CHAPTER-3

ANALYSIS OF EDUCATIONAL DATA MINING

3.1 INTRODUCTION

Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings and using those methods to better understand students and the settings which they learn in.

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [70]. It has been seen that educational data mining methods are often different from standard data mining methods, due to the need to explicitly account for (and the opportunities to exploit) the multi-level hierarchy and non-independence in educational data. For this reason, it is increasingly common to see the use of models drawn from the psychometrics literature in educational data mining publications [71,72, 73].

An Educational Data Mining system needs to focus on the collection, archiving and analysis of data related to student learning and assessment. In current scenario an Educational Data Mining system is a very new and very small academic field. As with all new fields, EDM has grown out of existing disciplines and is spreading to overlap with new ones. Many of the researchers who are shaping EDM hail from the Intelligent Tutoring System (ITS) community, where ready access to large quantities of educational data make EDM a logical direction to advance in. EDM research shares some commonalities with the Artificial Intelligence in Education (AIED) community. The analysis performed in EDM research is often related to techniques in psychometrics and educational statistics. EDM is poised to revolutionize, or at the very least enhance and expand, the statistical methods used in
education by bringing to bear the results of decades of research in data mining and machine learning.

EDM also borrows much from the machine learning and data mining communities. In truth, the term "Educational Data Mining" is a slight misnomer in that "data mining" is generally associated with enormous datasets and much of the research is focused on developing fast and efficient algorithms for finding meaning in the data. Although there are certainly datasets with thousands or even tens of thousands of records, it is just as common to work with datasets of tens or hundreds of records. It is likely that EDM will more commonly face problems of too little data, rather than the general data mining problem of too much data. General machine learning research, especially unsupervised or semi-supervised learning has a more direct influence on EDM. It is, however, important to note that EDM does share some usability features with general data mining. Most importantly, well-developed and generalized EDM techniques should have few parameters and require little or no user intervention.

The structure of most EDM systems can be broken down into three parts: collection, archiving and analysis. Collection refers to the tools and tutoring systems used to record the relevant information, be it student scores, answers to online quizzes, or events from an Intelligent Tutoring System (ITS). Archiving is the process of storing and browsing the collected data. For score data, this is a relatively minor issue, but for the vast quantities of data generated by some ITSs this can be a significant task. Analysis brings to bear the tools of machine learning and data mining on the collected data in an attempt to gain deeper understanding of student learning, discover the relationships among questions, and possibly develop deeper quantitative understanding of cognitive processes in general. Depending on the EDM system, these three tasks will shift in relative complexity and importance, but all three must be addressed in any EDM system.
Educational Data Mining is an emerging discipline, concerned with developing methods for exploring the unique types of data that come from educational settings, and using those methods to better understand students and the settings which they learn in [74]. Data mining is extraction of interesting (non-trivial, implicit, previously unknown and potentially useful) patterns or knowledge from huge amount of data. As we know large amount of data is stored in educational database, so in order to get required data and to find the hidden relationship, different data mining techniques are developed and used. There are varieties of popular data mining tasks within the educational data mining e.g. classification, clustering, outlier detection, association rule, prediction etc. We can use the data mining in educational system as: predicting drop-out student, relationship between the student university entrance examination results and their success, predicting student's academic performance, discovery of strongly related subjects in the undergraduate syllabi, knowledge discovery on academic achievement, classification of students' performance in computer programming course according to learning style, investing the similarity and difference between schools.

3.2. Role of Association Rule in Data Model

Association rules are used to show the relationship between data items. Mining association rules allows finding rules of the form: If antecedent then (likely) consequent where antecedent and consequent are item sets which are sets of one or more items. Association rule generation is usually split up into two separate steps: First, minimum support is applied to find all frequent item sets in a database. Second, these frequent item sets and the minimum confidence constraint are used to form rules. Figure 3.1 shows the generation of item sets and frequent item sets where the minimum support count is 2.
Support and confidence are the normal method used to measure the quality of association rule. Support for the association rule X→Y is the percentage of transaction in the database that contains XUY. Confidence for the association rule is X→Y is the ratio of the number of transaction that contains XUY to the number of transaction that contain X.
Association rule can be used in educational data mining for analyzing the learning data.

