An Empirical Study of the Applications of Data Mining Techniques in Higher Education

Dr. Varun Kumar
Department of Computer Science and Engineering
ITM University
Gurgaon, India
kumarvarun333@gmail.com

Anupama Chadha
Department of Computer Science and Engineering
ITM University
Gurgaon, India
anupamaluthra@gmail.com

***Abstract*— Few years ago, the information flow in education field was relatively simple and the application of technology was limited. However, as we progress into a more integrated world where technology has become an integral part of the business processes, the process of transfer of information has become more complicated. Today, one of the biggest challenges that educational institutions face is the explosive growth of educational data and to use this data to improve the quality of managerial decisions. Data mining techniques are analytical tools that can be used to extract meaningful knowledge from large data sets. This paper addresses the applications of data mining in educational institution to extract useful information from the huge data sets and providing analytical tool to view and use this information for decision making processes by taking real life examples.**

Keywords- Higher education,; Data mining; Knowledge discovery; Classification; Association rules; Prediction; Outlier analysis;

I. INTRODUCTION

In modern world a huge amount of data is available which can be used effectively to produce vital information. The information achieved can be used in the field of Medical science, Education, Business, Agriculture and so on. As huge amount of data is being collected and stored in the databases, traditional statistical techniques and database management tools are no longer adequate for analyzing this huge amount of data.

Data Mining (sometimes called data or knowledge discovery) has become the area of growing significance because it helps in analyzing data from different perspectives and summarizing it into useful information. [1]

There are increasing research interests in using data mining in education. This new emerging field, called Educational Data Mining, concerns with developing methods that discover knowledge from data originating from educational environments [1].

The data can be collected from various educational institutes that reside in their databases. The data can be personal or academic which can be used to understand students' behavior, to assist instructors, to improve teaching, to evaluate and improve e-learning systems , to improve curriculums and many other benefits.[1][2]

Educational data mining uses many techniques such as decision trees, neural networks, k-nearest neighbor, naive bayes, support vector machines and many others.[3]

Using these techniques many kinds of knowledge can be discovered such as association rules, classifications and clustering. The discovered knowledge can be used for organization of syllabus, prediction regarding enrolment of students in a particular programme, alienation of traditional classroom teaching model, detection of unfair means used in online examination, detection of abnormal values in the result sheets of the students and so on.

This paper is organized as follows: Section II describes the related work. Section III describes the research question. Section IV describes data mining techniques adopted. Section V discusses the application areas of these techniques in an educational institute. Section VI concludes the paper.

II. RELATED WORK

Data mining in higher education is a recent research field and this area of research is gaining popularity because of its potentials to educational institutes.

[1] gave case study of using educational data mining in Moodle course management system. They have described how different data mining techniques can be used in order to improve the course and the students' learning. All these techniques can be applied separately in a same system or together in a hybrid system.

[2] have a survey on educational data mining between1995 and 2005. They have compared the Traditional Classroom teaching with the Web based Educational System. Also they have discussed the use of Web Mining techniques in Education systems.

[3] have a described the use of k-means clustering algorithm to predict student's learning activities. The information generated after the implementation of data mining technique may be helpful for instructor as well as for students.

[4] discuss how data mining can help to improve an education system by enabling better understanding of the students. The extra information can help the teachers to manage their classes better and to provide proactive feedback to the students.

[6] have described the use of data mining techniques to predict the strongly related subject in a course curricula. This information can further be used to improve the syllabi of any course in any educational institute.

[8] describes how data mining techniques can be used to determine The student learning result evaluation system is an essential tool and approach for monitoring and controlling the learning quality. From the perspective of data analysis, this paper conducts a research on student learning result based on data mining.

III. Research Object

The object of the present study is to identify the potential areas in which data mining techniques can be applied in the field of Higher education and to identify which data mining technique is suited for what kind of application.

IV. Data Mining Definition and Techniques

Simply stated, data mining refers to extracting or “mining" knowledge from large amounts of data. [5] Data mining techniques are used to operate on large volumes of data to discover hidden patterns and relationships helpful in decision making. The sequences of steps identified in extracting knowledge from data are: shown in Figure 1.

