Data Mining Based Decision Support System for Students

Sajid Iqbal[1]
Dept. of Computer Sci.
Bahaudding Zakariya University, Multan, Pakistan
sajidiqbal.pk@gmail.com[1]

Dr. Shehbaz M.[2] Waqar M. M.[3]
Dept. of Computer Sci. and Engineering University of Engineering and Technology Lahore, Pakistan.
dr.muhammad *shahbaz@gmail.com[2]*,
mirzamwaqar@hotmail.com[3]

***Abstract* — Students have to make many decisions in their educational tenure. The study program offered at university level generally consists of two type of courses: Core and Electives. The core courses are compulsory for students to enroll whereas in elective subjects they are provided with options to choose according to their interest or specialization. Sometime it becomes bit difficult for students to formulate right decision in course selection according to their strengths and weakness. The proposed system help the students, at this course selection stage, with list of courses that are preferable for them based on their previous course performances. This recommendation system is based on ID3 decision tree algorithm and simply bases its decisions on probability calculations.**

Keywords: — Educational Data mining; classification; decision tree, knowledge discovery, data patterns.

I. INTRODUCTION

Educational data mining (EDM) is, now a days, a hot research area. The objective is to analyze the huge amount of educational data by using data mining techniques and extract useful and intelligent information that can help students and institutes to improve their performance and results [10]. This data is collected from the databases maintained by educational institutes that consists of both personal and academics information of students.

Today, the higher educational institutes not only attempt to provide quality education but also guide the students in their educational and career oriented decisions. This quality of education can be improved by measuring and analyzing students' previous performance quantitatively and discovering useful knowledge available in their educational databases. One use of this extracted knowledge is to guide the students in their decision making about right course/subject selection in the form of recommendations that can improve their overall performance and understanding. Generally, when students are offered a list of courses and they have the option to select among these offered courses, they base their selection on different unrelated or semi related factors like most class fellows taking which course? do I feel easy[1] with particular tutor? or the course is easy? Such decisions should be based on students' own aptitude and their performance in previous courses of related domain. This correlation provides the solid bases for students' better understanding and performance in recommended courses.

By applying the data mining algorithms on historical data of old students, new patterns can be discovered [11] that can help universities to estimate area of specialization of enrolled students. Using these patterns, the data mining models are built to predict individual behavior in a particular domain with high accuracy and help students in risk free decision making.

The universities design their degree programs consists of two major types of courses 1) Core Course 2) Electives. Core courses are basic courses of the domain which provide fundamental domain knowledge. For example in Computer Science degree program, the core courses are Introduction to Computing, Introduction to Programing, Operating Systems, Algorithms etc. and it is compulsory for all students, enrolled in that degree program, to study these core courses. The Higher Education Institutes (HEI) keep a pool of elective courses, defined in their study schemes, from which students can select number of courses according to their degree credit hours requirement.

It is not compulsory for students to enroll in all offered optional courses but this selection designates the area of specialization. In this decision making, students consult their seniors and fellows to get suggestions or follow the general trend without considering their own strengths. Therefore such decisions based on un-related factors lead students to either poor performance or failure in completing the courses resulting in dropping the degree program or completing with low grades. The low performance is due to lack of student personalized academic analysis. This gap indicates the needs of an intelligent system that can helps students in such situations and guide them to base their decisions on quantitative measurements.

We propose an intelligent system that helps the students by providing recommendations based on quantitative measures. The system takes the old students data and by use of data mining algorithms, suggests the best options to students appropriate for them in the particular scenario. The major objective of this system is to fetch a vital knowledge from huge

[1] Published in First International Young Engineering Convention FEIIC-IYEC 18-20 April, 2014, at University of Engineering and Technology, Lahore

http://www.uet.edu.pk/newsannouncement/newssection/uet_news.html?id=98

amount of old students' academic performance records and build a list of recommendations. Achieving good grades in majors directly impacts the overall grade of students and improvement in all these factors directly advances the excellence of higher learning institutes. The proposed system is equally useful for student and institutes.

