Clustering Academies: An Integrated Approach using Genetic Algorithm and Data Mining

Ashok M.V.
Associate Professor,
Department of MCA,
GIMS, RR Nagar,
Bangalore, Karnataka, India

Apoorva A.
Assistant professor
Department of MCA, GIMS,
RR Nagar, Bangalore,
Karnataka, India

G. Suganthi, PhD
Associate Professor,
Dept. of Computer Science,
Women's Christian College,
Nagercoil, Tamil Nadu, India

ABSTRACT
Educational Data Mining deals with developing methods to explore unique types of data in educational settings by applying a combination of approaches such as data mining, statistical and machine learning to get viable information. The objective of this paper is to help prospective students in selecting good academy during their enrollment to degree courses. In this paper, an integrated approach consisting of evolutionary approach i.e. genetic algorithm for preprocessing the data of 75 academies of Bangalore and data mining approach i.e. k-means for clustering the academies is used. Thus the cluster obtained as result will consist of academies that will be ranked as Excellent [E], Good [G], Average [A] and [Poor] according to the considered attributes. This work will help the prospective students in selecting the best academy during admission.

Keywords
Educational Data mining, evolutionary approach, Genetic algorithm, K-means algorithm, machine learning

1. INTRODUCTION
Every year new academies keep adding to the existing list of academies making students life difficult, as they will be in dilemma while selecting the academies during admission. More options in selection lead to confusion due to lack of information about the upcoming and existing academies. Success relies on choosing the right academy in student's career. The objective of this study is to help such students make right decision during admission. Bangalore being hub of educational institutions, educational data from 75 academies, in and around Bangalore are selected and collected. Huge data collected should be processed and patterns need to be compared manually, which is tedious and cumbersome. Hence data mining technique is used to mine the large data. A combination of evolutionary and data mining approaches is used, where in the former is used for preprocessing and later for clustering using association rule.

1.1 Problem Statement
Every student dreams to be successful in life. Hence choosing a right academy to pursue his studies is a key factor. A unique approach consisting of combination of both evolutionary and data mining is applied on data collected, which is processed and mined to get useful information that in turn help students select a right academy.

2. RELATED WORKS
Few of the related works are listed below:

2.1 Admission, genetic algorithm and k-means algorithm
Data mining concepts are applied to improve the quality of education and important component of quality is admission. According to Arora et al., 2013[1]; admission could be improved by identifying those admission inquires which most likely to turn into actual admissions. Arora et al., 2012 [2]; Arora et al., 2013 [3]; M. Sukanya et al., 2012 [4]; JayanthiRanjan et al., 2007 [5];Behrouz Minaei-Bidgoli et al.,2003 [6]; Deborah A. Kashy et al.,2003[7]; are few papers regarding admission of students. Providing genetic based personalized curriculum sequencing, class time table **Sandeep singhrawat, lakshmirajamani, 2010**[10]; Mu-Jung Huang et al.,[8];were implemented using genetic algorithm. Other applications of genetic algorithm are solving Traveler Salesman Problem (TSP) using a combined Hopfield neural network with genetic algorithm **AlirezaArabasadi et al., 2011**[11]; **AzinKhosravi Khorashad1et al., 2012** [9];Krishna K, Murty M N [12]; Propose a novel hybrid genetic algorithm (GA) viz., genetic K- means algorithm that finds worldwide optimal partition of a given data into a specified number of cluster. It is also observed that GKA search quicker than some of the other evolutionary algorithms used for clustering. Zhexue Huang[13]; focuses on the practical issues of extending the k-means algorithm to cluster data with categorical value. Outstanding property of k-means algorithm in data mining is its efficiency in clustering large data sets. However, it only works on numeric data limits its use in many data mining applications because of the involvement of categorical data. Leon Bottou, YoshuaBengio[14]; Studies the convergence properties of the well-known K- means clustering algorithm.

