ISSN:0975-9646

# Data Evaluation and Knowledge Extraction for Students Using Clustering and Classification

Tirtha Chavan[#1], Amruta Barretto[*2], Tanmay Chavan[*3]

[#]Department of Computer Engineering, Mumbai University, India

[*]Department of Information Technology, Mumbai University, India

*Abstract* — **This paper describes a methodology for classifying students planning for pursuing higher education in abroad countries on basis of their academic, GRE, TOEFL, IELTS, GMAT marks and also by grading their extracurricular activities. Starting from collection of databases, which was a result of survey, Data Mining (DM) techniques will be used to discover a set of students having similar academic profiles who got admits from the various universities, thus gaining information about the prospective universities that the students can apply for getting the admits. Past data collected of the students will be clustered using Two Step Clustering Algorithm to create clusters of universities according to each student's profile criteria. Then this knowledge will be used by the classification module to classify new students into the previously obtained classes.**

***Keywords*—Data Mining, Classification, Two Step Clustering Algorithm.**

## I. INTRODUCTION

In the current scenario the students who are looking for pursuing higher education abroad have to depend on study abroad consultancies, human decisions based on past experiences and web forums for taking suggestions regarding the university selection from current students and alumni .This is one of the major barrier for students who are planning to pursue higher education abroad.

Nowadays the large data available on internet is in disparate form, and the major barrier to obtain the required meaningful data is the growing size of database and the versatility of the domains. To overcome these barriers and to find the useful patterns of the data various clustering algorithms and data mining techniques are used.

Clustering is the task of assigning a set of objects into the groups called clusters so that the objects in the same cluster are identical to each other as compared to those in other clusters. Data mining is the process that extracts information from data set and discovers patterns in large data sets to transform it into understandable user structure for further use.

In this paper, a two-step method is applied for clustering dataset. In the first step, HAC (hierarchical agglomerative clustering) [1] algorithm is adopted to cluster the original dataset into some subsets. The formed subsets in this step along with adding additional features will be chosen to be the objects as an input to k-means in next step. Since every subset may contain several data points, applying chosen subsets as initial set of clusters in k-means clustering algorithm will be a better solution than selecting individual data. Another benefit is to reduce the influences of outlier, as the outlier will be smoothed by these features. The results show that this proposed method is a feasible solution for clustering our dataset which will be of great help to classify students planning for pursuing higher education in abroad on basis of their academic, GRE, TOEFL, IELTS, GMAT marks and also by grading their extracurricular activities.

## II. PROPOSED SYSTEM

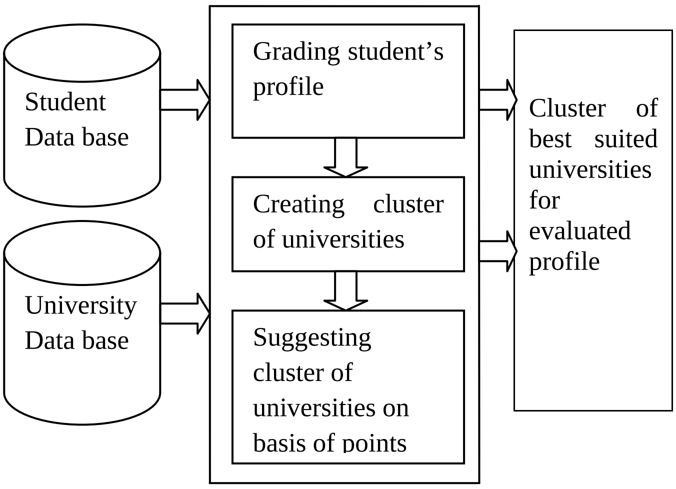
Student Data base
University Data base
Grading student's profile
Creating cluster of universities
Suggesting cluster of universities on basis of points
Cluster of best suited universities for evaluated profile

Fig. 1 Frame work showing an overview of the system

The first step depicts the collection of data from various sources including universities websites wherein they mention the minimum requirement of the GRE, TOEFL, IELTS, GMAT as well as GPA required in the academics [2]. The other sources include various consultancies and web portals. The data collected of past 5 years scan include both numeric as well as categorical attributes.

