. To use of C4.5 for decision tree this relies on the concept of entropy. The tree can be represented by a set of rules. Thus The tree is built taking a representative population and is used to predict values for new individuals.

**Association rules** find relations between items. Rules have the following form: $X \rightarrow Y$, support 40%, confidence 66%, which could mean 'if students get $X$ incorrectly, then they get also $Y$ incorrectly', with a support of 40% and a confidence of 66%. Support is the frequency in the population of individuals that contains both $X$ and $Y$. Confidence is the percentage of the instances that contains $Y$ amongst those which contain $X$ and implemented a variant of the standard Apriori algorithm [14] in TADA-Ed that takes temporality into account. Taking temporality into account produces a rule $X \rightarrow Y$ only if exercise $X$ occurred before $Y$.

### 4. DATA MINING AND IN E-LEARNING PROCESS

Some researchers have pointed out the close relation between the fields of Artificial Intelligence (AI) and Machine Learning (ML) are main sources of Data Mining techniques, methods and education processes [4, 26, 30, 49]. In [4], the author establishes the research opportunities in AI and education on the basis of three models of educational processes: models as scientific tool, are used as a means for understanding and forecasting some aspect of an educational situation; models as component: corresponding to some characteristic of the teaching or learning process and used as a component of an educative art effect; and models as basis for design of educational are facts: assisting the design of computer tools for education by providing design methodologies and system components, or by constraining the range of tools that might be available to learners.

In [49, 85], studies on how Data Mining techniques could successfully be incorporated to e-learning environments and how they could improve the learning tasks were carried out. In [85], data clustering was suggested as a means to promote group-based collaborative learning and to provide incremental student diagnosis.

A review of the possibilities of the application of Web Mining techniques to meet some of the current challenges in distance education was presented in [30]. The proposed approach could improve the effectiveness and efficiency of distance education in two ways: on the one hand, the
discovery of aggregate and individual paths for students could help in the development of effective customized education, providing an indication of how to best organize the educator organization’s courseware. On the other hand, virtual knowledge structure could be identified through Web Mining methods: The discovery of Association Rules could make it possible for Web-based distance tutors to identify knowledge patterns and reorganize the virtual course based on the patterns discovered.

An analysis on how ML techniques again, a common source for Data Mining techniques have been used to automate the construction and induction of student models, as well as the background knowledge necessary for student modeling, were presented in [79]. In this paper, the difficulty, appropriateness and potential of applying ML techniques to student modeling was commented.

5. THE CLASSIFICATION PROBLEM IN E-LEARNING

In classification problem aim is to model the existing relationships between a set of multivariate data items and a certain set of outcomes for each of them in the form of class membership labels.

5.1 FUZZY LOGIC METHODS

Fuzzy logic-based methods have only recently taken their first steps in the e-learning field [36, 39]. In [81], a neuro-fuzzy model for the evaluation of students in an intelligent tutoring system (ITS) was presented. Fuzzy theory was used to measure and transform the interaction between the student and the ITS into linguistic terms. Then, Artificial Neural Networks were trained to realize fuzzy relations operated with the max-min composition. These fuzzy relations represent the estimation made by human tutors of the degree of association between an observed response and a student characteristic. A further work by Hwang and colleagues [36, 40], a fuzzy rules-based method for eliciting and integrating system management knowledge was proposed and served as the basis for the design of an intelligent management system for monitoring educational Web servers. This system is capable of predicting and handling possible failures of educational Web servers, improving their stability and reliability. It assists students’ self-assessment and provides them with suggestions based on fuzzy reasoning techniques. A two-phase fuzzy mining and learning algorithm was described in [89].