3. PROPOSED MODEL
Algorithm of the proposed methodology with its computation process is explained below:

Data required for the study i.e., 75 institutions were collected from respective academy website for the year 2014. Data preprocessing is done in two steps

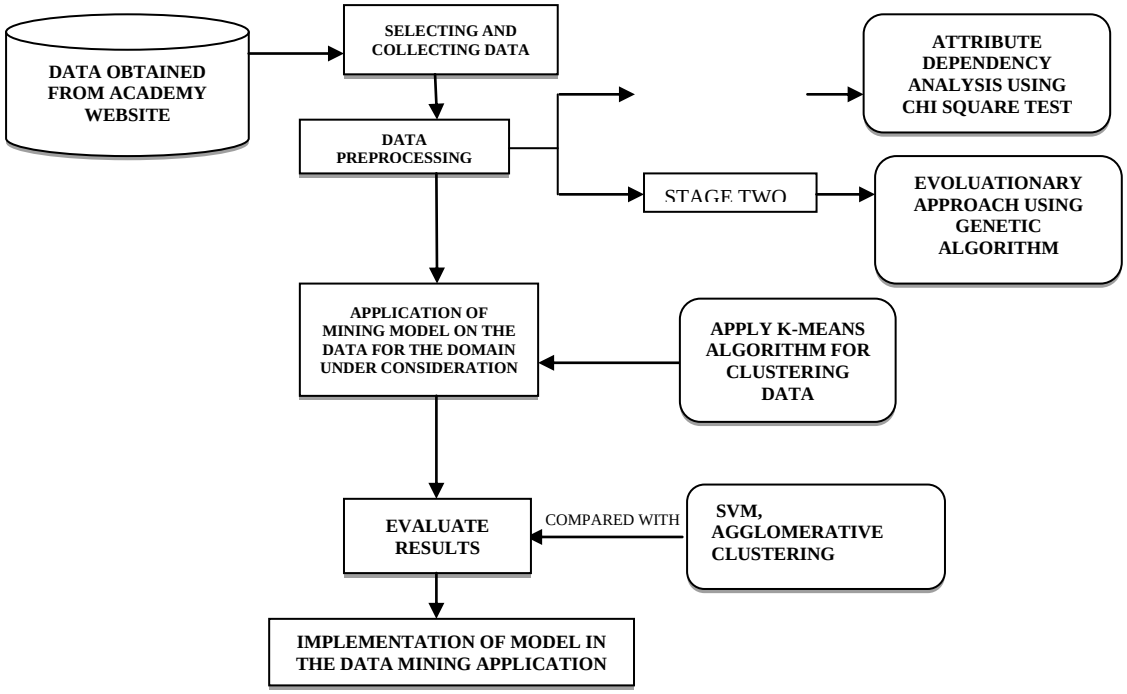

Fig.1 Proposed methodology for clustering institutions

Stage 1: **Apply chi- square test for the goodness of fit**
Attribute dependency analysis using statistical test is done wherein attributes that does not contribute to the study are eliminated using chi-square test for the goodness of fit. Variables like phone number, E-mail Id, Name and location are eliminated.

Stage 2: **Apply Genetic algorithm,** an evolutionary approach as it is robust and has ability to generate accurate result.

The resultant obtained after preprocessing were clustered using k-means algorithm, a data mining algorithm. Results obtained were compared with other clustering algorithms.

4. DATA DESCRIPTION

Data from 75 institutions have been collected and the data description is as follows

N(Name)- it represents the name of the academy and can take only text values.

E-ID- its email-id of the academy and can take alphanumeric values i.e., varchar.

PN- phone number of the academy and can take only digits from 0 to 9.

Location- address of the academy and will take textual values.

NE – number of events held in the academy. It is split in to 5 groups. If NE is between 10 and 15, allocation is 1. If NE is between 7 and 10, allocation is 2. If NE is between 5 and 7,

allocation is 3. If NE is between 3 and5, allocation is 4. If NE is between 0-3, allocation is 5. The possible values that it can take are {1, 2, 3, 4, and 5}.

RH – Number of rank holders. It is divided in to 5 groups. If a academy gets ranks between 5 and 6, allocation is 1. RH value is between 3 and 4, allocation is 2. RH value is between 2-1, allocation is 3. RH is 1, allocation 4 and RH is 0 then allocation is 5. The possible values that it can take are {1, 2, 3, 4, and 5}.

Pl(Placement) – number of students placed, expressed in terms of percentage. It can take values from 1% -100%.

R(Result) – total % of students passed. It can take values ranging from 1-100%.

A(Admission) – total number of students in academy. Possible values {1, 2, 3, 4, 5….}.Range is between 100% -

70%, allocation is 1. If the range is between 70% - 50%, allocation is 2. If the range is between 50% - 25%, allocation is 3. If the range is between 25% - 10%, allocation is 4 and if the range is between 10% - 0%, allocation is 5.