The structure of this paper is as follows. The next section presents literature review. In section III, the proposed model of data mining designed for this system is explained. Section IV discusses the case study. In section V data analysis and relevant results are given and finally in section VI the conclusion and future work is presented.

II. Related Work

Educational data mining (EDM) has become one of the hot research areas in last decade. Specialized courses like "Big Data in Education" [16] are being offered. Different universities have established their research groups [18], specialized conferences [17] and journals [19] are being published in EDM. EDM is being applied for different educational tasks i.e. prediction modeling, behavior analysis, relationship modeling and visualization. [3] Discusses major motivational factors for the application of DM in education i.e. to study analysis of dependencies between student assessment scores and final grades and formation of student project teams. Although lots of data mining tasks in education have been carried out using data sets collected from conventional learning systems. Still there is requirement of educational data analysis due to emergence of new learning mechanisms like e-learning, self-learning, blended and distance learning. In last century, education was considered as social service, a non-commercial activity, but for last couple of decades education has been emerged in the form of industry generating huge commercial activity [20]. This transformation has further widened the gap resulting in increased research requirement.

The techniques developed for commercial data mining have been successfully applied on educational data. [2] Presents few example applications of data mining in education with mapping of corporate world DM problems to education. One of the application, among few discussed in [2], is predicting the paths of students and alumni. An institute may need to know the answers of questions like which students will enroll in a particular program? Which students need assistance in their studies? Which type of students get their study program changed? And which types of students remain active with institute as alumni? Another application could be comprehensive analysis of student characteristics that is creating meaningful learning outcome typologies like what are different types of students who affect the institute enrollment in positive or negative way? and as a result of EDM the institutes can deploy resources and staff more effectively [3]. Other potential application areas may be academic planning and interventions, transfer predictions, drop outs and re-enrollment and designing marketing strategies using descriptive or predictive data analysis. EDM has also been used in measuring faculty performance using student feedback mechanisms and developing performance metrics [5]. These performance indicators are based on alumnae activity planning, course assessment, student assessment and student major counseling.

Commercial Data Mining (DM) is carried in four steps 1) Classification, 2) Categorization 3) Estimation 4) Visualization, with two major data mining approaches i) supervised and ii) unsupervised. OLTP data is transferred into OLAP data using Extract, Transform and Load (ETL) process and then one of the data mining technique is applied.

Researchers have used different DM algorithms for EDM like validity of application of DM in higher education using K-means algorithm [1]. [2] has used two step k-means, machine learning, and neural-net methods of DM to find out patterns. [4] Discusses the use of fuzzy mathematics in evaluating the teaching attitude, teaching content, teaching methods and teaching effects in English language teaching using the college data. Prediction, cluster analysis, classification and association classification methods of DM to extract information are used in [5]. The authors in [6] focus on discovering pedagogical relevant knowledge contained in data bases from web based educational systems to help the instructors to better understand the learning process of their students. For DM, [6] has used clustering, classification and association rules. Decision trees for classifying student evaluation data to analyze the student drop out and expected performance are used in [7]. Another interesting work is presented in [8] where authors have used DM with Association Rules to find out weaker students in class who may need remedial classes. [9] Has performed similar work as that of [1] but using different DM techniques i.e. decision trees and J48 algorithm.

Conventional education systems are different from the newly emerged IT tool assisted systems where availability of multi-dimensional data is rare. Most of the research in EDM has been done using web based learning systems or complete learning management systems. In this work, we have studied the conventional interactive learning system where available data is in unstructured and poor form.

III. Model Design for Student's Decision Support System

Today, universities and other higher education institutes are operating in highly competitive environment where quality of education is main business development factor but at the same time the institutes need to sustain students, once enrolled, along with exploring new markets. To remain competitive, they have to work on finding new techniques and tools for quicker analysis and solutions of their problems.

The proposed system intends to provide help for institutes in finding out the ways to improve their progression. It also helps the students to select those optional courses in which they may perform better by obtaining good grades leading to overall institutional education quality improvement.

