International Journal of Modern Agriculture, Volume 9, No. 4, 2020
ISSN: 2305-7246

APPLYING DATA MINING TECHNIQUE IN EDUCATION MANAGEMENT SYSTEMS

R.Karthikeyan

Asst professor Dept of Computer science and Engineering BIHER, Chennai

rkarthikeyan78@yahoo.com

Abstract

It is well known that in Information Technology (IT) driven society, knowledge is one of the most significant assets of any organization. The role of IT in education is well established. Association rules are very useful in Educational Data mining since they extract associations between educational items and present the results in an intuitive form to the teachers. In this paper, we survey the application of association rule mining in e-learning systems, and especially, learning management systems and we have found some drawbacks and some possible solutions to resolve them.

Key words: popular techniques in the data mining, association rule mining process in EMS, drawbacks and solutions.

Introduction

Knowledge discovery in databases is well-defined process consisting of several distinct steps.[3] Data mining is the core step, which results in the discovery of hidden but useful knowledge from massive databases. A formal definition of Knowledge discovery in databases is given as follows:

"Data mining is the non trivial extraction of implicit previously unknown and potentially useful information about data"[1]. Data mining technology provides a user- oriented approach to novel and hidden patterns in the data. Nowadays, Education Management Systems (EMS) are being installed more and more by universities, community colleges, schools, businesses, and even individual instructors in order to add web technology to their courses and to supplement traditional face-to-face courses [1]. EMS systems accumulate a vast amount of information which is very valuable for analyzing the students' behavior and could create a gold mine of educational data [2].

They can record whatever student activities it involves, such as reading, writing, taking tests, performing various tasks, and even communicating with peers. However, due to the vast quantities of data these systems can generate daily, it is very difficult to analyze this data manually. A very promising approach towards this analysis objective is the use of data mining techniques.

Association rules mining is one of the most well studied data mining tasks. It discovers relationships among attributes in databases, producing if-then statements concerning attribute-values [4]. An association rule X 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.

Association rule mining has been applied to e-learning systems for traditionally association analysis (finding correlations between items in a dataset), including, e.g., the following tasks: building recommender agents for on-line learning activities or shortcuts [5], automatically guiding the learner's activities and intelligently generate and recommend learning materials [6], identifying attributes characterizing patterns of performance

International Journal of Modern Agriculture, Volume 9, No. 4, 2020
ISSN: 2305-7246

disparity between various groups of students [7], discovering interesting relationships from student's usage information in order to provide feedback to course author [8], finding out the relationships between each pattern of learner's behavior [9], finding students' mistakes that are often occurring together [10], guiding the search for best fitting transfer model of student learning [11], optimizing the content of an e-learning portal by determining the content of most interest to the user [12], extracting useful patterns to help educators and web masters evaluating and interpreting on-line course activities [5], and personalizing e-learning based on aggregate usage profiles and a domain ontology [13].

The extraction of sequential patterns has been used in e-learning for evaluating the learners' activities and can be used in adapting and customizing resource delivery [14]; discovering and comparison with expected behavioral patterns specified by the instructor that describes an ideal learning path [15]; giving an indication of how to best organize the educational web space and be able to make suggestions to learners who share similar characteristics [16]; generating personalized activities to different groups of learners [17]; supporting the evaluation and validation of learning site designs [18]; identifying interaction sequences indicative of problems and patterns that are markers of success [19]. Finally, association rule mining has been used in the e-learning for classification [20]. From a syntactic point of view, the main difference to general association rules is that classification rules have a single condition in the consequent which is the class identifier name. They have been applied in learning material organization [21], student learning assessments [22, 23, 24], course adaptation to the students' behavior [25, 26] and evaluation of educational web sites [27].

Popular Techniques in the Data Mining

Several techniques have been proposed for solving a problem of extracting knowledge from explosive data, each of which adopts different algorithm. One of the areas where information plays an important role is that of education. The main ideas, used in data mining, can be categorized as follows [4]:

1. Association Mining

A main idea of association mining technique is to search a relationship of attributes and tuples, by discovering frequently occurring item sets in database. A result is patterns described as rules that represent one-way relationship. Furthermore, result rules consist of a confidential value and support value, a value of which is used to identify the pattern.

2. Classification and Prediction

The classification and prediction approaches are based on a divide-and-conquer technique by which the data are grouped and missing values are predicted.

3. Clustering

Clustering is a method to group data into classes with identical characteristics in which the similarity of intra-class is maximized or minimized.