5.2 ARTIFICIAL NEURAL NETWORKS AND EVOLUTIONARY COMPUTATION

Some research on the use of Artificial Neural Networks and Evolutionary Computation models to deal with e-learning topics can be found in [53, 55, 87]. A navigation support system based on an Artificial Neural Network was put forward in [55] to decide on the appropriate navigation strategies. The Neural Network was used as a navigation strategy decision module in the system. Evaluation has validated the knowledge learned by the Neural Network and the level of effectiveness of the navigation strategy.
In [53, 87], evolutionary algorithms were used to evaluate the students’ learning behaviour. A combination of multiple classifiers (CMC), for the classification of students and the prediction of their final grades, based on features extracted from logged data in an education web-based system, was described in [53]. The classification and prediction accuracies are improved through the weighting of the data feature vectors using a Genetic Algorithm. In [87], a random code generation and mutation process suggested as a method to examine the comprehension ability of students.

5.3 GRAPHS AND TREES

Graph and/or tree theory was applied to e-learning in [9, 13, 14, 29, 42, 47, 48]. An e-learning model for the personalization of courses, based both on the student’s needs and capabilities and on the teacher’s profile, was described in [9]. Personalized learning paths in the courses were modeled using graph theory. In [47, 48], Decision Trees (DT) as classification models were applied. A discussion of the implementation of the Distance Learning Algorithm (DLA), which uses Rough Set theory to find general decision rules, was presented by [47]: A DT was used to adequate the original algorithm to distance learning issues. On the basis of the obtained results, the instructor might consider the reorganization of the course materials. System architecture for mining learners’ online behaviour patterns was put forward in [13]. A framework for the integration of traditional Web log mining algorithms with pedagogical meanings of Web pages was presented.

Also in [48], an automatic tool, based on the students’ learning performance and communication preferences, for the generation and discovery of simple student models was described, with the ultimate goal of creating a personalized education environment. The approach was based on the PART algorithm, which produces rules from pruned partial DTs. In [97], a tool that can help trace deficiencies in students’ understanding was presented. It resorts to a tree Abstract Data Type (ADT), built from the concepts covered in a lab, lecture, or course. Once the tree ADT is created, each node can be associated with different entities such as student performance, class performance, or lab development. Using this tool, a teacher could help students by discovering concepts that needed additional coverage, while students might discover concepts for which they would need to spend additional working time. A tool to perform a quantitative analysis based on students’ learning performance was introduced in [14]. It proposes new courseware diagrams, combining tools provided by the theory of conceptual maps [63] and influence diagrams [75]. In [29, 42], personalized Web-based learning systems were defined, applying Web usage mining techniques to personalized recommendation services. The approach is based on a Web page classification method, which uses attribute-oriented induction according to related domain knowledge shown by a concept hierarchy tree.

142
5.4 ASSOCIATION RULES

Association Rules for classification, applied to e-learning, have been investigated in the areas of learning recommendation systems [18, 98, 99], learning material organization [89], student learning assessments [38, 45, 52, 54, 69, 70], course adaptation to the students’ behaviour [19, 35, 50], and evaluation of educational web sites [21]. Data Mining techniques such as Association Rule mining, and inter-session and intra-session frequent pattern mining, were applied in [98, 99] to extract useful patterns that might help educators, educational managers, and Web masters to evaluate and interpret on-line course activities. A similar approach can be found in [54], where contrast rules, defined as sets of conjunctive rules describing patterns of performance disparity between groups of students, were used. A computer-assisted approach to diagnosing student learning problems in science courses and offer students advice was presented in [38], based on the concept effect relationship (CER) model (a specification of the Association Rules technique).

A hypermedia learning environment with a tutorial component was described in [19]. It is called Logiocando and targets children of the fourth level of primary school (9-10 years old). It includes a tutor module, based on if-then rules, that emulates the teacher by providing suggestions on how and what to study. In [52], the description of a learning process assessment method that resorts to Association Rules, and the well-known ID3 DT learning method. A framework for the use of Web usage mining to support the validation of learning site designs was defined in [21], applying association and sequence techniques [80].