The proposed system used the data set of computer science graduate students of Bahauddin Zakariya University, Multan, Pakistan [12]. The data set consists of 1600 students' data spread over 10-year duration from 2003 to 2013 with record count reaching up to 100,000. The proposed system can be implementing in any department and in any level of education.

The system consists of three main modules discussed as follow (see Fig 1):

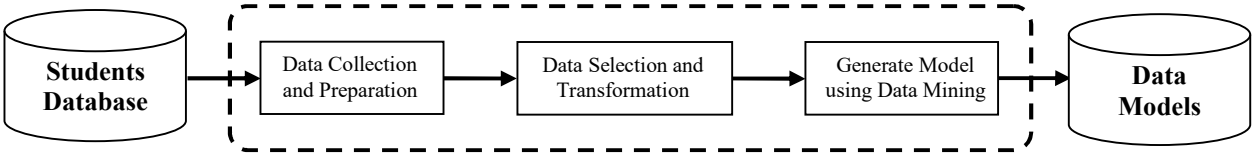

Fig. 1. Abstract architecture of Student's Decision Support System

A. Data Collection and Preparation

Data preprocessing is one of the major process before data mining operation. To get better results of data mining techniques, some preprocessing is performed on data set to prepare data as a decent input to a system [13]. The data in original database is organized in semi normalized form stored in independent tables. In the first step we collected all useful attributes based on our objective, from different tables and stored them into a single table after performing joining process on those tables. The attributes collected from different tables are related to student's demographic and academic information.

Data collected from different sources into one target database normally contains redundancy with missing values, noises and outliers. So the next step is to remove all possible errors produced in data by using data cleaning operation to produce good structured quality data. The selected attributes belong to quantitative category. The redundant data is discarded, inconsistent data is replaced with most probable value and the missing values are replaced in two ways either by calculating again (if those are calculated/formula fields) or putting most probable values.

B. Data Selection and Transformation

Once the data has been prepared the next step is to further explore that collected attributes by identifying their meanings and descriptions and also determining the relation of different attributes. This operation helps in selecting a right attribute because choosing a right attribute in data mining leads to a good data model. So in this step we selected only those attributes which were required for our system. Once the attributes are selected then the next step is to transform the values of attributes if necessary. The GPA of a core course of an individual student is in the form of continues value. So to make the processing to generate model of that data, these kinds of attributes is transformed into categorical form. As an example we transformed all GPAs values into three qualitative values i.e. High, Average and Low. Table 1 shows the name, description and possible values of selected attributes for our system.

The number of core courses is depending on the graduate program for which a data model is created. We selected only those core courses which are related to program domain, not those which are studied for supporting the program. For our case we selected computer science program as case study so the domain based core courses of computer science are Programming, Data structure, Algorithm Analysis, Theory of automata, Computer Architecture etc. Here our concern is only study program core courses because these courses provide foundation for specialized courses. For each specialize course a data model is generated and stored. Whenever a student want to know about any specialize course he should take or not, the relevant model is fetched from database and results is presented.

TABLE I. ATTRIBUTES LIST OF STUDENT'S DECISION SUPPORT SYSTEM

Attribute Name	Description	Possible Value
Gender	Gender of Student	M, F
Province	Province of student	Punjab, KPK, Sindh, Baluchistan
Core_1	1st Core subject selected	H, A, L
Core_2	2nd Core subject selected	H, A, L
Core_3	3rd Core subject selected	H, A, L
Core_4	4th Core subject selected	H, A, L
….	…..	…..
Core_n	nth Core subject selected	H, A, L
Specialize_course	Class Attribute	H, A, L

C. Generate Model using Data Mining

Once the preprocessing is performed, the obtained attributes from the previous phase are used as input to this phase to generate a model. In the proposed system we used ID3 algorithm [14] to generate decision tree as data model. In ID3 algorithm the decision tree is constructed by employing top down greedy search on the selected data to evaluate each attribute at every node of tree.ID3 algorithm uses information gain function to select most appropriate attribute as tree node for optimum data split. At each level of tree, the information gain of all possible attributes is calculated and the attribute which have the maximum gain is chosen as splitting attribute. All possible values of that attribute generate tree branches. The advantage of using information gain function is that it splits the data in most balanced way and also it minimizes the depth of the tree. The tree constructed using Information Gain function is very optimal and efficient.