The Association Rule Mining Process in EMS

 Collecting data. Most of the current EMSs do not store logs as text files. Instead, they normally use a relational database that stores all the systems information: personal information of the users (profile), academic results, the user's interaction data, etc.. Databases are more powerful, flexible and bug -prone than the typically textual log files for gathering detailed access and high level usage information from all the services available in the EMS. The EMSs keep detailed logs of all activities that students perform. Not only every click that students make for navigational purposes (low level information) is stored, but also test scores, elapsed time, etc. (high level information).

 Data pre-processing. Most of the traditional data pre-processing tasks, such as data cleaning, user

International Journal of Modern Agriculture, Volume 9, No. 4, 2020
ISSN: 2305-7246

identification, session identification, transaction identification, data transformation and enrichment, data integration and data reduction are not necessary in EMS. Data pre-processing of EMS data is simpler due to the fact that most EMS store the data for analysis purposes, in contrast to the typically observational datasets in data mining, that were generated to support the operational setting and not for analysis in the first place. EMSs also employ a database and user authentication (password protection) which allows identifying the users in the logs.

- **Applying the mining algorithms.** In this step it is necessary: 1) to choose he specific association rule mining algorithm and implementation; 2) to configure the parameters of the algorithm, such as support and confidence threshold and others; 3) to identify which table or data file will be used for the mining; 4) and to specify some other restrictions, such as the maximum number of items and what specific attributes can be present in the antecedent or consequent of the discovered rules.

- **Data post-processing.** The obtained results or rules are interpreted, evaluated and used by the teacher for further actions. for making decisions about the students and the EMS activities of the course in order to improve the students' learning. So, data mining algorithms have to express the output in a comprehensible format by e.g., using standardized e-learning metadata.

It is important to notice that traditional educational data sets are normally small [28] if we compare them to databases used in other data mining fields such as e-commerce applications that involve thousands of clients. This is due to the fact that the typical size of one classroom is often only between 10-100 students, depending on the type of the course (elementary, primary, adult, higher, tertiary, academic and special education). In the distance learning setting, the class size is usually larger, and it is also possible to pool data from several years or from several similar courses. Furthermore, the total number of instances or transactions can be quite large depending on how much information the EMS stores about the interaction of each student with the system (and at what levels of granularity). In this way, the number of available instances is much higher than the number of students. And, as we have said previously, educational data has also one advantage compared to several other domains [28]: the data sets are usually very clean.

Drawbacks and Solutions

Over the past decade a variety of algorithms that address these issues through the refinement of search strategies, pruning techniques and data structures have been developed. While most algorithms focus on the explicit discovery of all rules that satisfy minimal support and confidence constraints for a given dataset, increasing consideration is being given to specialized algorithms that attempt to improve processing time or facilitate user interpretation by reducing the result set size and by incorporating domain knowledge [30].

Most of the current data mining tools are too complex for educators to use and their features go well beyond the scope of what an educator might require. As a result, the courses administrator is more likely to apply data mining techniques in order to produce reports for instructors who then use these reports to make decisions about how to improve the student's learning and the online courses. However, it is most desirable that teachers participate directly in the iterative mining process in order to obtain more valuable rules. But normally, teachers only use the feedback provided by the obtained rules to make decisions about modification to improve the course, detect activities or students with problems, etc. and the main drawback here is the used algorithms have too many parameters , non-interesting and with low comprehensibility. In the following subsections, we will tackle these problems.

4.1 Finding the appropriate parameter settings of the mining algorithm

Association rule mining algorithms need to be configured before to be executed. So, the user has to give appropriate values for the parameters in advance (often leading to too many or too few rules) in order to obtain a good number of rules. A comparative study between the main algorithms that are currently used to discover association rules can be found in [31]: Apriori [32], FP-Growth [33], MagnumOpus [34], Closet [35]. Most of these algorithms require the user to set two thresholds, the minimal support and the minimal confidence, and

International Journal of Modern Agriculture, Volume 9, No. 4, 2020
ISSN: 2305-7246

find all the rules that exceed the thresholds specified by the user. Therefore, the user must possess a certain amount of expertise in order to find the right settings for support and confidence to obtain the best rules.

One possible solution to this problem can be to use a parameter-free algorithm or with less parameters. For example, the Weka [36] package implements an Apriori-type algorithm that solves this problem partially. This algorithm reduces iteratively the minimum support, by a factor delta support (Δs) introduced by the user, until a minimum support is reached or a required number of rules (NR) has been generated.

