Volume 4, Issue 6, June 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

Admission Forecasting: Using Unsupervised Learning

[1]Deepika Sharma, [2]Aman Paul
[1]Research Scholar, [2]Assistant Professor
Department of CSE, Eternal University, Baru Sahib
Himachal Pradesh-173101, India

ABSTRACT- Most of the institutions is now in self financing mode. This paper uses data mining techniques in the field of education so as to improve the quality of admissions in the institute. It needs a lot of effort to have good strength of students and meet the expenses as well. This paper focuses on identifying the success ratio i.e. actual admissions taken. The result of analysis will help in knowing the best of the used techniques and will also assist in making a proper plan to have quality admissions along with the good strength of students.

KEYWORDS - Data mining, Clustering, Clustering Techniques, Success Ratio, Accuracy.

I. INTRODUCTION

The educational institutes that are in self financing mode face many problems to have good strength of students.
All the institutes follow their own admission procedures like written exam, group discussions and interviews. Some of the institutions get the enquiry forms filled by the admission seekers. In this paper we are using Data Mining technique to find out the enquiry forms that have actually converted into admissions. Data mining is the process of extracting interesting information from large amount of data stored in different databases or data warehouses. Data mining tools can be used to predict future in the field of business, knowledge driven systems. Several Data mining techniques are present like classification, association, clustering etc. This paper makes use of Clustering which means identifying and making groups. There are various clustering algorithms namely Simple K-Means, EM, Filtered Clustering, Farthest First, Hierarchical, DBSCAN etc.

II. METHODOLOGY

In the proposed research work we have used the data mining technique namely clustering. **Clustering** or **cluster analysis** refers to grouping of a set of objects such that the objects in the one group/**cluster** are more similar (in some sense or another) to each other than to those in other groups /clusters. Various clustering algorithms have been used in the proposed research.

Simple K-Means:

- The K-Means algorithm is an method to cluster objects based on their attributes into k partitions.
- It assumes that the *k* clusters exhibit Gaussian distributions.
- It assumes that the object attributes form a vector space.
- The objective it tries to achieve is to minimize total intra-cluster variance.

EM:

- The EM algorithm estimates the parameters of a model iteratively.
- Start with initial values for the parameters
- Calculate the cluster probabilities for each instance
- Re-estimate the values for the parameters
- Repeat

Farthest First:

Farthest first is a variant of K-Means that places each cluster centre in turn at the point farthest from the existing cluster centres. This point must lie within the data area. Farthest-point heuristic method is suitable for large-scale data mining applications. Farthest-point heuristic based method has the time complexity O (nk), where n is number of objects in the dataset and k is number of desired clusters.

Filtered Clustering:

This algorithm is used for the purpose of filtering the information or pattern. In this the user provides the keywords or a sample set of relevant information. On the arrival of new information they are compared against the filtering profile and the information matching the keywords is presented to the user. The user is not provided with the details of filtering algorithms used by the system. Filtering of information or pattern by collaboration of multiple agents, data sources and viewpoints is referred to as collaborative filtering.

Hierarchical Clustering:

- Produces a nested sequence of clusters, a tree, also called Dendrogram.
  Types:
- Agglomerative (bottom up) clustering: It builds the dendrogram (tree) from the bottom level, and
  - merges the most similar (or nearest) pair of clusters
  - stops when all the data points are merged into a single cluster (i.e., the root cluster).
- Divisive (top down) clustering: It starts with all data points in one cluster, the root.
  - Splits the root into a set of child clusters. Each child cluster is recursively divided further
  - stops when only singleton clusters of individual data points remain, i.e., each cluster with only a single point

Make Density Based Clustering:

The cluster will be constructed based on the density properties of the database are derived from a human natural clustering approach. The clusters and consequently the classes are easily and readily identifiable because they have an increased density with respect to the points they possess.The elements of the database can be classified in two different types: the border points, the points located on the extremities of the cluster, and the core points, which are located on its inner region.