In [50], a framework for personalized e-learning based on aggregate usage profiles and a domain ontology were presented, and a combination of Semantic Web and Web mining methods was used. The Apriori algorithm for Association Rules was applied to capture relationships among URL references based on the navigational patterns of students. A test result feedback (TRF) model that analyzes the relationships between student learning time and the corresponding test results was introduced in [35]. The objective was twofold: on the one hand, developing a tool for supporting the tutor in reorganizing the course material; on the other, a personalization of the course tailored to the individual student needs. The approach was based in Association Rules mining. A rule-based mechanism for the adaptive generation of problems in ITS in the context of web-based programming tutors was proposed in [45]. In [18], a web-based course recommendation system, used to provide students with suggestions when having trouble in choosing courses, was described.

5.5 MULTI-AGENT SYSTEMS

Multi Agents Systems (MAS) for classification in e-learning have been proposed in [2, 28]. In [28] this takes the form of an adaptive interaction system based on three MAS: the Interaction MAS captures the user preferences applying some defined usability metrics (affect, efficiency, helpfulness, control and learnability). The Learning MAS shows the contents to the user.
according to the information collected by the Interaction MAS in the previous step; and the Teaching MAS offers recommendations to improve the virtual course. A multi-agent recommendation system, called InLix, was described in [2]; it suggests educational resources to students in a mobile learning platform.

6. THE CLUSTERING PROBLEM IN E-LEARNING

Unlike in classification problems, in data grouping or clustering are not interested in modeling a relation between a set of multivariate data items and a certain set of outcomes for each of them (being this in the form of class membership labels). Instead, aim to discover and model the groups in which the data items are often clustered, according to some item similarity measure.

To find a first application of clustering methods in [37], where a network-based testing and diagnostic system was implemented. It entails a multiple-criteria test sheet-generating problem and a dynamic programming approach to generate test sheets. The proposed approach employs fuzzy logic theory to determine the difficulty levels of test items according to the learning status and personal features of each student, and then applies an Artificial Neural Network model: Fuzzy Adaptive Resonance Theory (Fuzzy ART) [10] to cluster the test items into groups, as well as dynamic programming [22] for test sheet construction.

In [60, 61], an in-depth study describing the usability of Artificial Neural Networks and, more specifically, of Kohonen’s Self-Organizing Maps (SOM) [43] for the evaluation of students in a tutorial supervisor (TS) system, as well as the ability of a fuzzy TS to adapt question difficulty in the evaluation process, was carried out. An investigation on how Data Mining techniques could be successfully incorporated to e-learning environments, and how this could improve the learning processes was presented in [85]. Here, data clustering is suggested as a means to promote group based collaborative learning and to provide incremental student diagnosis.

In [86], user actions associated to students’ Web usage were gathered and preprocessed as part of a Data Mining process. The Expectation-Maximization (EM) algorithm was then used to group the users into clusters according to their behaviours. These results could be used by teachers to provide specialized advice to students belonging to each cluster. The simplifying assumption that students belonging to each cluster should share web usage behaviour makes personalization strategies more scalable. The system administrators could also benefit from this acquired knowledge by adjusting the e-learning environment they manage according to it. The EM algorithm was also the method of choice in [82], where clustering was used to discover user behaviour patterns in collaborative activities in e-learning applications. Some researchers [23,31,83] propose the use of clustering techniques to group similar course materials: An ontology-based tool, within a Web Semantics framework, was implemented in [83] with the goal of helping e-learning users to find and organize distributed courseware resources. An element of this tool was the implementation of the Bisection K-Means algorithm, used for the grouping of similar learning materials. Kohonen’s well-known SOM algorithm was used in [23]
to devise an intelligent searching tool to cluster similar learning material into classes, based on its semantic similarities. Clustering was proposed in [31] to group similar learning documents based on their topics and similarities. A Document Index Graph (DIG) for document representation was introduced, and some classical clustering algorithms (Hierarchical Agglomerative Clustering, Single Pass Clustering and k-NN) were implemented.

Different variants of the Generative Topographic Mapping (GTM) model, a probabilistic alternative to SOM, were used in [11, 12, 94] for the clustering and visualization of multivariate data concerning the behaviour of the students of a virtual course. More specifically, in [11, 94] a variant of GTM known to behave robustly in the presence of atypical data or outliers was used to successfully identify clusters of students with atypical learning behaviours. A different variant of GTM for feature relevance determination was used in [12] to rank the available data features according to their relevance for the definition of student clusters.