Another improved version of the Apriori algorithm is the Predictive Apriori algorithm, which automatically resolves the problem of balance between these two parameters, maximizing the probability of making an accurate prediction for the data set. In order to achieve this, a parameter called the exact expected predictive accuracy is defined and calculated using the Bayesian method, which provides information about the accuracy of the rule found. In this way the user only has to specify the maximal number or rules to discover.

In experimental tests were performed on a Moodle course by comparing the two previous algorithms. The final results demonstrated better performance for Predictive Apriori than Apriori-type algorithm using the Δs factor.

4.2 Discovering too many rules

The application of traditional association algorithms will be simple and efficient. However, association rule mining algorithms normally discover a huge quantity of rules and do not guarantee that all the rules found are relevant. Support and confidence factors can be used for obtaining interesting rules which have values for these factors grater than a threshold value. Although these two parameters allow the pruning of many associations, another common constraint is to indicate the attributes that must or cannot be present in the antecedent or consequent of the discovered rules.

Another solution is to evaluate, and post-prune the obtained rules in order to find the most interesting rules for a specific problem. Traditionally, the use of objective interestingness measures has been suggested , such as support and confidence, mentioned previously, as well as others measures such as Laplace, chi-square statistic, correlation coefficient, entropy gain, gini, interest, conviction, etc. These measures can be used for ranking the obtained rules in order than the user can select the rules with highest values in the measures that he/she is more interested.

Subjective measures are becoming increasingly important, in other words measures that are based on subjective factors controlled by the user. Most of the subjective approaches involve user participation in order to express, in accordance with his or her previous knowledge, which rules are of interest. Some suggested subjective measures are:

- Unexpectedness: Rules are interesting if they are unknown to the user or contradict the user's knowledge.
- Action ability: Rules are interesting if users can do something with them to their advantage.

The number of rules can be decreased by only showing unexpected and actionable rules to the teacher and not all the discovered rules. In an Interestingness Analysis System (IAS) is proposed. It compares rules discovered with the user's knowledge about the area of interest. Let *U* be the set of user's specifications representing his/her knowledge space, *A* be the set of discovered association rules, this algorithm implements a pruning technique for removing redundant or insignificant rules by ranking and classifying them into four categories:

- Conforming rules: a discovered rule A_i A conforms to a piece of user's knowledge U_j if both the antecedent and the consequent parts of A_i match those of U_j U well.
- Unexpected consequent rules: a discovered rule A_i A has unexpected consequents with respect to U_j U if the antecedent part of A_i matches that of U well.

International Journal of Modern Agriculture, Volume 9, No. 4, 2020
ISSN: 2305-7246

- Unexpected condition rules: a discovered rule A_i A has unexpected conditions with respect to U_j U if the consequent part of A_i matches that of U_j well, but not the antecedent part.

- Both-side unexpected rules: a discovered rule A_i A is both-side unexpected with respect to U_j U if the antecedent and consequent parts of A_i don't match those of U_j well.

The degrees of membership into each of these four categories are used for ranking the rules. Using their own specification language, they indicate their knowledge about the matter in question, through relationships among the fields or items in the database.

Finally, we can also use the knowledge database as a rule repository on the basis of which subjective analysis of the rules discovered is performed. Before running the association rule mining algorithm, the teacher could download the relevant knowledge database, in accordance with his/her profile. The personalization of the rules returned is based on filtering parameters, associated with the type of the course to be analyzed such as: the area of knowledge; the level of education; the difficulty of the course, etc. The rules repository is created on the server in a collaborative way where the experts can vote for each rule in the repository, based on the educational considerations and their experience gained in other similar e-learning courses.

4.3 Discovery of poorly understandable rules

A factor that is of major importance in determining the quality of the extracted rules is their comprehensibility. Although the main motivation for rule extraction is to obtain a comprehensible description of the underlying model's hypothesis, this aspect of rule quality is often overlooked due to the subjective nature of comprehensibility, which can not be measured independently of the person using the system. Prior experience and domain knowledge of this person play an important role in assessing the comprehensibility. This contrasts with accuracy that can be considered as a property of the rules and which can be evaluated independently of the users.

There are some traditional techniques that have been used in order to improve the comprehensibility of discovered rules. For example, we can reduce the size of the rules by constraining the number of items in the antecedent or consequent of the rule. Simplicity of the rule is related with its size, and as such, the shorter the rule is, the more comprehensible it will be. Another technique is performing a discretization of numerical values. Discretization divides the numerical data into categorical classes that are easier to understand for the instructor (categorical values are more user-friendly for the instructor than precise magnitudes and ranges).