F(Facility) – facilities available in academy such as lab, library, campus and other infrastructure. Possible values it can take A-average, G-good, E-excellent, and P-poor.

P(Popularity) - represents popularity of the academy in society. Possible values it can take from 1% to 100%.

Table 1. Data Description And Input Table

Variables	Description	Possible Values
N	Name of the academy	{Text}
E-Id	Email-Id of the academy	{alphanumeric}
PN	Phone number of academy	{1, 2, 3, 4, 5…}
Location	Address of the academy	{Text}
NE	No of Events held	{1, 2, 3, 4, 5...}

	in academy	
RH	Number of rank holders	{1, 2, 3, 4, 5...}
Pl	Number of students placed	{ 1% - 100% }
R	Total percentage of passed students	{ 1% - 100% }
A	Number of students in the academy	{1, 2, 3, 4, 5...}
F	Laboratory, library, campus and other equipment	A-average, G-good, E-excellent, p-poor
P	Popularity of the academy in the society	{ 1% - 100% }

5. METHODOLOGY

5.1 Data Pre-processing:

Basically data preprocessing is done using two stages.

5.1.1 ***Stage1***:

Application of chi-square test for the goodness of fit.

5.1.2 ***Stage2***:

Application of Genetic algorithm (Evolutionary approach)

Table 2. Extract Of Input Table With Data

Id	N	NE	RH	Pl	R	A	F	P
1	DSBS	2	4	78%	68%	3	E(4)	90%
2	DBIT	3	3	77%	60%	2	E(3)	99%
3	PES	1	1	99%	98%	1	G(10)	83%
4	BHS	1	1	98%	99%	1	G(9)	67%
5	MC	2	3	80%	81%	2	G(8)	65%
6	AIMS	4	4	72%	75%	3	G(6)	65%
7	HCW	5	5	65%	60%	4	G(5)	65%
8	EWCS	4	3	73%	65%	3	G(2)	58%
9	AIT	3	2	76%	76%	2	G(1)	60%
10	BMS	3	2	75%	76%	2	A(7)	50%

Genetic Algorithm working:

Step 1: Initialize population

The data table 5.1.1 is the input for the genetic algorithm.

Step 2: Evaluate population

This step helps in finding out useful variables used as attributes for further processing. The resultant of this step is same as 5.1.1, eliminating useless variables.

The above table represents data extract of the 75 academies acting as input.

Step 3: perform crossover and mutation

```
While (Termination Criteria Not Satisfied)
        {
        //select attribute for reproductions//
                Current_attribute = next_attribute.
        While (! EOF)
         {
        If (compare each values with target)
                {
                New target value
                }
         }
       }
```

Table 3 .Resultant of step 3

Id	NE	RH	Pl	R	A	F	P
1	1(4)	1(4)	99%(3)	99%(4)	1(4)	E(4)	90%
2	1 (3)	1(3)	98%(4)	98%(3)	1(2)	E(3)	99%
3	2(5)	2(10)	80%(5)	81%(5)	2(10)	G(10)	83%
4	2(1)	2(9)	78%(1)	76%(10)	2(9)	G(9)	67%
5	3(10)	3(8)	77%(2)	76%(10)	2(5)	G(8)	65%
6	3 (9)	3(5)	76%(9)	75%(6)	2(2)	G(6)	65%
7	3(2)	3(2)	75%(10)	68%(1)	3(8)	G(5)	65%
8	4(8)	4(6)	73%(8)	65%(8)	3(6)	G(2)	58%
9	4(6)	4(1)	72%(6)	60%(7)	3(1)	G(1)	60%
10	5 (7)	5(7)	65%(7)	60%(2)	4(7)	A(7)	50%

Note: values within brackets are ids representing academies.

Processing of data starts with the first attribute ‘no of events (NE)’ in table 5.1.1. The first element will be compared with other values in the same attribute and the resultant will be displayed in the descending order. The process will be continued till last attribute i.e., research projects resulting in the above table.

Table 4 Output Table Extract Of Genetic Algorithm

Id	No of events	Rank Holders	Placement	Results	Admission	Facility	Popularity
4	1	1	98%	99%	1	E	90%
3	1	1	99%	98%	1	E	99%
5	2	3	80%	81%	2	G	83%
9	3	2	76%	76%	2	G	67%
10	3	2	75%	76%	2	G	65%
6	4	4	72%	75%	3	G	65%
2	3	3	77%	60%	2	G	65%
8	4	3	73%	65%	3	G	58%
1	2	4	78%	68%	3	G	60%
7	5	5	65%	60%	4	A	50%

Search the id appearing maximum number of times in a row in the input table 5.1.2, and then locate the record headed by that id. In the above table, id number 4 is repeated maximum number of times compared to id number 3. Hence row headed by id 4 in the input table i.e., 5.1.1 is selected as the topmost row in the resultant table. Same procedure will be applied for all the rows in table 5.1.2 to get the output of the genetic algorithm.

