International Conference on "Recent Research Development in Science, Engineering and Management"
Mahratta Chamber of Commerce Industries & Agriculture, Pune (ICRRDSEM-16)
29th May 2016, www.conferenceworld.in ISBN: 978-81-931039-0-6

ADMISSION MANAGEMENT USING RELATIONAL K-MEANS CLUSTERING

Ameya Joshi[1], Manjusha Singh[2]

[1,2] *Computer Science and Engineering,*

Chhatrapati Shivaji Institute of Technology, Bhilai, (India)

ABSTRACT

Data Mining is the activity of analyzing data from different perspectives and incorporating it into useful information. From the last decades, number of students is interested and focuses their carrier in engineering meanwhile many private as well as government colleges came into existence. Now the students is having various viewpoint and thoughts regarding the colleges related to different perspectives. In this paper, the technique of data mining is used to find out the willingness of the students and to analyze the activities related to their performance. This will help the college management to get fruitful admissions in the successive years.

Keywords: Clustering, Data Mining, KDD, WEKA

I. INTRODUCTION

The term data mining is formally known as Knowledge mining from data or Knowledge data mining. It is the process of analyzing data from different vision and detailing it into useful information. Data Mining is an analytic process designed to explore large amounts of data typically business oriented or market related also known as "big data" in search of consistent patterns and/or appropriate relationships between objects, and then to validate the findings by applying the detected patterns to new subsets of data. The overall goal of Data mining process is to extract hidden information which are potentially useful and structure it for further use. Business questions can be easily answered by data mining tools that traditionally were too time taking to resolve.

Data mining techniques can be used rapidly on various existing software and hardware platforms to enhance the quality of existing information resources that can be integrated with upcoming products and systems as they are brought on-line. In this paper we are focusing on techniques of data mining: Clustering technique is the crucial task for distinguish the data in to sub sets, There are various applications of Data mining such as in marketing, web mining, scientific data analysis and many more.

Data mining can be defined as extracting the hidden predictive information from huge amount of data set. Alternatively, it has been called exploratory data analysis, data driven discovery and deductive learning. It is the computational process of discovering patterns in large data sets. The main goal of data mining is prediction and Classification. Data mining is the most common type of direct business applications. The fundamental and perspective view provided by data mining move beyond the past event of retrospective tools typical of decision support systems.

The two models of Data mining are as follows:

- Predictive
- Descriptive

A **predictive model** makes a prediction about values of data using known results found from different data. Predictive modeling may be made based on the use of other historical data. A **descriptive model** identifies patterns or relationships in data. Unlike the predictive model, a descriptive model serves as a way to explore the properties of the data examined, not to predict new properties.[Margaret book].

Descriptive model comprises of four sub models like clustering, summarization, association rules & sequence discovery. In this project we are focusing on one of the sub models of descriptive modeling. Clustering is the identification of classes, called clusters or groups for a set of objects whose classes are unknown. The definition of clustering can also be express as the process of nominating a set of objects into batch (called cluster) so that the objects in the same clusters are more identical to each other to those in other cluster.

In this paper, the real time data is introduced from the college of Nasik, the details of the students were collected at the time of admissions. The number of students involved is 250 and their around 14 parameters through which the analysis takes place. The parameters used for the analysis with their values are shown in Table 1.

II. LITERATURE REVIEW

Rakesh kumar arora, (krishna engineering college ghaziabad, up) Dr.Dharmendra badal (bundelkhand university, ghaziabad, up)(2013) had given the procedure to classify the set of data through the certain number of clusters. They found the number of cluster for classifying the students according to their parametric value. Also determine the reasons for decline in quality of admissions taken in the institute over the year. In this paper, author explained and clearly indicated that 57 & 51 out of 61 inquiries are being converted into admission. In this paper author conclude that the tools are quite easy to work with and the methods applied for admission management procedure were simple to execute.[1]