DBSCAN:

Density based clustering algorithm has played a vital role in finding non linear shapes structure based on the density. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is most widely used density based algorithm. It uses the concept of density reachability and density connectivity.

III. RESULT

3.1 DATA SET AND TOOL USED

In order to compare the different approaches I have used admission dataset of "Rayat & Bahra College of Engineering and Bio Technology for women, Mohali" on which various algorithms from Weka tool are applied. Table shows the selection of dataset and its properties. The dataset was analysed at 199 instances with 12 attributes in the area of Education.

Table 3.1: The dataset used for the experiments and its properties

Data Set	Instance	Attributes	Area	Missing Value
Admission	199	12	Education	0

For clustering algorithms, I have explored admission dataset and performed comparisons on the basis of Success Ratio and accuracy. In this the confusion matrix is used for the computation of success ratio. In this way we find the algorithm with the highest success ratio and accuracy.

3.2 RESULTS OF CLUSTERING ALGORITHMS

3.2.1 SIMPLE K-MEANS ALGORITHM

Table 3.2: Success Ratio in case of Simple K-Means Algorithm

Data Set	Cluster	Success Ratio(in percentage)
Admission	Cluster0	90.10
	Cluster1	79.62

I have applied Simple K-Means clustering algorithm on Admission data set which contains 199 instances. After applying the algorithm we have:

Clustered instances-

0	91
1	108`

Class to clusters-

0	**1**
82	86
9	22

The success ratio for cluster0 i.e., 90.10% is selected in case of simple k-means algorithm.

3.2.2 EM ALGORITHM

Table 3.3: Success Ratio in case of EM Algorithm

Data Set	Cluster	Success Ratio(in percentage)
Admission	Cluster0	95.58
	Cluster2	78.40

On applying EM clustering algorithm on Admission data set with 199 instances we have:

Clustered instances-

0	68
1	43 (No class)
2	88

Class to clusters-

0	**1**	**2**
65	34	69
3	9	19

Neglecting cluster1 since no class is formed so the success ratio for cluster0 i.e., 95.58% is selected in case of EM algorithm.

3.2.3 FARTHEST FIRST ALGORITHM

Table 3.4: Success Ratio in case of Farthest First Algorithm

Data Set	Cluster	Success Ratio(in percentage)
Admission	Cluster0	82.85
	Cluster1	86.17

I have applied Farthest First clustering algorithm on Admission data set which contains 199 instances. After applying the algorithm we have:

Clustered instances- **0** 105 **Class to clusters-** **0** **1**
1 94 87 81
18 13

The success ratio for cluster1 i.e., 86.17% is selected in case of Farthest First algorithm.

3.2.4 FILTERED CLUSTERED ALGORITHM

Table 3.5: Success Ratio in case of Filtered Clustered Algorithm

Data Set	Cluster	Success Ratio(in percentage)
Admission	Cluster0	90.10
	Cluster1	79.62

On applying Filtered clustered algorithm on Admission data containing 199 instances we get:

Clustered instances- **0** 91 **Class to clusters-** **0** **1**
1 108 82 86
9 22

The success ratio for cluster0 i.e., 90.10% is selected in case of Filtered Clustered algorithm.

3.2.5 HIERARCHICAL CLUSTERING ALGORITHM

Table 3.6: Success Ratio in case of Hierarchical Algorithm

Data Set	Cluster	Success Ratio(in percentage)
Admission	Cluster0	84.77
	Cluster1	50

On applying Hierarchical clustering algorithm on Admission data set which contains 199 instances, we get:

Clustered instances- **0** 197 **Class to clusters -** **0** **1**
1 2 167 1
30 1

The success ratio for cluster0 i.e., 84.77% is selected in case of Hierarchical algorithm.

3.2.6 MAKE DENSITY BASED CLUSTERED ALGORITHM

Table 3.7: Success Ratio in case of Make Density Based Algorithm

Data Set	Cluster	Success Ratio(in percentage)
Admission	Cluster0	92.13
	Cluster1	78.18

Here I have applied Make Density Based clustering algorithm on Admission dataset which contains 199 instances. After applying the algorithm we have:

Clustered instances- **0** 89 **Class to clusters-** **0** **1**
1 110 82 86

7 24

The success ratio for cluster0 i.e., 92.13% is selected in case of Make Density Based algorithm.