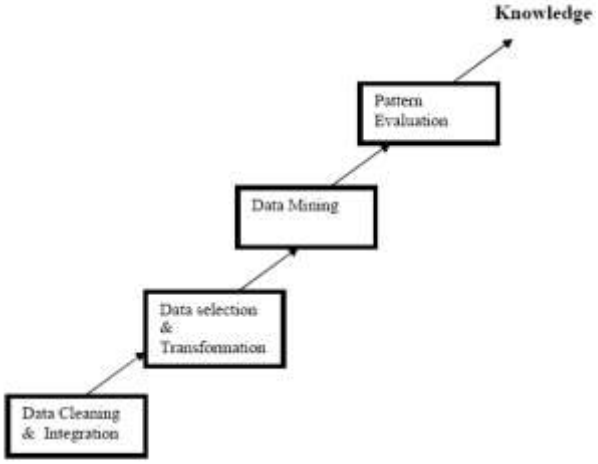


Figure 1. The steps of extracting knowledge from data

The various techniques used in Data Mining are:

A. Association analysis

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis.

More formally, association rules are of the form $X \Rightarrow Y, i.e., "A_1 \wedge ----\wedge A_m \rightarrow B_1 \wedge ----\wedge B_n"$, where Ai (for i to m) and Bj (j to n) are attribute-value pairs.

(I)

The association rule X=>Y is interpreted as database tuples that satisfy the conditions in X are also likely to satisfy the conditions in Y ".

B. Classification and Prediction

Classification is the processing of finding a set of models (or functions) which describe and distinguish data classes or concepts, for the purposes of being able to use the model to predict the class of objects whose class label is unknown. The derived model may be represented in various forms, such as classification (IF-THEN) rules, decision trees, mathematical formulae, or neural networks. Classification can be used for predicting the class label of data objects. However, in many applications, one may like to predict some missing or unavailable data values rather than class labels. This is usually the case when the predicted values are numerical data, and is often specifically referred to as prediction.

IF-THEN rules are specified as **IF condition THEN conclusion**

e.g. IF age=youth and student=yes then buys_computer=yes

C. Clustering Analysis

Unlike classification and predication, which analyze class-labeled data objects, clustering analyzes data objects without consulting a known class label. In general, the class labels are not present in the training data simply because they are not known to begin with. Clustering can be used to generate such labels. The objects are clustered or grouped based on the principle of maximizing the intraclass similarity and minimizing the interclass similarity.

That is, clusters of objects are formed so that objects within a cluster have high similarity in comparison to one another, but are very dissimilar to objects in other clusters. Each cluster that is formed can be viewed as a class of objects, from which rules can be derived. [5]

Application of clustering in education can help institutes group individual student into classes of similar behavior. Partition the students into clusters, so that students within a cluster (e.g. Average) are similar to each other while dissimilar to students in other clusters (e.g. Intelligent, Weak).

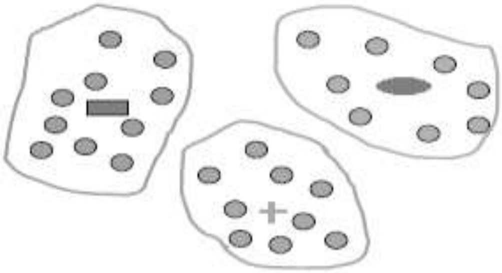


Figure 2. Picture showing the partition of students in clusters

D. Outlier Analysis

A database may contain data objects that do not comply with the general behavior of the data and are called outliers. The analysis of these outliers may help in fraud detection and predicting abnormal values.

