



Applying Data Mining in Higher Education Sector

¹Archana Sharma, ²Prof. Vibhakar Mansotra, ³Rhea Mahajan

¹Department of Computer Application, Jodhpur National University, Jodhpur, Rajasthan, India.

²Department of Computer Science & IT, University of Jammu, Jammu, India.

³Department of Computer Engineering, Yogananda College Of Engineering and Technology, India.

Abstract: The new interesting subject that offered by institution to interact more student is "DATA MINING". In this paper we will discuss about the problem that are faced by students how to choose the best institutes for learning. One of the biggest challenges that student's faced tough time selecting the right engineering college that opens doors to exciting careers. Students would like to know, which college provides better quality education and its alumni are successful in the real world. Data Mining helps to students to take decision more accurately. Data mining is better tool to predict the general information of the college, courses offered and no. of seats, selection criteria, infrastructure, faculty performance, industry interface, placement, and potential to network, exchange programs and global exposure and national and international alumni chapter. In this paper we will discuss about data mining, their different phase's, advantages and also we classify data using weak data mining tool which helps to understand the data. In this paper we will use decision tree algorithm to predict the status of colleges, faculty performance, student feedback, student performance, infrastructure, placements and emotion states of students.

Key terms: Induction Algorithm, Knowledge Discovery, Information Gain

I. INTRODUCTION

Data Mining, also popularly known as Knowledge Discovery in Databases (KDD), refers to the nontrivial extraction of implicit, previously unknown and potentially useful information from data in databases. While data mining and knowledge discovery in databases (or KDD) are frequently treated as synonyms, data mining is actually part of the knowledge discovery process.[7] The following figure (Figure 1) shows data mining as a step in an iterative knowledge discovery process.

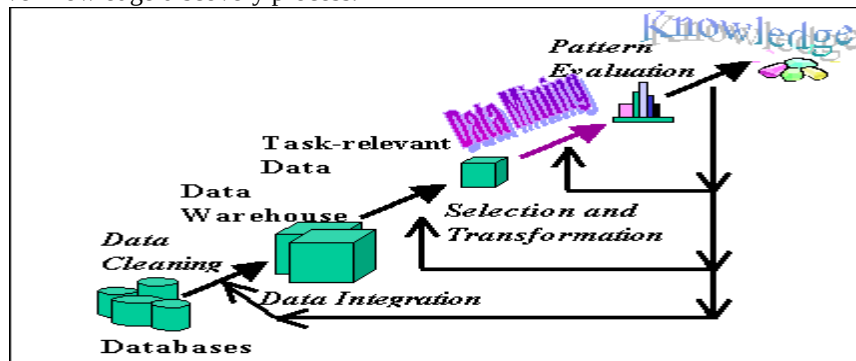


Fig. 1: Data mining Process

The Knowledge Discovery in Databases process comprises of a few steps leading from raw data collections to some form of new knowledge. The iterative process consists of the following steps:

1. Data cleaning:
Also known as data cleansing, it is a phase in which noise data and irrelevant data are removed from the collection.
2. Data integration:
At this stage, multiple data sources, often heterogeneous, may be combined in a common source.
3. Data selection:
At this step, the data relevant to the analysis is decided on and retrieved from the data collection.
4. Data transformation:
Also known as data consolidation, it is a phase in which the selected data is transformed into forms appropriate for the mining procedure.
5. Data mining:
It is the crucial step in which clever techniques are applied to extract patterns potentially useful.
6. Pattern evaluation:
In this step, strictly interesting patterns representing knowledge are identified based on given measures.

7. Knowledge representation:

Is the final phase in which the discovered knowledge is visually represented to the user. This essential step uses visualization techniques to help users understand and interpret the data mining results.

It is common to combine some of these steps together. For instance, data cleaning and data integration can be performed together as a pre-processing phase to generate a data warehouse. Data selection and data transformation can also be combined where the consolidation of the data is the result of the selection, or, as for the case of data warehouses, the selection is done on transformed data.