### 3.3. Role of Classification in Data Model

Classification is a data mining task that maps the data into predefined groups and classes. It is also called as supervised learning. It consists of two steps:

#### 3.3.1 Model construction:

It consists of a set of predetermined classes. Each tuple/sample is assumed to belong to a predefined class. The set of tuple used for model construction is training set. The model is represented as classification rules, decision trees, or mathematical formulae. This model is shown in Figure 3.3.

#### 3.3.2 Model usage:

This model is used for classifying future or unknown objects. The known label of test sample is compared with the classified result from the model. Accuracy rate is the percentage of test set samples that are correctly classified by the model. Test set is independent of training set, otherwise over-fitting will occur. This model is shown in Figure 3.4.

In educational data mining, given works of a student, one may predicate his/her final grade. The decision tree is used to represent logical rules of student final grade.

<table>
<thead>
<tr>
<th>TID</th>
<th>ItemSet</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Bread,Butter,Milk</td>
</tr>
<tr>
<td>2</td>
<td>Bread,Milk</td>
</tr>
<tr>
<td>3</td>
<td>Bread,Coke,Diaper,Milk</td>
</tr>
<tr>
<td>4</td>
<td>Bread,Diaper,Milk</td>
</tr>
<tr>
<td>5</td>
<td>Coke,Diaper,Milk</td>
</tr>
</tbody>
</table>

Figure- 3.3: Association rule

Figure- 3.2: Association rule

{Bread}→ {Milk}

{Diaper,Milk}→ {Bread}
Figure 3.3: Learning step or model construction

Figure 3.4: Model Usage (Classification)
3.3.3 Prediction

It is used to model continuous-valued functions, i.e., predicts unknown or missing values. In this model we deduce single aspect of data from some combination of other aspect of data. In educational data mining prediction can be used to detect student behavior, predicting or understanding student educational outcomes. This model is shown in Figure 3.3.

3.3.4 Clustering

Clustering is finding groups of objects such that the objects in one group will be similar to one another and different from the objects in another group [75] Clustering can be considered the most important unsupervised learning technique. Clustering and its classification is shown in Figure 3.6.
In educational data mining, clustering has been used to group the students according to their behavior e.g. clustering can be used to distinguish active student from non-active student according to their performance in activities.

3.4. Analytical Study

In [76], author presents an approach to classifying students in order to predict their final grade based on features extracted from logged data in an education web-based system. They design, implement, and evaluate a series of pattern classifiers and compare their performance on an online course dataset. Four classifiers were used to segregate the students. The combination of multiple classifiers leads to a significant improvement in classification performance.
They used the genetic algorithm (GA) to improve the prediction accuracy and using the genetic algorithm, the accuracy of combine classifier performance is about 10 to 12% as compared to the non-GA. This method is of considerable usefulness in identifying students at risk early, especially in very large classes, and allows the instructor to provide appropriate advising in a timely manner. In paper [77], researchers analyzed how association rule are useful in Educational data mining for analyzing learning data. They explained the cosine and added value (or equivalently lift) are well suited to educational data and that teacher can interpret their result easily. They provide the case study with data from LMS (Learning Management System). In paper [75], researchers explained how data mining is useful in higher education particularly to improve the performance of the student. For that they used the Database course and also collected all available data including their usage of Model e-learning facility. They used association rule, classification rule using decision tree, clustered the student into the group using EM-clustering and using outlier analysis detected the outlier in the data. They used this knowledge to improve the performance. In paper [78], authors studied the relationship between the student university entrance examination result and their success using cluster analysis and K-means algorithm techniques. The university students were grouped according to their characteristic, forming cluster and clustering process carried out using the K-means clustering. In paper [79], researchers surveyed the application of data mining to traditional educational system particularly web-based courses, intelligent web-based educational system, learning content management system. Each of these systems used data source and objectives for knowledge discovery. In each case, data mining technique such as statistics and visualization, clustering, classification, outlier detection, association rule mining pattern mining and text mining were applied. It was explained how the cycle of applying data mining in educational system as shown below worked. The cycle of applying data mining in educational system is shown in Figure 3.7.
In [80], Kifaya explained how associative classification and clustering is effective in finding the relation and association between the students. Author evaluated the student progress according to the association between different factors using the data collected. In [81], authors analyzed the log file of elementary school student studied with science web-based module. They also produced Learn gram-the graphical representation tool that visualized the student learning process for each student.