3.2.7 DBSCAN ALGORITHM

Table 3.8: Success Ratio in case of DBSCAN Algorithm

Data Set	Cluster	Success Ratio(in percentage)
Admission	Cluster1	80.95
	Cluster2	100

I have applied DBSCAN clustering algorithm on Admission dataset with 152 instances. After applying the algorithm we have:

Clustered instances-			**Class to clusters -**			
	0	18 (No Class)		**0**	**1**	**2**
	1	42		15	34	33
	2	33		3	8	0

Neglecting cluster0 since no class is formed so the success ratio for cluster2 i.e., 100% is selected in case of DBSCAN algorithm.

Table 3.9: Combined Success Ratio of all clustering algorithms

Data Set	Algorithm	Success Ratio (in percentage)
Admission	Farthest First	86.17
	Hierarchical	84.77
	EM	95.58
	Make Density Based	92.13
	Simple K-Means	90.10
	Filtered	90.10
	DBSCAN	100

In the Table 3.9 there is combined success ratio of all algorithms on the Admission dataset, which makes it clear that Hierarchical algorithm has lowest success ratio and DBSCAN has highest success ratio of all the algorithms used.

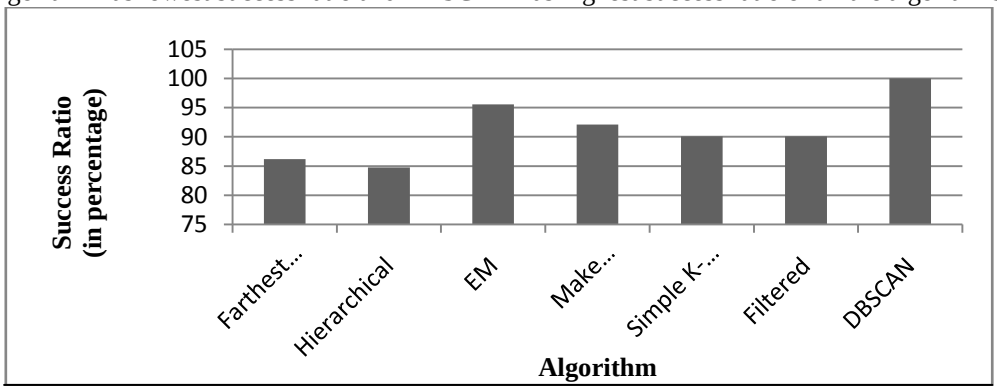

Fig. 3.1 Graphical Representation of Success Ratio of all Clustering Algorithms

Table 3.10: Accuracy comparison of all clustering algorithms

Data Set	Algorithm	Accuracy (in percentage)
Admission	EM	42.21
	Filtered Clustered	52.26
	Simple K-Means	52.26
	Make Density Based	53.26
	Farthest First	50.26
	DBSCAN	73.86
	Hierarchical	84.42

The Table 3.10 contains the combined success ratio of all algorithms on the Admission data set, which shows that EM has lowest accuracy and Hierarchical algorithm has highest accuracy of all the algorithms used.