The KDD is an iterative process. Once the discovered knowledge is presented to the user, the evaluation measures can be enhanced, the mining can be further refined, new data can be selected or further transformed, or new data sources can be integrated, in order to get different, more appropriate results.

Data mining tools and algorithms

Machine Learning

Computer science, heuristics and Induction algorithms

Artificial Intelligence

Emulating human intelligence

Neural Networks

A. Classification by Decision Tree Induction

The basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner. The algorithm is a version of ID3, a well-known decision tree induction algorithm. [6] The basic strategy is as follows:

1. The tree starts as a single node representing the training samples.
2. If the samples are all of the same class, then the node becomes a leaf and is labeled with that class.
3. Otherwise, the algorithm uses an entropy based measure known as information gain as a heuristic for selecting the attribute that will best separate the samples into individual classes. This attribute becomes the “test” or “decision” attributes at the node. In this version of the algorithm, all attributes are categorical, that is, discrete-valued. Continuous-valued attribute must be discretized.
4. A branch is created for each known value of the test attribute, and the samples are portioned accordingly.
5. The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once an attribute has occurred at a node, it need not be considered in any of the nodes descendants.
6. The recursive partitioning stops only when any one of the following conditions is true:
 - a) All samples for a given node belong to the same class.
 - b) There are no remaining attribute on which the samples may be further partitioned. In this case, majority voting is employed. This involves converting the given node into a leaf and labeling it with the class in majority among samples. Alternatively, the class distribution of the node samples may be stored.
 - c) There are no samples for the branch test-attribute= a_i . In this case, a leaf is created with the majority class in samples.

Attribute selection measure is computed by Information gain Used by the ID3, C4.5 and C5.0 tree-generation algorithms. Information gain is based on the concept of entropy from information theory.

$$I_E(f) = - \sum_{i=1}^m f_i \log_2 f_i$$

B. Tools of Data Collection & Analysis

Various tools are needed for that project some for analyzing data, some for designing, implementation and some developing software tool these are:

- MYSQL DATABASE
- EXCEL
- MS ACCESS
- SPSS
- METLAB TOOL
- WEKA DATA MINING TOOL
- TANGARA DATA MINING TOOL
- WEB MINER
- V.B 6.0

II. DATA MINING EXPERIMENT

In this research work I will collect data of three thousand students from three colleges but in this example I have chosen fourteen students from three colleges. In the first step we clean and integrate data. For our problem we

chose five attributes these converted into its equivalent values which are given below in the table.

Table I: Selected Attributes

S.No	Selected attributes	Description
1	College_name	X,Y,Z
2	Faculty_performance	Excellent,good,average
3	Campus_placement	Yes,no
4	Student_performance	Pass,reappear
5	Class:take_admission	Yes,no

	A	B	C	D	E	F	G	H
	COLLEGE_NAME	FACULTY_PERFOR	CAMPUS_PLACEM	STUDENT_PERFOR	CLASS:TAKE_ADMISSION			
1								
2	X	Excellent	No	Reappear	No			
3	X	Excellent	No	Pass	No			
4	Y	Excellent	No	Reappear	Yes			
5	Z	Good	No	Reappear	Yes			
6	Z	Average	Yes	Reappear	Yes			
7	Z	Average	Yes	Pass	No			
8	Y	Average	Yes	Pass	Yes			
9	X	Good	No	Reappear	No			
10	X	Average	Yes	Reappear	Yes			
11	Z	Good	Yes	Reappear	Yes			
12	X	Good	Yes	Pass	Yes			
13	Y	Good	No	Pass	Yes			
14	Y	Excellent	Yes	Reappear	Yes			
15	Z	Good	No	Pass	No			
16								
17								
18								
19								
20								
21								
22								
23								
24								

Fig. 2: csv file of database

After collecting and cleaning the data we classify data using weka mining tool. For classifying and prediction we use ID3 algorithm. in this experiment table 2 presents a training set of data tuples taken from the college1 database. The class label attribute, take_admission, has two distinct values (namely yes or no).let class C1 correspond to yes and class c2 correspond to no. There are 9 samples of class yes and 5 samples of class no.