6. K-MEANS ALGORITHM (DATA MINING APPROACH)

After the initial preprocessing using evolutionary approach, academies need to be grouped or clustered. Hence K-means algorithm, a clustering algorithm is applied for grouping academies as excellent, good, average and poor. Extract of output of k-means algorithm is given below.

Table 5 Extract of the output represented as clusters

Id	Rank
4,3,5	E
9,10,6	G
2,8,1	A
7	P

From the table we can infer that ids 4, 3, and 5 is grouped as excellent, ids 9, 10, and 6 as good , 2, 8, and 1 as average, and 7 as poor.

7. CONCLUSION AND FUTURE ENHANCEMENT

In this study, an integrated approach using genetic and data mining has been used. Genetic algorithm is used for preprocessing and the data thus obtained is mined using k-means algorithm to get clusters consisting of the academies as its elements. Result proved that k-means with 81.5% outperformed compared to other algorithm SVM(81%) in terms of accuracy. Clusters thus obtained as excellent, good, average and poor not only help the students in the selection of a right academy during admission but also act as a good aid in easing the decision making process.

8. REFERENCES

[1] Arora et al., 'Admission Management through Data Mining using WEKA' International Journal of Advanced Research in Computer Science and Software Engineering, Volume 3, Issue 10, pp. 674-678, October 2013

[2] Arora et al., 'Location wise Student Admission Analysis', International Journal of Computer Science and Information Technology & Security (IJCSITS), ,Vol. 2, No.6, ISSN: 2249-9555, December 2012

[3] Arora et al. 'subject distribution using data mining' International Journal of Research in Engineering and Technology Volume (IJRET) 02, Issue: 12 , eISSN: 2319-1163 | ISSN: 2321-7308, Dec-2013,

[4] M. Sukanya, S. Biruntha, Dr.S. Karthik, T. Kalaikumaran 'Improving the Performance in Education Sector using Data Mining Techniques' Data mining and knowledge engineering, Vol 4, No 8 (2012).

[5] JayanthiRanjan, Kamna Malik, "Effective educational process: a data-mining approach" ,VINE, Vol. 37 Iss: 4, pp.502 – 515, 2007

[6] Behrouz Minaei-Bidgoli, William F. Punch, 'Using Genetic Algorithms for Data Mining Optimization in an Educational Web-Based System', Genetic and Evolutionary Computation Conference Chicago, Volume 2724, pp 2252-2263, 2003

[7] Behrouz Minaei-Bidgoli , Deborah A. Kashy , GerdKortemeyer , William F. Punch 'predicting student performance: an application of data mining methods with the educational web-based system LON-CAPA', 33rd ASEE/IEEE Frontiers in Education Conference, 2003

[8] Mu-Jung Huang a, Hwa-Shan Huang a, Mu-Yen Chen 'Constructing a personalized e-learning system based on genetic algorithm and case-based reasoning approach' Expert Systems with Applications (2006)

[9] AzinKhosravi Khorashad1, KianooshZakerHaghighy, 'Application of Genetic Algorithm in Regional Planning', J. Basic. Appl. Sci. Res., 2(11)11428-11433, 2012

[10] Sandeep singhrawat, lakshmirajamani 'Timetable prediction for technical educational system using genetic algorithm', Journal of Theoretical and Applied Information Technology, 2010

[11] AlirezaArabasadi et al., 'A New Hybrid Algorithm for Traveler Salesman Problem based on Genetic Algorithms and Artificial Neural Networks' International Journal of Computer Applications (0975 – 8887) Volume 24– No.5, June 2011

[12] Krishna.k, Murty M.N "Genetic k-means algorithm", volume 29, issue 3, 1999 pages 435-439.

[13] Zhexue Huang "Extensions to the k-means algorithm for clustering large data sets with categorical values", volume 2, issue 3, pages 283-304, 1998.

[14] Leon Bottou, YoshuaBengio "Convergence properties of the k-means algorithms", 1995.