Educational data mining methods are drawn from a variety of literatures, including data mining and machine learning, psychometrics and other areas of statistics, information visualization and computational modeling. Romero and Ventura (2007) categorize work in educational data mining into the following categories:

- Statistics and visualization
- Web mining
- Clustering, classification and outlier detection
This viewpoint is focused on applications of educational data mining to web data, a perspective that accords with the history of the research area. To a large degree, educational data mining emerged from the analysis of logs of student-computer interaction. This is perhaps most clearly shown by the name of an early EDM workshop (according to the EDM community website, the third workshop in the history of the community – the workshop at AIED2005 on Usage Analysis in Learning Systems [82]. The methods listed by Romero and Ventura (2007) as web mining methods are quite prominent in EDM today, both in mining of web data and in mining other forms of educational data.

A second viewpoint on educational data mining is given by Baker [in press], which classifies work in educational data mining as follows:

- Prediction
- Classification
- Regression
- Density estimation
- Clustering
- Relationship mining
- Association rule mining
- Correlation mining
- Sequential pattern mining
- Causal data mining
- Distillation of data for human judgment
- Discovery with models
The first three categories of Baker’s taxonomy of educational data mining methods would look familiar to most researchers in data mining: the first set of sub-categories are directly drawn from Moore’s categorization of data mining methods [83]. The fourth category, though not necessarily universally seen as data mining, accords with Romero and Ventura’s category of statistics and visualization, and has had a prominent place both in published EDM research [84], and in theoretical discussions of educational data mining [85].

The fifth category of Baker’s EDM taxonomy is perhaps the most unusual category, from a classical data mining perspective. In discovery with models, a model of a phenomenon is developed through any process that can be validated in some fashion (most commonly, prediction or knowledge engineering), and this model is then used as a component in another analysis, such as prediction or relationship mining. Discovery with models has become an increasingly popular method in EDM research, supporting sophisticated analyses such as which learning material sub-categories of students will most benefit from [86], how different types of student behavior impact students’ learning in different ways [87], and how variations in intelligent tutor design impact students’ behavior over time [88].

Historically, relationship mining methods of various types have been the most prominent category in EDM research. In Romero & Ventura’s survey of EDM research from 1995 to 2005, 60 papers were reported that utilized EDM methods to answer research questions of applied interest (according to a post-hoc analysis conducted for the current article). 26 of those papers (43%) involved relationship mining methods. 17 more papers (28%) involved prediction methods of various types. Other methods were less common. The full distribution of methods across papers is shown in Figure 3.8.
Figure 3.8. The proportion of papers involving each type of EDM method, in Romero & Ventura’s (2007) 1995-2005 survey.

(Note that papers can use multiple methods, and thus some papers can be found in multiple categories)

Educational Data Mining researchers study a variety of areas, including individual learning from educational software, computer supported collaborative learning, computer-adaptive testing (and testing more broadly), and the factors that are associated with student failure or non-retention in courses.

Across these domains, one key area of application has been in the improvement of student models. Student models represent information about a student’s characteristics or state, such as the student’s current knowledge, motivation, meta-cognition and attitudes.
Modeling student individual differences in these areas enables software to respond to those individual differences, significantly improving student learning [89]. Educational data mining methods have enable researchers to model a broader range of potentially relevant student attributes in real-time, including higher-level constructs than were previously possible. For instance, in recent years, researchers have used EDM methods to inter whether a student is gaining the system [90], experiencing poor self-efficacy [91], off-task [92], or even if a student is bored or frustrated [93]. Researchers have also been able to extend student modeling even beyond educational software, towards figuring out what factors are predictive of student failure or non-retention in college courses or in college altogether [94, 95, 96].