Now we compute the information gain of each attribute, we first use equation to compute the expected information needed to classify a given sample:

$$I(S_1, S_2) = I(9, 5) = -9/14 \log_2 9/14 - 5/14 \log_2 5/14 = 0.940.$$

Next we compute the entropy of each attribute.

For college_name="X"

$$S_{11}=2 \quad S_{21}=3 \quad I(S_{11}, S_{21})=0.971$$

For college_name="Y"

$$S_{12}=4 \quad S_{22}=0 \quad I(S_{12}, S_{22})=0$$

For college_name="Z"

$$S_{13}=3 \quad S_{23}=2 \quad I(S_{13}, S_{23})=0.971$$

Using equation 2, the expected information needed to classifying a given sample, if the samples are partitioned according to college_name is

$$E(\text{college_name}) = 5/14 I(S_{11}, S_{21}) + 4/14 I(S_{12}, S_{22}) + 5/14 I(S_{13}, S_{23}) = 0.694.$$

Hence, the gain in information from such a partitioning would be

$$\text{Gain}(\text{college_name}) = I(S_1, S_2) - E(\text{college_name}) = 0.246.$$

Similarly we can compute $\text{Gain}(\text{campus_placement})=0.151$, $\text{Gain}(\text{faculty_performance})=0.029$, and $\text{Gain}(\text{student_performance})=0.048$. since college_name has the highest information gain among the attributes, it is selected as the test attribute.

The samples are then partitioned as shown in figure below:

College_Name			
X			
Faculty_performance	Campus_placement	Student_performance	Class
Excellent	No	Reappear	No
Excellent	No	Pass	No
Good	No	Reappear	Yes
Average	Yes	Reappear	Yes
Good	Yes	Pass	Yes
Y			
Faculty_performance	Campus_placement	Student_performance	Class
Good	No	Reappear	Yes
Average	Yes	Reappear	Yes
Average	Yes	Pass	No
Good	Yes	Reappear	Yes
Good	No	Pass	No
Z			
Faculty_performance	Campus_placement	Student_performance	Class
Excellent	No	Reappear	Yes
Average	Yes	Pass	Yes
Good	No	Pass	Yes
Excellent	Yes	Reappear	Yes

Fig. 3: Decision table

The samples falling into the partition for college_name="Y" all belong to the same class. Since they are all belong to class yes, a leaf should therefore be created at the end of this branch and labeled with yes.

The final decision tree returned by the algorithm is shown below

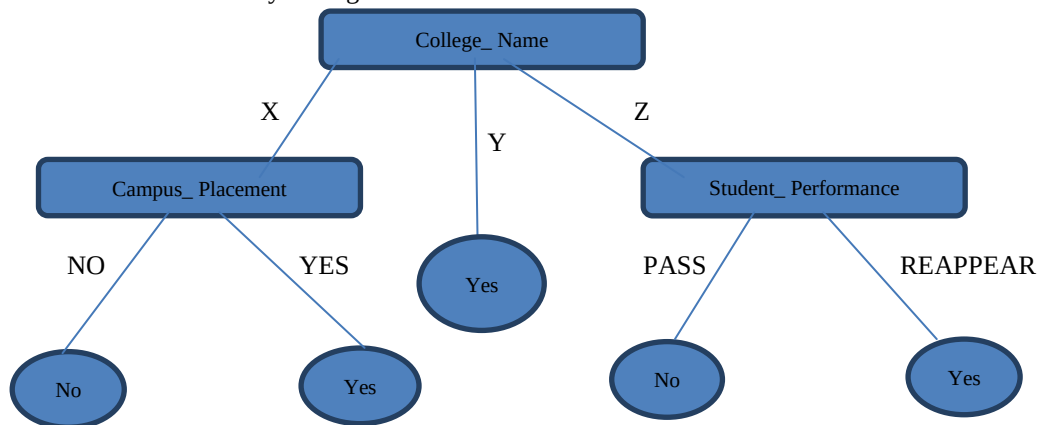


Fig. 4: Decision tree

The classification learning was also used to predict the student's failure/success to pass the academic exam based on their present behavioral profile. For the ID3 classification learning based on training set, there was an 85.71% success rate (the correctly classified instances) which is higher value of prediction. We have taken a sample of 14 instances from which 12 are correctly classified and 2 are not correctly identified. From the decision tree we are easily identify the weak institute and whose chances of fail are maximum. After identifying the weak institute we can work hard on that institute to minimize the failure result and we can improve overall result and performance of the institute.