The data pre processing stage will detect and correct bad data. In this cleaning phase missing values found will be cleaned. With this procedure the major problems of real databases will be minimized.

Now the numeric attribute values will be normalized into the range of zero and one. In the Data Mining stage DM techniques will be applied on the formatted data. One or more DM techniques can be used to retrieve knowledge about the student's intake in the various universities. The details will be described in the next section.

Finally the knowledge discovered from the data mining models will be used by the students to know about which universities to apply in order to get admits from them.

## III. DATA MINING PROCESS

Initially a data set is collected which includes both the categorical and the numeric data so we assign numeric values to the categorical data. DM process consists of data selection and transformation and clustering algorithms that are used to discover similar patterns among data.

### *A. Data Selection and Transformation*

In this step only those fields will be selected which will be required for data mining. A few derived variables will be selected.

TABLE I
STUDENT PROFILE EVALUATION VARIABLES

| Variable | Description | Possible values |
|---|---|---|
| GRE | Marks obtained in the GRE exam | {>=260 & <=340} |
| TOEFL | Marks obtained in TOEFL exam for each sections | {>=0 & <=120} |
| GMAT | Marks obtained in GMAT exam | {>=0 & <=750} |
| IELTS | Marks Obtained in IELTS exam for each section | {1-9} |
| ACADS | Grade Point Aggregate | {>=0 & <=4} |
| RND | Research and development work. If the student has done research and development work then the student will get 5 profile points. | 5 |
| TE | Participation in various technical events and also published research papers. If student has participated then the student will get 4 profile points. | 4 |
| NGO'S | Participation in various social drives and working for various NGO's. If student has participated then the student will get 3 profile points | 3 |
| SPORTS | Participation in various sports. If student has participated then the student will get 2 profile points. | 2 |
| CC | Participation in college committee. If student has participated then the student will get 1 profile point. | 1 |
| OTHER | This includes the other achievements and other parameters. For other parameters the student will get 1 profile point. | 1 |

### *B. Evaluating the students profile*

This module will give the details of evaluating the student's profile, considering all the parameters. (Here it is considered student's profile applying for master's program in United States of America)

Calculation of the student's profile points

Student's profile competency rate
$$= \text{GRE/GMAT/IELTS SCORE} + \text{TOEFL SCORE} + 10*(\text{RND}+\text{TE}+\text{NGO'S}+\text{SPORTS}+\text{CC}+\text{OTHER}) + 15*(\text{GPA})$$

This gives us the basis for evaluating the student's competency.

Maximum possible values that can be considered for evaluating student's profile are as below.
GRE= 340
TOEFL=120
ACADS=60 (15 X GPA where 15 is multiplication factor)
Profile Points=160 (10 X summation of Evaluating Variable)

Let us consider 3 profiles of Student A, Student B, Student C after evaluation based on specific variables

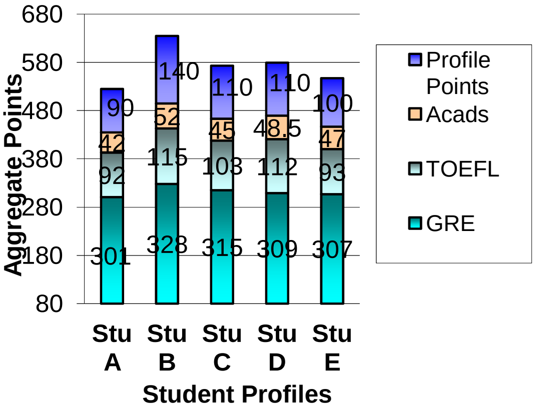

Fig. 2 A sample graph showing the competency of three student's profiles

### *C. Clustering*

In this stage, clustering will be done on collected data of students to form clusters of universities according to profile criteria. The choice and selection of clustering algorithm is decisive in this stage [3].