Dr. Sudhir B. Jagtap SVITN (Udgir) M.H. Dr. Kodge B. G. SVITN (Udgir) M.H analyzed that, the data mining tool WEKA has tremendous power to present analytical facts and knowledge extraction from database. With the help of hidden knowledge one can explore the possibilities of the outcomes for their organizational development & realted issues and plays a vital role in e-governance for future planning and development issues.[2]

Narendra Sharma, Aman Bajpai , Ratnesh Litoriya Jaypee University of Engg. & Technology(2012) had concluded that k-means clustering algorithm is simplest algorithm as compared to other algorithms. Each and every algorithm is having certain boundaries with which they perform their tasks. However, all the clustering algorithm is having many disadvantages related to either assumption of the values or its boundaries. But apart from that WEKA is the powerful tool for data mining & we can't require deep knowledge of algorithms for working in WEKA. That's why WEKA is more suitable tool for data mining applications.[3]

R. ROBU, C.HORA University "Politechnica" Romania (2012) analyzed that in the field of medical data mining it is necessary to evolve all the possibilities from the given input and collect those outcomes such as images, signals, and related information which will help to improve diagnosis from all the diseases for efficiently prevent them and treat them more easily at the given frame of time. In this paper author improved the classification interface of WEKA and also improved the diagnostic 3D instances which aims to extract useful knowledge from the data. [4]

Suhem Parack, Zain Zahid, Fatima Merchant(2012) has analyzed that K-means clustering algorithm and Apriori algorithm in WEKA plays the same role more efficiently just like data mining techniques in the field of

education. It is very difficult to sort out huge data sets manually. With the help of these two algorithms authors can easily analyzed the hidden patterns about their learning style and the performance and behavior of students. [5]

III. OBJECTIVE OF STUDY

- To Study of existing data analysis clustering technique.
- To analyze complexity and outlier issue in Algorithm
- To deploy the clustering methodology.
- To compare the results with K-Means algorithm in terms of **efficiency and throughput.**

Here efficiency is in terms of number of student converting their enquiries in to admission and throughput deals with the conversion rate of the students.

3.1 Methodology

From the reviews and survey, it is clear that clustering is the most vital task in data mining. In the Educational data mining, clustering is used to explore the dimensions of the student's performance. Clustering is an ultimate task of explorative data mining, and a common technique for statistical data analysis used in various fields such as information extraction, pattern evaluation, machine learning techniques, image analysis. Clustering is the activity of composing objects into groups whose members are similar in some properties. A cluster is a group of objects which are "identical" between them and are "dissimilar" to the objects belonging to other clusters. In this regard previously object clustering is used and the value of the cluster is calculated. But in this paper we want to compare the results of object clustering with the relational clustering. The real time data is collected from the school of Nasik with the attributes such as (school name, 12th percentage, scholarship, family_income, family_member, etc.)

Table:1 Parameters to be used

In this paper K-Means algorithm is used. In the past work K-Means is being used for the analysis of the data obtained from the admission procedure. The data set is collected from the college of Ghaziabad. The number of student involved is 129 and is done on the basis of different parameters. The author concluded that this model will help to assist the management persons to filter good student who are willing to join the institute.[1].The concept of object clustering is to find out the relationship between the students with the attributes but the interesting fact is that at the time of comparison with relational clustering, relational clustering gives better

accuracy and result than object clustering. But each and every good thing come up with little disadvantages, K-Means clustering also facing difficulties while comparing quality of the cluster and it does not work efficiently with non-globular clusters.

3.2 Indexes

- Davies-Bouldin index (DB)

$$DB = \frac{1}{c}\sum_{i=1}^{c} Max_{i\neq j}\left\{\frac{d(x_i)+d(x_j)}{d(c_i,c_j)}\right\}$$

c = number of clusters, i, j are cluster labels, $d(Xi)$ and d(Xj) are all samples in clusters i and j to their respective cluster centroids, $d(ci, cj)$ is the distance between these centroid.