Advantage of Data mining in Academics

Data mining gives the answer of questions like:

- Q1. Which college provides quality education?
- Q2. How qualified is the faculty, i.e. Ph.D., MTECH, B.TECH?
- Q3. How active is the faculty in research?
- Q4. Are there many visitors giving seminars?
- Q5. Do they arrange workshops and conferences regularly?
- Q6. The number of faculty members?
- Q7. How good are the labs in the discipline of your choice?
- Q8. How many books are there per students in the library?
- Q9. What e-journals do they subscribe?
- Q10. How much bandwidth on a per capita basis do they have?
- Q11. Are lecture hall equipped with Pc's, LCD projector screen and audio facilities?

Q12. What sports facilities exist for the student boys?

Q13. Which college provides best branch?

Statically result given by ID3 algorithm

```
Time taken to build model: 0 seconds

=== Stratified cross-validation ===
=== Summary ===

Correctly Classified Instances      12          85.7143 %
Incorrectly Classified Instances     2          14.2857 %
Kappa statistic                     0.6889
Mean absolute error                 0.1429
Root mean squared error             0.378
Relative absolute error              30 %
Root relative squared error         76.6097 %
Total Number of Instances           14

=== Detailed Accuracy By Class ===

      TP Rate  FP Rate  Precision  Recall  F-Measure  ROC Area  Class
      0.8      0.111    0.8        0.8     0.8        0.844    No
      0.889    0.2      0.889    0.889   0.889      0.844    Yes
Weighted Avg.  0.857    0.168    0.857    0.857   0.857      0.844

=== Confusion Matrix ===

a b  <-- classified as
4 1 | a = No
1 8 | b = Yes
```

III. CONCLUSION

The analysis about the Institute success rate varies for students in choosing the right college. The domain of the current patterns in colleges about the success rate identifies in this paper we will concluded about the problem faced by students how to choose the best institutes for learning. Whether provides better quality education and its alumni are successful in the real world. Since the model shows the weak institute, it also helps the students to identify the best college to build their future. In end it helps students to take decision more accurately about the college and courses. Proposed system also shows the data graphically according to the need of the students in particular fields. For future work we also use clustering, with the help of clustering we can see the domain and emotion states of students.

REFERENCES

- [1] Hideko Kitahama, "Data Mining through Cluster Analysis Evaluation on Internationalization of Universities in Japan".
- [2] Bruce L. Golden R. H. Smith School of Business University of Maryland College Park, MD 20742 "An Example of Visualization in Data Mining"
- [3] Jing Luan, PhD Chief Planning and Research Officer, Cabrillo College Founder, Knowledge Discovery Laboratories "Data Mining Applications in Higher Education".
- [4] Thulasi Kumarthulasi.kumar@uni.edu, University of Northern Iowa " Theoretical Basis for Data Mining Approach to Higher Education Research".
- [5] N.V.Anand Kumar Research Scholar, Department of Computer Science and engineering Anna university, Chennai G.V.Uma Assistant professor, Department of Computer Science and Engineering Anna university, Chennai "Improving Academic Performance of Students by Applying Data Mining Technique".
- [6] Han, J. and Kamber, M., (2006), "Data Mining: Concepts and Techniques", 2nd edition. The Morgan Kaufmann Series in Data Management Systems", Jim Gray, Series Editor, 2006.
- [7] C. Romero, S. Ventura, E. Garcia (2008), "Data mining in course management Systems: Moodle case study and tutorial", Computers & Education, Vol. 51, No. 1, pp. 368- 384, 2008.