V. POTENTIAL APPLICATIONS

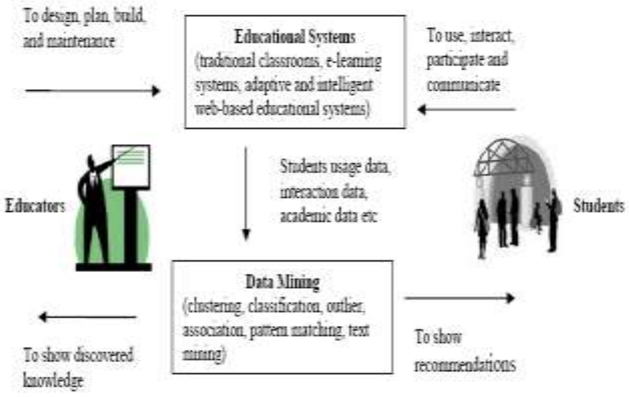

To design, plan, build, and maintenance
Educational Systems
(traditional classrooms, e-learning systems, adaptive and intelligent web-based educational systems)
To use, interact, participate and communicate
Students usage data, interaction data, academic data etc
Educators
Students
Data Mining
(clustering, classification, outlier, association, pattern matching, text mining)
To show discovered knowledge
To show recommendations

Figure 3. The cycle of applying data mining in education system [4]

The above figure illustrates how the data from the traditional classrooms and web based educational systems can be used to extract knowledge by applying data mining techniques which further helps the educators and students to make decisions.

A. Organization of Syllabus

It is important for educational institutes to maintain a high quality educational programme which will improve the student's learning process and will help the institute to optimize the use of resources. A typical student at the university level completes a number of courses (i.e. "course" and "subject" are used synonymously) prior to graduation.

Presently, organization of syllabi is influenced by many factors such as affiliated, competing or collaborating programmes of universities, availability of lecturers, expert judgments and experience. This method of organization may not necessarily facilitate students' learning capacity optimally. Exploration of subjects and their relationships can directly assist in better organization of syllabi and provide insights to existing curricula of educational programmes.

One of the application of data mining is to identify related subjects in syllabi of educational programmes in a large educational institute.[6]

A case study has been performed where the student data collected over a period of time at the Sri Lanka Institute of Information Technology (SLIIT) [7]. The main aim of the study was to find the strongly related subjects in a course offered by the institute. For this purpose following methodology was followed to:

- Identify the possible related subjects.
- Determine the strength of their relationships and determine strongly related subjects.

METHODOLOGY

In the first step, association rule mining is used to identify possibly related two subject combinations in the syllabi which also reduce our search space. In the second step(see[6]), Pearson Correlation Coefficient was applied to determine the strength of the relationships of subject combinations identified in the first step.

TABLE I. THE SUBJECTS CHOSEN BY STUDENTS

Student id	Subject 1	Subject 2	Subject 3
1	Databases	Advanced Databases	Data mining
2	Databases	Advanced Databases	Data mining
3	Databases	Advanced Databases	Data mining
4	Databases	Advanced Databases	Visual Basic
5	Databases	Advanced Databases	Web Designing

Association Rules that can be derived from Table 1 are of the form:

$$(X, subject1) \Rightarrow (X, subject2) \qquad (2)$$

$$(X, subject1)^\wedge(X, subject2) \Rightarrow (X, subject3) \quad (3)$$

$$(X, "Databases") \Rightarrow (X, "AdvancedDatabases")$$

[support=2% and confidence=60%] (4)

$$(X, "Databases")^\wedge(X, "AdvancedDatabases" \Rightarrow (X, "DataMining")$$

[support=1% and confidence=50%] (5)

Where support factor of the association rule shows that 1% of the students have taken both the subjects "Databases" and "Advanced Databases" and confidence factor shows that there is a chance that 50% of the students who have taken "Databases" will also take "Advanced Databases"

This way we can find the strongly related subjects and can optimize the syllabi of an educational programme.

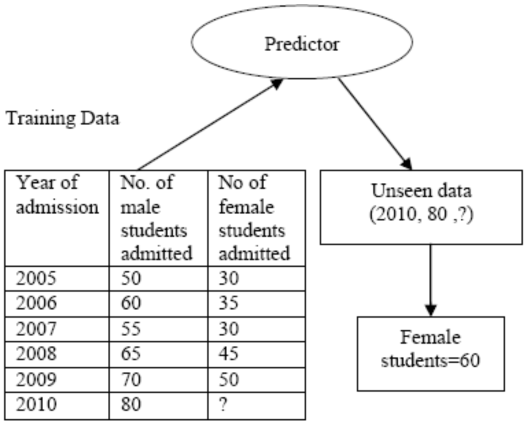
B. Predicting The Registration Of Students in an Educational Programme

Now a days educational organization are getting strong competition from other Academic competitors. To have an edge over other organizations, needs deep and enough knowledge for a better assessment, evaluation, planning, and decision making.