The information gain G(A) can be defined as the difference of the entropy of original data set with the entropy of data set when it is splits on the bases of an attribute A. The information gain G(A) of an attribute A relevant to data set D is calculated as:

$$G(A) = E(D) - \sum_{v \in V} \frac{|D_v|}{|D|} E(D_v)$$

E(D) is Entropy of data set D that is used to measure the degree of impurity [15]. V is the set of all possible values of attribute A. D_v is the set of all records where the attribute A has value v. $E(D_v)$ is the Entropy of subset D_v and $\frac{|D_v|}{|D|}$ is the

proportion of the number of records in in data set having value v to the number of records in data set D.

The Entropy needed to calculate information gain of data set D is calculated as:

$$E(D) = \sum_{i \in c} - p_i \log_2 p_i$$

Where c is the set of all possible values of class attribute and p_i is the probability of data set having class value i to data D. Here the log is used with base 2 because the information is encoded in bits. If the data table is pure i.e. it has a single value for all records of class attributes, then the entropy is 0 as the probability is 1 and log(1) = 0. In other case when all the values of class attribute have equal probability then the entropy reaches to its maximum. The process of selecting an attribute for partitioning the tree is repeated for each non terminal descendant node of the tree and continued till either every attribute have been added in a current path of that tree or all the values of the attribute at that particular node are same.

In our case we have created number of data models for each specialized subject offered in a particular program and stored each in system's database for future use. Whenever a student wants to know what will be his expected progress in any specialized subject, the relevant model is fetched from the system, attributes of that student are parsed through the model and result is presented. The complete procedure of ID3 algorithm is given below.

```
ID3_tree(Dataset, AttrList)
{
  IF Dataset = NULL THEN
    Return Null
  END IF

  N = new node
  IF AttrList = NULL THEN
    Label N as leaf with most common value of class attribute
    Return

  ELSE IF all rows in Dataset has same class value THEN
    Label N as leaf node with that class value
    Return

  ELSE
    HGA = NULL   // Attribute having highest Gain
    FOREACH attribute A in AttrList
      Calculate gain(A)
      IF HGA < gain(A) THEN
        HGA = A
      END IF
    END FOR
    Label N with HGA
    REMOVE HGA from AttrList
    FOREACH value V in HGA
      CREATE a branch with label V
      CREATE Subset from Dataset having value V
      IF Subset = NULL THEN
        ATTACH leaf having most common value for HGA
      ELSE
        ATTACH subtree ID3_tree(Subset, AttrList)
      END IF
    END FOR

  END IF
  Return Root
}
```

IV. A CASE STUDY

As case study, we used the database of BZU's students. At first level of this system we selected students of computer science department. The dataset contains 1600 records of different students enrolled in past 10 years. This data is further divided into sub parts according to the specialization courses studied by students because not all students took all specialized courses offered in a term/semester. As an example the students who studied "Web Engineering" is 656 and similarly the students who taken "Expert System" were 288 etc. The proposed system have created the model of each specialized course offered by university on the bases of their dataset. To describe the process, we just show the model of one subject named "Data Mining" and we have selected 6 core courses which are: Programming Fundamentals (PF), Data Structures and Algorithms (DSA), Operating Systems (OS), Introduction to Database Systems (IDB), Introduction to Software Engineering (ISE) and Artificial Intelligence (AI).

The number of students who enrolled in Data Mining in past 10 years were 423 from which we selected 350 records to generate model and train it. Remaining 73 were used as test data. The proposed algorithm uses unsupervised technique for predictive modeling. Table 2 shows the sample data selected for creating model for Data Mining course.