Two Step Cluster Method will be applied to perform clustering. First the HAC (Hierarchical Agglomerative Clustering) will be applied which takes the numeric data as the input and generates the hierarchical partitions as the output. Thus HAC algorithm is applied in the first clustering step to group the data into subsets.

```
INPUT     : D: the training data set;
OUTPUT  : A: hierarchy of clusters;
PSEUDOCODE:
Let each object be a cluster;
Compute proximity matrix;
do {
    Merge two "closest" clusters based on a certain criterion;
    Update proximity matrix;
}
while (until only one cluster remains);
```

In the second step K-Means algorithm which takes numeric data as the input and generates crispy partitions (i.e. every object only belongs to one cluster) as the output will be applied

Clusters are segregated on criteria of Major courses and Score Range which makes Clustering algorithm more efficient.

TABLE II
CLUSTERS FORMED USING THE CLUSTERING ALGORITHMS

| Cluster ID | Score Range | Major |
|---|---|---|
| . | . | . |
| 3 | Class(631- 640) | CS |
| . | . | . |
| . | . | . |
| 13 | Class(571-580) | CS |
| 14 | Class(561-570) | CS |
| . | . | . |
| . | . | . |
| 31 | Class(571-580) | MIS |
| 32 | Class(561-570) | MIS |
| . | . | . |
| . | . | . |
| 59 | Class(571-580) | EE |
| 60 | Class(571-580) | EE |
| . | . | . |
| . | . | . |
| 71 | Class(521-530) | MPH |
| . | . | . |
| . | . | . |

Universities distribution using the clustering algorithm is given in the table below [4] [5] [6].

TABLE III
UNIVERSITIES DISTRIBUTION USING THE CLUSTERING ALGORITHMS

| Cluster ID | University Cluster |
|---|---|
| 1 | Carnegie Mellon University |
| | Massachusetts Institute of Technology |
| | Stanford University |
| | University of California-Berkeley |
| | Cornell University |
| . | . |
| 12 | Stony Brook University-SUNY |
| | Virginia Tech |
| | North Carolina State University |
| | Rensselaer Polytechnic Institute |
| | Texas A&M University-College Station |
| . | . |

### D. Classification

Classification model as shown in Fig. 2 is used to classify new students to previously obtained clusters. This section explains one of the algorithms used to create Univariate DT's which can test one attribute per test node. It is based on the ID3 algorithm that tries to find small DT's.

Some premises on which this algorithm is based are presented, and after that the inference of the weights and tests on the nodes of the trees is discussed. ID3 stands for Iterative Dichotomiser 3. The classification model Fig 2. uses supervised learning, based on the knowledge about relation between characteristics of the consumer and its corresponding class, obtained with the clustering operation.

```
INPUT:      k: the number of clusters;
            D: the training data set;
OUTPUT: A set of k clusters;
PSEUDOCODE:
Select first k objects as the initial cluster centroids;
do {
    Assign all objects in D to the nearest centroids;
    Update centroid for each cluster, i.e., compute the  mean value of objects for each cluster;
}
while(until no change to all centroids or Maximum Iteration has been reached or terminate condition matched);
```

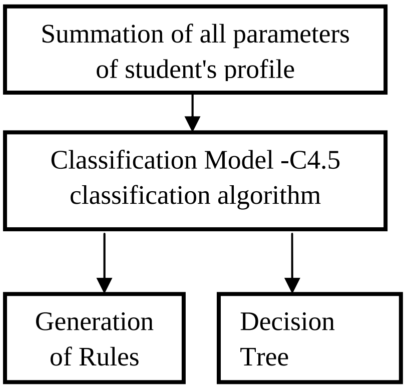
Summation of all parameters of student's profile
Classification Model -C4.5 classification algorithm
Generation of Rules
Decision Tree

Fig. 3 Classification model

For classification C4.5 classification algorithm which is a decision tree of a rule based modeling technique was used.

The pseudo code for C4.5 classification algorithm is as follows.

1. Check for base cases.
2. For each attribute a
    1. Find the normalized information gain from splitting on a
3. Let a_best be the attribute with the highest normalized information gain.

4. Create a decision node that splits on a_best.
5. Recurs on the sub lists obtained by splitting on a_best, and add those nodes as children of node.