Another way to improve the comprehensibility of the rules is to incorporate domain knowledge and semantics, and to use a common and well-know vocabulary for the teacher. In the context of web-based educational systems, we can identify some common attributes to a variety of e-learning systems such as EMS and adaptive hypermedia courses. As we can see in table 1, these attributes could be present in many sections or levels of the course. For example, a unit could be a chapter, or a lesson, or an exercise, or a collaborative resource.

Table 1. Examples of attributes common to a variety of e-learning systems.

In this context the use of standard metadata for e-learning allows the creation and maintenance of a common knowledge base with a common vocabulary susceptible of sharing among different communities of instructors. For example, the SCORM standard describes a content aggregation model and a tracking model for reusable learning objects. Although SCORM provides a framework for the representation and processing of the metadata, it falls short in including the support needed for other, more specific, pedagogical tracking such as the use of collaborative resources. "SCORM is essentially about a single-learner, self-paced and self-directed.

Finally, another proposal is to use domain specific interactive data mining in which the user is involved in the discovery process to find iteratively the most interesting results. Domain and problem specific representation are also added to the mining process. The user is not just evaluating the result of an automatic data mining process,

International Journal of Modern Agriculture, Volume 9, No. 4, 2020
ISSN: 2305-7246

but he or she is actively involved in the design of a new representation and the search for patterns.

Conclusions and Future Trends:

In our survey after applying association rule mining in Education management systems we have met certain problems and some solutions to overcome it. Future research work can be done on developing association rule mining tools that can more easily be used by educators; proposing new specific measures of interest with the inclusion of domain knowledge and semantic; embedding and integrating mining tools into EMS in order to enable the teacher to use the same interface to create/maintain courses and to carry out the mining process/obtain direct feedback/make modifications in the course; developing iterative and interactive or guided mining to help educators to apply KDD processes, or even developing an automatic mining system that can perform the mining automatically in an unattended way, such that the teacher only has to use the proposed recommendations in order to improve the students' learning.

References

1. Rice, W.H.: Moodle E-learning Course Development. A complete guide to successful learning using Moodle. Packt publishing (2006).

2. Mostow, J., Beck, J., Cen, H., Cuneo, A., Gouvea, E., Heiner, C.: An educational data mining tool to browse tutor-student interactions: Time will tell! In: Proc. of the Workshop on Educational Data Mining (2005) 15–22.

3. Klosgen, W., & Zytkow, J.: Handbook of data mining and knowledge discovery. Oxford University Press, New York (2002).

4. Agrawal R., Imielinski, T., Swami, A.N.: Mining Association Rules between Sets of Items in Large Databases. In: Proc. of SIGMOD (1993) 207-216.

5. Zaïane, O.: Building a Recommender Agent for e-Learning Systems. In: Proc. of the Int. Conf. in Education (2002) 55-59.

6. Lu, J.: Personalized e-learning material recommender system. In: Proc. of the Int. Conf. on Information Technology for Application (2004) 374–379.

7. Minaei-Bidgoli, B., Tan, P., Punch, W.: Mining interesting contrast rules for a web-based educational system. In: Proc. of the Int. Conf. on Machine Learning Applications (2004) 1-8.

8. Romero, C., Ventura, S., Bra, P. D.: Knowledge discovery with genetic programming for providing feedback to courseware author. User Modeling and User-Adapted Interaction: The Journal of Personalization Research, 14:5 (2004) 425–464.

9. Yu, P., Own, C., Lin, L.: On learning behavior analysis of web based interactive environment. In: Proc. of the Int. Conf. on Implementing Curricular Change in Engineering Education (2001) 1-10.

10. Merceron, A., & Yacef, K.: Mining student data captured from a web-based tutoring tool. Journal of Interactive Learning Research, 15:4 (2004) 319–346.

11. Freyberger, J., Heffernan, N., Ruiz, C.: Using association rules to guide a search for best fitting transfer models of student learning. In: Workshop on Analyzing Student-Tutor Interactions Logs to Improve Educational Outcomes at ITS Conference (2004) 1-10.

12. Ramli, A.A.: Web usage mining using apriori algorithm: UUM learning care portal case. In: Proc. of the Int. Conf. on Knowledge Management (2005) 1-19.