TABLE II. DATASET OF STUDENT TOOK DATA MINING

Sr.	Gender	State	PF	DSA	OS	IDB	ISE	AI	DM
1	M	Punjab	H	L	M	H	H	H	H
2	M	KPK	H	M	M	H	M	H	H
3	F	Punjab	H	M	H	H	L	L	L
4	M	Sindh	M	H	M	H	L	M	H
5	M	Punjab	L	M	H	L	M	H	M
6	F	Punjab	H	L	L	M	H	M	M
.	.	.	.	.	.	.	.	.	.
.	.	.	.	.	.	.	.	.	.
350	M	Bolichistan	L	M	L	M	M	H	H

Here the class attribute is subject DM. In order to create a tree we had to make the decision that which attribute would at the root node of tree. For this we calculated Information Gain of each attribute one by one. To calculate gain of any attribute we first had to calculate entropy of data. In our case we had three possible values of class attribute (H, M, and L). In our data set of 350 records there are 72 H, 217 M and 61 L.

$$E(D) = -p_H \log_2(p_H) - p_M \log_2(p_M) - p_L \log_2(p_L)$$

Then the information gain of each attribute was calculated. The information of attribute state is as follow:

$$G(State) = E(D) - \frac{|D_{Punjab}|}{|D|} E(D_{Punjab}) - \frac{|D_{Sindh}|}{|D|} E(D_{Sindh}) - \frac{|D_{KPK}|}{|D|} E(D_{KPK}) - \frac{|D_{Bolochistan}|}{|D|} E(D_{Bolochistan})$$

Similarly the gain of all other attributes is calculated. In our case study the information gain of State is highest and therefore it selected as root node of the tree. This process was continued till all the attributes were classified perfectly in the tree and at the end of this step we will get a data model in the form of decision tree. This gained model is then stored in the database and the model of next subjects is created and one by one we got models of all subjects in our database. As screen shot of the generated tree is shown in figure 2.

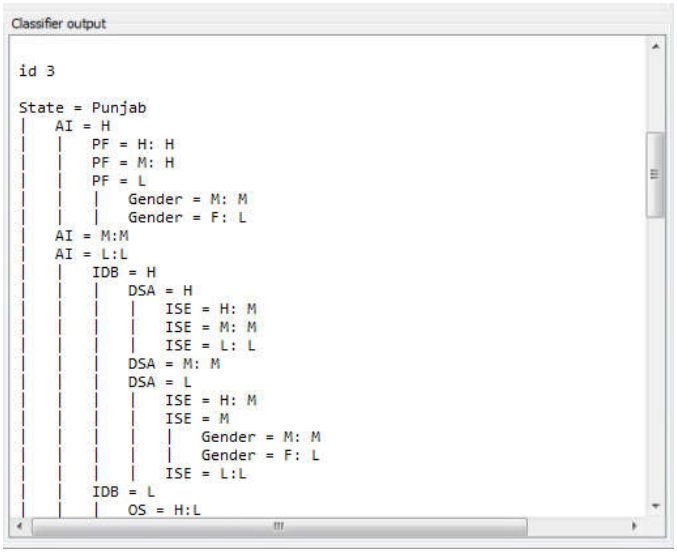

Fig. 2. Decision Tree for Data Mining

V. Experiments and Resutls

In this section, the models generated in previous section are analyzed and discussed. The models are validated through test data to find out whether the same patterns exists or not. The available dataset was divided into training data and test data. Training data was used to train the system quantitatively whereas the test data was used to test the system performance. In available data set of 423 students, we selected 350 for training and reaming 73 for testing purpose. In the 73 samples of test data there are 9 records having class value H, 47 rows have L and 17 students have Data Mining result L. The selected features of chosen student records were fed one by one in the system and obtained results were tallied with the actual results of the students.

The performance of the system is shown in table 3. The fourth column in the table shows the accuracy of each class attribute founded through the model we have generated. The accuracy is calculated by taking the ratio of true results identified by the proposed system to the total number of test record fed into the system. Last row of the table shows the overall performance of the system which is 85%, a reasonable result. The results obtained from the system are very encouraging and it can improve the performances of students as well as the organization.