### 1) Construction

The algorithm is constructed as follows

1. If all cases are of the same class, the tree is a leaf, the tree is a leaf and so the leaf is returned with this class
2. For each attribute calculate the potential information provided by a test on the attribute. Also calculate the gain in information that would result from a test on the attribute
3. Depending on the current selection criterion, find the best attribute to branch on

### 2) Counting gain

This process uses the “Entropy”, i.e. a measure of the disorder of the data. The Entropy of $\bar{y}$ is calculated by

$$Entropy\ (\bar{y}) = -\sum_{j=1}^{n} \frac{|y_j|}{|\bar{y}|} \log \frac{|y_j|}{|\bar{y}|}$$

(1)

Iterating over all possible values of $\bar{y}$. The conditional Entropy is

$$Entropy\,(j|\bar{y}) = \frac{|y_j|}{|\bar{y}|} \log \frac{|y_j|}{|\bar{y}|}$$

(2)

And finally, we define Gain by

$$Gain\,(\bar{y}, j) = Entropy(\bar{y} - Entropy(j|\bar{y}))$$

(3)

The aim is to maximize the Gain, dividing by over all entropy due to split argument $\bar{y}$ by value j.

### 3) Pruning

This is an important step to the result because of the outliers. All data sets contain a little subset of instances that are not well-defined, and differ from the others in the neighborhood. After the complete creation of the tree, that must classify all the instances in the training set, it is pruned. This is done to reduce classification errors and make the tree more general.

Statistical Representation for data assumed in fig.4

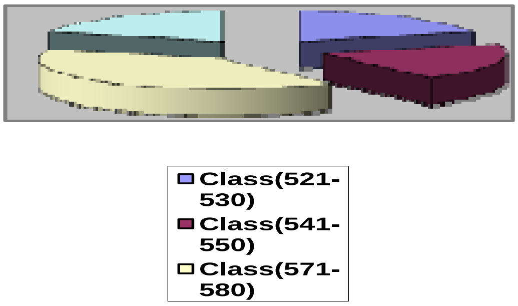

Fig. 4 Classification of students based on their profile into the predefined clusters

## IV. MERITS

The proposed system has many benefits. By using this system, prospective students do not have to only depend on human decisions, study abroad consultancies and inputs from current students and alumni for information regarding the entire admission process. This system will not only help students to take the decision of which all universities they can apply but it will also give students access to the first hand information about the profiles of students who got the admit for the same course and university.

## V. LIMITATIONS

The statement of purpose and letter of recommendation are also considered by the universities for student profile evaluation. These are the documents which are required to be evaluated by humans so system functionality becomes partially human functioned instead of completely automated. Humans can’t maintain the accuracy while grading these 300-400 words documents which might affect systems overall functionality. But in the entire evaluation of the students profile the statement of purpose and the student’s evaluation only affects 2 % of actual evaluation. Algorithms used to develop system do not evaluate the text data submitted by student and recommender.

## VI. CONCLUSION

This system will ease the student’s efforts as after inputting the profile details like the academics, extracurricular activities information, GRE, TOEFL, GMAT, IELTS score, relevant work experience information along with course for which one is applying as system will search in the database where profiles of students on the basis of the course and university student has admit from is stored.

This system can be used by prospective students along with the study abroad consultancies. This system will give accurate results on the basis of the information provided by student as factors such as minimum scores required in each section of GRE, TOEFL and minimum grade point average (GPA) required for a particular university and its course will also be taken into consideration. [2]

## REFERENCES

[1] Ming-Yi Shih, Jar-Wen Jheng and Lien-Fu Lai, “A Two-Step Method for Clustering Mixed Categorical and Numeric Data” ,Tamkang Journal of Science and Engineering, Vol. 13, No. 1, pp. 1119 (2010)

[2] http://www.heinz.cmu.edu/admissions/application process/english-language-proficiency/index.aspx

[3] Jain, A. K. and Dubes, R. C., Algorithms for Clustering Data, New Jersey: Printice Hall (1988).

[4] http://www.msinus.com/content/usa-university-rankings-computer science-527/

[5] http://grad-schools.usnews.rankingsandreviews.com/best graduate-schools/top-science-schools/computer-science-rankings

[6] http://www.topuniversities.com/university-rankings/world university- rankings/2012/subject-rankings/technology