Data Mining helps organizations to identify the hidden patterns in databases; the extracted patterns are then used to build data mining models, and hence can be used to predict performance and behavior with high accuracy. As a result of this, universities are able to allocate resources more effectively.

METHODOLOGY

One of the application of data mining can be for example, to efficiently assign resources with an accurate estimate of how many male or female will register in a particular program by using the Prediction techniques.



Year of admission	No. of male students admitted	No of female students admitted
2005	50	30
2006	60	35
2007	55	30
2008	65	45
2009	70	50
2010	80	?

Figure 4. Prediction of female students in the coming year

In real scenario couple of other associated attributes like type of course, transport facility, hostel facility etc can be used to predict the registration of students.

C. Predicting Student Performance

One of the question whose answer almost every stakeholder of an educational system would like to know "Can we predict student performance?"

Over the years, many researchers applied various data mining techniques to answer this question.

In modern times, learning is taking on a more important role in the development of our civilization. Learning is an individual behavior as well as a social phenomenon.

It is a difficult task to deeply investigate and successfully develop models for evaluating learning efforts with the combination of theory and practice. University goals and outcomes clearly relate to "promoting learning through effective undergraduate and graduate teaching, scholarship, and research in service to University."

Student learning is addressed in some goals and outcomes related to the development of overall student knowledge, skill, and dispositions. Collections of randomly selected student work are examined and assessed by small groups of faculty teaching courses within some general education categories. [8]

With the help of data mining techniques a result evaluation system can be developed which can help teachers and students to know the weak points of the traditional classroom teaching model. Also it will help them to face the rapidly developing real-life environment and adapt the current teaching realities.

METHODOLOGY

We can use student participation data as part of the class grading policy. An instructor can assess the quality of student by conducting an online discussion among a group of students and use the possible indicators such as the time difference between posts, frequency distribution of the postings, duration between postings and replies etc.

Given this data, we can apply classification algorithms to classify the students into possible levels of quality.

Table II. Student participation data and their grades

Time difference between posts (in min)	Duration between postings and replies (in min)	Grade of the student
3	1	A
3	2	B
4	1	A
5	2	B
6	1	C
6	2	C

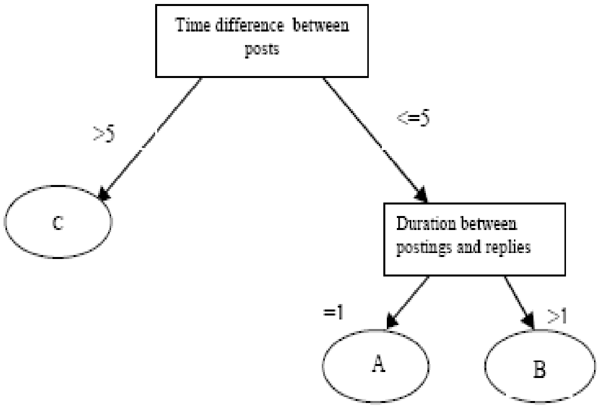

Time difference between posts
>5
<=5
C
Duration between postings and replies
=1
>1
A
B

Figure 5. The Decision Tree built from the data in Table II

D. Detecting Cheating in Online Examination

We can say that online assessments are useful to evaluate students' knowledge; they are used around the world in schools -since elementary to higher education institutions- and in recognized training centers like the Cisco Academy [9].

Now a days exams are conducted online remotely through the Internet and if a fraud occurs then one of the basic problems to solve is to know: who is there? Cheating is not only done by students but the recent scandals in business and journalism show that it has become a common practice.

Data Mining techniques can propose models which can help organizations to detect and to prevent cheats in online assessments. The models generated use data comprising of different student's personalities, stress situations generated by online assessments, and common practices used by students to cheat to obtain a better grade on these exams.