TABLE III. Performance Meaurement of System

Class Value	Total Data Set	True Results	Accuracy
High	9	7	78%
Midiam	47	41	88%
Low	17	14	83%
Total	**73**	**62**	**85%**

VI. Conclusion and Future Work

Currently there is no automated system available, in organization under study, for educational data mining that can provide recommendation on subject selection to students based on their previous performance.

In this paper, data mining techniques are used for educational data mining. We used ID3 algorithm to generate decision trees. The number of experiments are performed to evaluate the performance of the system and the results obtained are very promising. The proposed system helps the students to improve their grades and also helps the institutes to improve their student's success ratio. As a future work, we plan to implement other data mining techniques to improve the performance of current system. In this work, for analysis and result calculation, dropped out students are not considered. In future work, we will attempt to include more features for better results and data of dropped out students.

As other direction of work, we will focus on applying EDM techniques in lower level academic institutes like schools and colleges.

References

[1] ERDOĞAN, Şenol Zafer, and Mehpare TİMOR. "A data mining application in a student database." Journal of aeronautics and space technologies 2.2, pp. 53-57, 2005.

[2] Luan, Jing. "Data mining and its applications in higher education" new directions for institutional research, vol. 2002 no. 113, pp. 17-36, 2002.

[3] Ivančević, Vladimir, et al. "An Application of Educational Data Mining Techniques at Faculty of Technical Sciences in Novi Sad" Proc. of the The 5th International Conference on Information Technology: ICIT 2011.

[4] Zhimei, Zhu. "Application of Data Mining Technology in the Information Technology of College English Teaching", Advance Journal of Food Science and Technology, vol. 5 no.7, pp. 969-975, 2013.

[5] Beikzadeh, Mohammad Reza, Somnuk Phon-Amnuaisuk, and Naeimeh Delavari. "Data mining application in higher learning institutions." Informatics in Education-An International Journal, vol. 7 no.1, pp 31-54, 2008.

[6] Agathe Merceron, Kalina Yacef, "Educational Data Mining: a Case Study", Proceedings of the 2005 conference on Artificial Intelligence in Education: Supporting Learning through Intelligent and Socially Informed Technology, pp.467-474, May 06, 2005.

[7] Baradwaj Brijesh Kumar and Saurabh Pal, "Mining educational data to analyze students' performance", International Journal of Advanced Computer Science and Applications (IJACSA), vol. 2 no. 6, pp. 63-69, 2011.

[8] Yiming Ma , Bing Liu , Ching Kian Wong , Philip S. Yu , Shuik Ming Lee, "Targeting the right students using data mining", Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining, p.457-464, August 20-23, 2000, Boston, Massachusetts, USA.

[9] Bhullar, Manpreet Singh, and Amritpal Kaur. "Use of Data Mining in Education Sector" Proceedings of the World Congress on Engineering and Computer Science, vol. 1, October 24-26, 2012, San Francisco, USA.

[10] Romero C., Ventura S. and Garcia E. "Data mining in course management systems: Moodle case study and tutorial", Computers & Education, vol. 51, no. 1, pp. 368-384, 2008.

[11] M.-S. Chen, J. Han, and P.S. Yu, "Data Mining and: an overview from a database perspective", IEEE Transactions on Knowledge and Data Engineering, vol. 8 no. 6, pp. 866-883, 1996.

[12] http://www.bzu.edu.pk

[13] Mohammed M. Abu Tair, Alaa M. El-Halees, "Mining Educational Data to Improve Students' Performance: A Case Study", International Journal of Information and Communication Technology Research, vol. 2 no. 2, pp. 140-146, February 2012

[14] J.R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, no. 1, pp. 81–106, 1986

[15] Nikhil Rajadhyax, Rudresh Shirwaikar, Data Mining on Educational Domain, 2012.

[16] https://www.coursera.org/course/bigdata-edu

[17] https://sites.google.com/a/iis.memphis.edu/edm-2013-conference/

[18] http://www.educationaldatamining.org/

[19] http://www.educationaldatamining.org/JEDM/index.php/JEDM

[20] Tooley, James. "The global education industry." IEA Hobart Paper 141 (2005).