A second key area of application of EDM methods has been in discovering or improving models of a domain’s knowledge structure. Through the combination of psychometric modeling frameworks with space-searching algorithms from the machine learning literature, a number of researchers have been able to develop automated approaches that can discover accurate domain structure models, directly from data. For instance, [71] has developed algorithms which can automatically discover a Q-Matrix from data, and [97, 73] have developed algorithms for finding partial order knowledge structure (POKS) models that explain the interrelationships of knowledge in a domain.

A third key area of application of EDM methods has been in studying pedagogical support (both in learning software, and in other domains, such as collaborative learning behaviors), towards discovering which types of pedagogical support are most effective, either overall or for different groups of students or in different situations [86, 94]. One popular method for studying pedagogical support is learning decomposition [86]. Learning decomposition fits exponential learning curves to performance data, relating a student’s later success to the amount of each type of pedagogical support the student received up to that point. The relative weights for each type of pedagogical support, in the
best-fit model, can be used to infer the relative effectiveness of each type of support for promoting learning.

A fourth key area of application of EDM methods has been in looking for empirical evidence to refine and extend educational theories and well-known educational phenomena, towards gaining deeper understanding of the key factors impacting learning, often with a view to design better learning systems. For instance in [98] investigated the impact of self-discipline on learning and found that, whilst it correlated to higher incoming knowledge and fewer mistakes, the actual impact on learning was marginal. In [99] used the Big 5 theory for teamwork as a driving theory to search for successful patterns of interaction within student teams. In [100] investigated the relationship between consistency and student performance with the aim to provide guidelines for scaffolding instruction, basing their work on prior theory on the implications of consistency in student behavior.

3.5. Observations and Recommendations

In this section, we consider how educational data mining has developed in recent years and investigate what some of the major trends are in EDM research. In order to investigate what the trends are, we analyzed what researchers were studying previously, and what they are studying now, towards understanding what is new and what attributes EDM research has had for some time.

One way to see where EDM has been is to look at which articles were the most influential in its early years. We have an excellent resource, in Romero and Ventura's (2007) survey. This survey gives us a comprehensive list of papers, published between 1995 and 2005, which are seen as educational data mining by a prominent pair of authorities in EDM (beyond authoring several key papers in EDM, Romero and Ventura were conference chairs of EDM2009). To determine which articles were most influential, we use how many citations each paper received, a bibliometric or scientometric measure often used to indicate influence of papers, researchers, or institutions. As [102] have noted, Google Scholar, despite imperfections in its
counting scheme, is the most comprehensive source for citations—particularly for the conferences which are essential for understanding Computer Science research.

The top 8 most cited applied papers in Romero and Ventura’s survey (as of September 9, 2009) are listed in Table 1. These articles have been highly influential, both on educational data mining researchers, and on related fields; as such, they exemplify many of the key trends in our research community.

The most cited article, [103], suggests an application for data mining, using it to study on-line courses. This article proposes and evangelizes EDM’s usefulness, and in this fashion was highly influential to the formation of our community.

The second and fourth most cited articles, [104] and [105] center around how educational data mining methods (specifically association rules and clustering to support collaborative filtering) can support the development of more sensitive and effective e-learning systems. As in his other paper in this list, Zaiane makes a detailed and influential proposal as to how educational data mining methods can make an impact on e-learning systems. Tang and McCalla (2005) report an instantiation of such a system, which integrates clustering and collaborative filtering to recommend content to students. The authors present a study conducted with simulated students; successful evaluation of the system with real students is presented in [106].

The third most-cited article, [90] gives a case study on how educational data mining methods (specifically prediction methods) can be used to open new research areas, in this case the scientific study of gaming the system (attempting to succeed in an interactive learning environment by exploiting properties of the system rather than by learning the material). Though this topic had seen some prior interest (including [107, 108, 109]), publication and research into this topic exploded after it became clear that educational data mining now opened this topic to concrete, quantitative and fine-grained analysis.
The fifth and sixth most cited articles, [110,111], present tools that can be used to support educational data mining. This theme is carried forward in these groups’ later work [112,113] and in EDM tools developed by other researchers [114].

The seventh most cited article [115] shows how educational data mining prediction methods can be used to develop student models. Researchers used a variety of variables to predict whether a student will make a correct answer. This work has inspired a great deal of later educational data mining work.