In [60, 61], an in-depth study describing the usability of Artificial Neural Networks and, more specifically, of Kohonen’s Self-Organizing Maps (SOM) [43] for the evaluation of students in a tutorial supervisor (TS) system, as well as the ability of a fuzzy TS to adapt question difficulty in the evaluation process, was carried out. An investigation on how Data Mining techniques could be successfully incorporated to e-learning environments, and how this could improve the learning processes was presented in [85]. Here, data clustering is suggested as a means to promote group based collaborative learning and to provide incremental student diagnosis.

7. DATA MINING PROBLEMS IN E-LEARNING

As previously stated most of the current research deals with problems of classification and clustering in e-learning environments. However, there are several applications that tackle other Data Mining problems such as prediction and visualization, which are discussed in the following subsections.

7.1 PREDICTION TECHNIQUES

Prediction is often also an interesting problem in e-learning, although it must be born in mind that it can easily overlap with classification and regression problems. The forecasting of students’ behaviour and performance when using e-learning systems bears the potential of facilitating the improvement of virtual courses as well as e-learning environments in general.

A methodology to improve the performance of developed courses through adaptation was presented in [72, 73]. Course log-files stored in databases could be mined by teachers using evolutionary algorithms to discover important relationships and patterns, with the target of discovering relationships between students’ knowledge levels, e-learning system usage times and students’ scores.
A system for the automatic analysis of user actions in Web-based learning environments, which could be used to make predictions on future uses of the learning environment, was presented in [59]. It applies a C4.5 DT model for the analysis of the data.

Some studies apply regression methods for prediction [5, 27, 44]. In [27], a study that aimed to find the sources of error in the prediction of students’ knowledge behaviour was carried out. Stepwise regression was applied to assess what metrics help to explain poor prediction of state exam scores. Linear regression was applied in [5] to predict whether the student’s next response would be correct, and how long he or she would take to generate that response.

In [44], a set of experiments was conducted in order to predict the students’ performance in e-learning courses, as well as to assess the relevance of the attributes involved. In this approach, several Data Mining methods were applied which includes Naïve Bayes, kNN, MLP Neural Network, C4.5, Logistic Regression and Support Vector Machines. With similar goals in mind, experiments applying the Fuzzy Inductive Reasoning (FIR) methodology to the prediction of the students’ final marks in a course taken at a virtual campus were carried out in [62]. The relative relevance of specific features describing course online behaviour was also assessed. This work was extended in [25] using Artificial Neural Networks for the prediction of the students’ final marks. In this work, the predictions made by the network were interpreted using Orthogonal Search-based Rule Extraction (OSRE) a novel rule extraction algorithm [24]. Rule extraction was also used in [72, 73] with the emphasis on the discovery of interesting prediction rules in student usage information, in order to use them to improve adaptive Web courses. Graphical models and Bayesian methods have also been used in this context. For instance, an open learning platform for the development of intelligent Web-based educative systems, named MEDEA, was presented in [88]. Systems developed with MEDEA guide students in their learning process, and allow free navigation to better suit their learning needs. A Bayesian Network model lies at the core of MEDEA. In [3] an evaluation of students’ attitudes and their relationship to students’ performance in a tutoring system was implemented. Starting from a correlation analysis between variables, a Bayesian Network that inferred negative and positive students’ attitudes was built. Finally, a Dynamic Bayes Net (DBN) was used in [15], for modeling students’ knowledge behaviour and predict future performance in an ITS.

In [90, 91], a tool for the automatic detection of a typical behaviours on the students’ use of the e-learning system was defined. It resorts to a Bayesian predictive distribution model to detect irregular learning processes on the basis of the students’ response time. Note that some models for the detection of atypical student behavior were also referenced in the section reviewing clustering applications [11, 94].