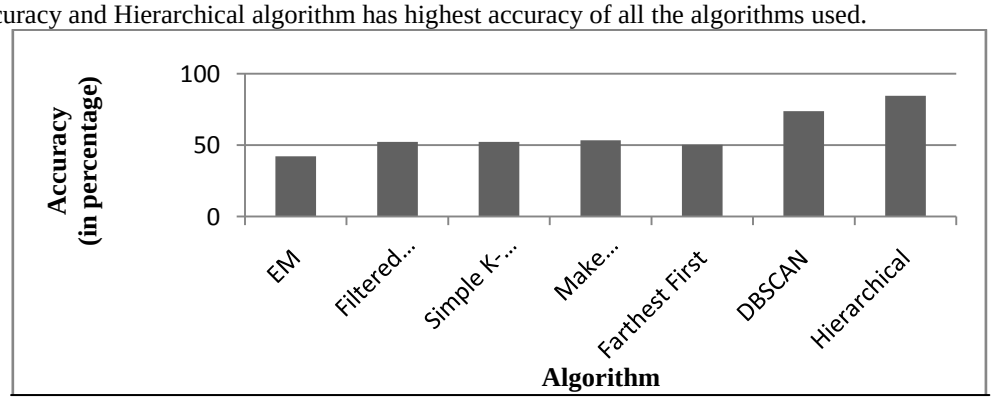

Fig. 3.2 Graphical Representation of Accuracy of all Clustering Algorithms

IV. DISCUSSION

Various Clustering Algorithms (Simple K-Means, EM, Farthest First, Filtered Clustered, Hierarchical, Make Density Based and DBSCAN) are applied on Admission data set to find out the success ratio. The success ratio depends on how many enquiry forms filled by the students for admission purpose turn successfully into the admission taken.

When Simple K-Means algorithm is applied on the given data set the clustered instances are obtained. In cluster0 there are 91 instances and in cluster1 there are 108 instances. Success Ratio is computed by using the confusion matrix. In this case, the Success Ratio comes out to be 90.10% for cluster0 and 79.62% for cluster1. So the cluster0 with success ratio 90.10% is selected. So, it means that 90.10% enquiry forms have turned into success i.e., admission taken.

When EM algorithm is applied on the given data set the clustered instances are obtained. In cluster0 there are 68 instances, in cluster1 there are 43 instances and in cluster2 there are 88 instances. In this case, the Success Ratio comes out to be 95.58% for cluster0 and 78.40% for cluster2. So the cluster0 with success ratio 95.58% is selected. So, it means that 95.58% enquiry forms have turned into success i.e., admission taken. Cluster1 is neglected since no class was formed.

When Farthest First algorithm is applied on the given data set the clustered instances are obtained. In cluster0 there are 105 instances and in cluster1 there are 94 instances. The Success Ratio comes out to be 82.85% for cluster0 and 86.17% for cluster1. So the cluster1 with success ratio 86.17% is selected. So, it is clear that 86.17% enquiry forms have turned into admission taken.

Filtered Clustered algorithm is applied on the given data set the clustered instances are obtained as- cluster0 contains 91 instances and in cluster1 contains 108 instances. In this case, the Success Ratio comes out to be 90.10% for cluster0 and 79.62% for cluster1. So the cluster0 with success ratio 90.10% is selected which means that 90.10% enquiry forms have turned into success.

Hierarchical algorithm is applied on the given data set the clustered instances are obtained as- cluster0 has 197 instances and in cluster1 there are 2 instances. In this case, the Success Ratio comes out to be 84.77% for cluster0 and 50% for cluster1. So the cluster0 with success ratio 84.77% is selected which means that 84.77% enquiry forms have turned into success.

Make Density Based algorithm is applied on the Admission data set the clustered instances are obtained as- cluster0 contains 89 instances and cluster1 contains 110 instances. In this case, the Success Ratio comes out to be 92.13% for cluster0 and 78.18% for cluster1. So the cluster0 with success ratio 92.13% is selected which means that 92.13% enquiry forms have turned into admission taken.

When DBSCAN algorithm is applied on the given data set the clustered instances are obtained as- in cluster0 there are 18 instances, in cluster1 there are 42 instances and in cluster2 there are 33 instances. The Success Ratio is 80.95% for cluster1 and 100% for cluster2. Cluster0 is neglected since no class is formed. So the cluster2 with success ratio 100% is selected which means that 100% enquiry forms have turned into success.

On combining the results of all the used clustering algorithms i.e., the success ratio of all the algorithms we conclude that Hierarchical algorithm has lowest success ratio and DBSCAN has highest Success Ratio. Also we computed the accuracy for these algorithms and the result shows that EM algorithm has lowest accuracy and Hierarchical algorithm has highest accuracy.