Student modeling is a key theme in modern educational data mining, and the paradigm of testing EDM models’ ability to predict future correctness, advocated strongly by Beck & Woolf, has become very common (e.g. [115, 116].

Table 1. The top 8 most cited papers, in Romero & Ventura’s 1995-2005 survey. Citations are from Google Scholar, retrieved 9 September, 2009.
As discussed earlier, relationship mining methods of various types were the most prominent type of EDM research between 1995 and 2005. 43% of papers in those years involved
relationship mining methods. Prediction was the second most prominent research area, with 28% of papers in those years involving prediction methods of various types. Human judgment/exploratory data analysis and clustering followed with (respectively) 17% and 15% of papers.

A very different pattern is seen in the papers from the first two years of the Educational Data Mining conference [117, 118], as shown in Figure 2.9. Whereas relationship mining was dominant between 1995 and 2005, in 2008-2009 it slipped to fifth place, with only 9% of papers involving relationship mining. Prediction, which was in second place between 1995 and 2005, moved to the dominant position in 2008-2009, representing 42% of EDM2008 papers. Human judgment/exploratory data analysis and clustering remain in approximately the same position in 2008-2009 as 1995-2005, with (respectively) 12% and 15% of papers.

A new method, significantly more prominent in 2008-2009 than in earlier years, is discovery with models. Whereas no papers in Romero & Ventura's survey involved discovery with models, by 2008-2009 it has become the second most common category of EDM research, representing 19% of papers.

Another key trend is the increase in prominence of modeling frameworks from Item Response Theory, Bayes Nets, and Markov Decision Processes. These methods were rare at the very beginning of educational data mining, began to become more prominent around 2005 (appearing, for instance, in [71, 72], and were found in 28% of the papers in EDM2008 and EDM2009. The increase in the commonality of these methods is likely a reflection of the integration of researchers from the psychometrics and student modeling communities into the EDM community.
Figure 2.9: The proportion of papers involving each type of EDM method, in the proceedings of Educational Data Mining 2008 and 2009 [48, 49].

It is worth noting that educational data mining publications in 2008 and 2009 are not limited solely to those appearing in the proceedings of the conference (though our analysis in this paper was restricted to those publications). One of the notable metrics of our community’s growth is that the proceedings of EDM2008 and EDM2009 alone accounted for approximately as many papers as were published in the first 10 years of the community’s existence (according to Romero & Ventura’s review). Hence, EDM appears to be growing in size rapidly and the next major review of the field is likely to be a time-consuming process. However, we encourage future researchers to conduct such a survey. In general, it will be very interesting to see how the methodological trends exposed in Figures 1 and 2 develop in the next few years.
One interesting difference between the work in EDM2008 and EDM2009, and earlier educational data mining work, is where the educational data comes from. Between 1995 and 2005, data almost universally came from the research group conducting the analysis—that is to say, in order to do educational data mining research, a researcher first needed to collect their own educational data.

This necessity appears to be disappearing in 2008, due to two developments. First, the Pittsburgh Science of Learning Center has opened a public data repository, the PSLC Data Shop [119], which makes substantial quantities of data from a variety of online learning environments available, for free, to any researcher worldwide. 14% of the papers published in EDM2008 and EDM2009 utilized data publicly available from the PSLC Data Shop.

Second, researchers are increasingly and frequently instrumenting existing online course environments used by large numbers of students worldwide, such as Model and Web CAT. 12% of the papers in EDM2008 and EDM2009 utilized data coming from the instrumentation of existing online courses.

Hence, around a quarter of the papers published at EDM2008 and EDM2009 involved data from these two readily available sources. If this trend continues, there will be significantly benefits for the educational data mining community. Among them, it will become significantly easier to externally validate an analysis. If a researcher does an analysis that produces results that seem artifactual or "too good to be true," another researcher can download the data and check for themselves. A second benefit is that researchers will be more able to build on others' past efforts. As reasonably predictive models of domain structure and student moment-to-moment knowledge become available for public data sets, other researchers will be able to test new models of these phenomena in comparison to a strong baseline, or to develop new models of higher grain-size constructs that leverage these existing models. The result is a science of education that is more concrete, validated and progressive than was previously possible.