### 7.2 VISUALIZATION TECHNIQUES

One of the most important phases of a Data Mining process is that of data exploration through visualization methods. Visualization was understood in [68] in the context of Social Network
Analysis adapted to collaborative distance-learning, where the cohesion of small learning groups is measured. The cohesion is computed in several ways in order to highlight isolated people, active sub-groups and various roles of the members in the group communication structure. Note the links between this goal and that of a typical student behaviour described in previous sections. The method allows the display of global properties both at individual level and at group level, as well as to efficiently assist the virtual tutor in following the collaboration patterns within the group.

An educational Data Mining tool is presented in [57, 58] that shows, in a hierarchical and partially ordered fashion, the students’ interaction with the e-learning environment and their virtual tutors. The tool provides case analysis and visualizes the results in an event tree, exploiting MySQL databases to obtain tutorial events. One main limitation to the analysis of high-dimensional multivariate data is the difficulty of representing those data faithfully in an intuitive visual way. Latent methods (of which Principal Component Analysis, or PCA, is perhaps the most widely known) allow such representation. One such latent method was used in [11, 12, 94] to display high-dimensional student behaviour data in a 2-dimensional representation. This type of visualization helps detecting the characteristics of the data distributions and their grouping or cluster structure.

### 7.3 OTHER DATA MINING TECHNIQUES

Not all Data Mining in e-learning concerns advanced AI or ML methods: traditional statistics are also used in [1, 32, 74, 77], as well as Semantic Web technologies [34], ontologies [46], Case-Based Reasoning [33] and/or theoretical modern didactical approaches [6, 7, 41, 96].

Although it could have been included in the section devoted to classification, Naïve Bayes, the model used in [78, 84], also fits in the description of general statistical method. An approach to automate the classification process of Web learning resources was developed in [78]. The model organizes and labels learning resources according to a concept hierarchy extracted from the extended ontology of the ACM Computing Curricula 2001 for Computer Science. In [84], a method to construct personalized courseware was proposed. It consists of the building of a personalized Web tutor tree using the Naïve algorithm, for mining both the context and the structure of the courseware.

Statistical methods were applied in [8, 56, 64]. In [64], the goals were the discovery and extraction of knowledge from an e-learning database to support the analysis of student learning processes, as well as the evaluation of the effectiveness and usability of Web-based courses. Three Web Mining-based evaluation criteria were considered: session statistics, session patterns and time series of session data.

An experiment combining a MAS and self-regulation strategies to allow flexible and incremental design, and to provide a more realistic social context for interactions between students and the
teachable agent, were presented in [6]. In [41], a model called Learning Response Dynamics that analyzes learning systems through the concepts of learning dynamics, energy, speed, force, and acceleration, was described. In [7, 96], the problems of developing versatile adaptive and intelligent learning systems that could be used in the context of practical Web-based education were discussed.

MAS has also been applied to e-learning beyond classification problems. In [76], one called IDEAL was designed to support student-centered, self-paced, and highly interactive learning. The analysis was carried out on the students’ learning-related profile, which includes learning style and background knowledge in selecting, organizing, and presenting the learning material to support active learning. IDEAL supports personalized interaction between the students and the learning system and enables adaptive course delivery of educational contents. The student learning behaviour (student model) is inferred from the performance data using a Bayesian Belief Network model. In [66, 67], a MAS called Cooperative Intelligent Distance Learning Environments (CIDLE) was described. It extracts knowledge from domain knowledge and students’ behaviour during a learning discussion. It therefore infers the learners’ behaviour and adapts to them the presentation of course material in order to improve their success rate in answering questions. In [51], software agents were proposed as an alternative for data extraction from e-learning environments, in order to organize them in intelligent ways.

8. CONCLUSION

This paper presented a general and up-to-date survey on Data Mining application in e-learning, as reported in the field of education. This survey will be useful for Data Mining practitioners and e-learning system managers and developers, but also even for members and users, teachers and learners, of the e-learning community at large in the research field of education and e-learning. Data mining techniques and its usages played vital role in Education based prediction analysis and in e-learning system. This will motivate to develop new algorithms and useful applications to predict the hidden and useful information as future scope.

REFERENCES