V. CONCLUSION

In this research paper, various clustering algorithms are applied to the admission dataset. As a result of which DBSCAN algorithm provides the highest success ratio as compared to other clustering algorithms. This technology helps to analyse the students' details and which of the students are most likely to take admission in the institute. It also helps the in making a proper plan to have a good strength of students and have the quality admissions as well.

REFERENCES

[1] Oded Maimon and Lior Rokach, "DATA MINING AND KNOWLEDGE DISCOVERY HANDBOOK", *Springer*, pp. 321-352, 2005.

[2] Amandeep Kaur Mann and Navneet Kaur, "Survey Paper on Clustering Techniques", *International Journal of Science, Engineering and Technology Research*, vol. 2, pp. 803-806, April 2013.

[3] Pradeep Rai and Shubha Singh, " A Survey of Clustering Techniques", *International Journal of Computer Applications*, October 2010.

[4] B.S. Everitt, S. Landau and M. Leese, "Cluster Analysis*", Oxford University* Press, fourth edition, 2001.

[5] A. K. Jain and R. C. Dubes, "Algorithms for Clustering Data", *Printice Hall*, 1988.

[6] Pragati Shrivastava and Hitesh Gupta, "A Review of Density-Based clustering in Spatial Data", *International Journal of Advanced Computer Research*, pp. 2249-7277, September 2012.

[7] M. R. Anderberg, "Cluster Analysis for Applications", *Academic Press*, 1973.

[8] Tan P, Steinbach M and Kumar, "Introduction to Data Mining", *Addison Wesley*, vol.1, pp. 157-169, 2006.

[9] Kalyani M Raval, "Data Mining Techniques", *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 2, pp. 439-442, 2012.

[10] Bhoj Raj Sharma and Aman Paul (2013), "Clustering Algorithms: Study and Performance Evaluation Using Weka Tool", *International Journal of Current Engineering and Technology*, vol. 3, pp. 1094-.
1098, 2001.

[11] S. Revathi and Dr. T.Nalini (2013) "*Performance Comparison of Various Clustering Algorithm",* International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 2, 2013, pp. 67-72.

[12] Rakesh Kumar Arora and Dr. Dharmendra Badal (2013) "*Admission Management through Data Mining using WEKA",* International Journal of Advanced Research in Computer Science and Software Engineering, vol. 3, 10, 2013, pp. 674-678.

[13] Pallavi and Godara S (2011) "A Comparative Performance Analysis of Clustering Alogrithms". *International Journal of Engineering Research and Application*, 1: 441-445.

[14] Shreya Jain and Samta Gajbhiye (2012) "A Comparative Performance Analysis of Clustering Algorithms", International Journal of Advanced Research in Computer Science and Software Engineering , vol. 2, pp. 441-445.

[15] Sunila Godara and Ritu Yadav (2013) *"Performance Analysis Of Clustering Algorithms For Character Recognition Using Weka Tool"* , International Journal of Advanced Computer and Mathematical Sciences, vol. 4, pp.119-123.

[16] D. Prabhu, Dr.K.Vivekanandan, R. Vijayanandh (2012) *" Performance Analysis of Clustering Techniques to Normal and Uniform Distribution of Data Points",* International Journal of Engineering Research and Applications, vol.2, pp.861-870.

[17] Vibekananda Dutta, Krishna Kumar Sharma and Deepti Gahalot (2012) *" Performance Comparison of Hard and Soft Approaches for Document Clustering",* International Journal of Computer Applications, vol.41, pp.44-48.

[18] Er. Arpit Gupta , Er.Ankit Gupta and Er. Amit Mishra, "RESEARCH PAPER ON CLUSTER TECHNIQUES OF DATA VARIATIONS*", International Journal of Advance Technology & Engineering Research*, vol. 1, pp. 39-47, 2011.

[19] PAULRAJ PONNIAH, "DATA WAREHOUSING FUNDAMENTALS FOR IT PROFESSIONALS", *A John Wiley & Sons, Publication*, 2nd Edition New Jersey.