E. Identifying Abnormal/ Erroneous Values

The data stored in a database may reflect outliers-|noise, exceptional cases, or incomplete data objects. These objects may confuse the analysis process, causing over fitting of the data to the knowledge model constructed. As a result, the accuracy of the discovered patterns can be poor. [5]

One of the applications of Outlier Analysis can be to detect the abnormal values in the result sheet of the students. This may be due many factors like a software fault, data entry operator negligence or an extraordinary performance of the student in a particular subject.

Table III. The result of students in four subjects

Student roll no	Marks in subject1	Marks in subject2	Marks in subject3	Marks in subject4
101	30	35	45	30
102	67	75	78	67
103	89	90	78	77
104	30	35	45	**99**

In the table shown above the result of the student in subject4 with roll no 104 will be detected as an exceptional case and can be further analyzed for the cause.

VI. Conclusion

In the present study, we have discussed the various data mining techniques which can support education system via generating strategic information.. Since the application of data mining brings a lot of advantages in higher learning institution, it is recommended to apply these techniques in the areas like optimization of resources, prediction of retainment of faculties in the university, to find the gap between the number of candidates applied for the post, number of applicants responded, number of applicants appeared, selected and finally joined. Hopefully these areas of application will be discussed in our next paper.

References

[1] C. Romero, S. Ventura, E. Garcia, "Datamining in course management systems: Moodle case study and tutorial", Computers & Education, Vol. 51, No. 1, pp. 368-384, 2008

[2] C. Romero, S. Ventura "Educational data Mining: A Survey from 1995 to 2005", Expert Systems with Applications (33), pp. 135-146, 2007

[3] Shaeela Ayesha, Tasleem Mustafa, Ahsan Raza Sattar, M. Inayat Khan, **“**Data Mining Model for Higher Education System”, Europen Journal of Scientific Research, Vol.43, No.1, pp.24-29, 2010

[4] K. H. Rashan, Anushka Peiris, “Data Mining Applications in the Education Sector”, MSIT, Carnegie Mellon University, retrieved on 28/01/2011

[5] Han Jiawei, Micheline Kamber, *Data Mining: Concepts and Technique.* Morgan Kaufmann Publishers,2000

[6] W.M.R. Tissera, R.I. Athauda, H. C. Fernando “Discovery of Strongly Related Subjects in the Undergraduate Syllabi using Data Mining”, IEEE International Conference on Information Acquisition, 2006

[7] Sri Lanka Institute of Information Technology, *http://www.sliit.lk/*, retrieved on 28/02/2011

[8] Sun Hongjie, “Research on Student Learning Result System based on Data Mining”, IJCSNS International Journal of Computer Science and Network Security, Vol.10, No. 4, April 2010

[9] Academy Connection – Training Resources In html, http:*//www.cisco.com/web/learning/netacad/index*, December 28th, 2005.

[10] Wayne Smith, "Applying Data Mining to Scheduling Courses at a University", Communications of the Association for Information Systems, Vol. 16, Article 23, 2005

[11] Firdhous, M. F. M. (2010). Automating Legal Research through Data Mining. International Journal of Advanced Computer Science and Applications - IJACSA, 1(6), 9-16.

[12] The Result Oriented Process for Students Based On Distributed Data Mining. (2010). International Journal of Advanced Computer Science and Applications - IJACSA, 1(5), 22-25.

[13] Jadhav, R. J. (2011). Churn Prediction in Telecommunication Using Data Mining Technology. International Journal of Advanced Computer Science and Applications - IJACSA, 2(2), 17-19.

Authors Profile

Presently a Research student at ITM Gurgaon , she is armed with a degree in Electronics followed by a Master of Technology and has had an accomplished academic record all through her education and career. After having a stint in software development she has taught for over 11 years to PG and UG students in Computer Applications and Information Technology. She also has to her credit a number of systems improvement projects for her previous employers besides having administrative experience relating to functioning of placement cell and coordination of various academic programmes. A distinguished faculty, her field of specialization is Software Engineering, Data Mining and Object Oriented Analysis and Design concepts.