Smaller value of DB indicates a better clustering solution.

- RMSSTD (root – mean– square standard deviation)

The values of RS range from 0 to 1 where 0 means there are no difference among the clusters and 1 indicates that there are significant difference among the clusters.

- Dunn index

$$Dunn = min_{1\leq i\leq c}\left\{min\left\{\frac{d(c_i,c_j)}{max_{1\leq k\leq c}(d(X_k))}\right\}\right\}$$

$d(ci, cj)$ defines the inter-cluster distance between cluster Xi and Xj ;

$d(Xk)$ represents the intra-cluster distance of cluster (Xk) and c is the number of cluster of dataset.

Large values of index Dunn correspond to good clustering solution.

Figure:1 GUI for student analysis system

In the above figure, the four labels help us to display the attributes of the student which is selected from the dropdown list. In the review panel the results are shown after clustering is performed.

Figure: 2 Selection of Data

In this model, the data is loaded from the excel file. The file contains the records and relevant attributes.

Figure: 3 Displaying student information

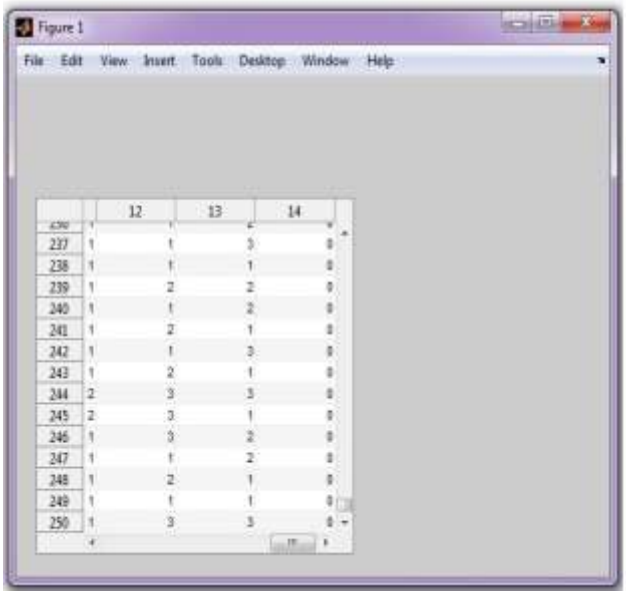


Figure: 4 Data Generated for Object Clustering

The matrix is generated for object clustering of 250*14. Here 250 is the number of students and having 14 attributes. On this object matrix clustering is applied i.e. K-Means algorithm.

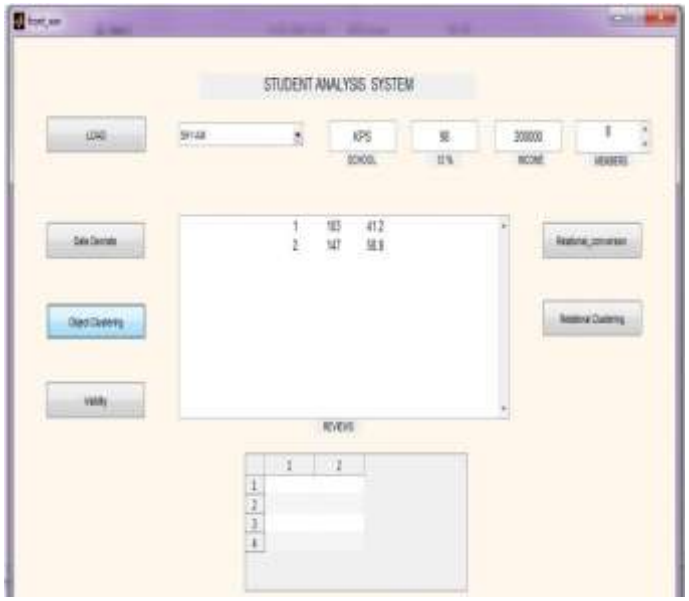
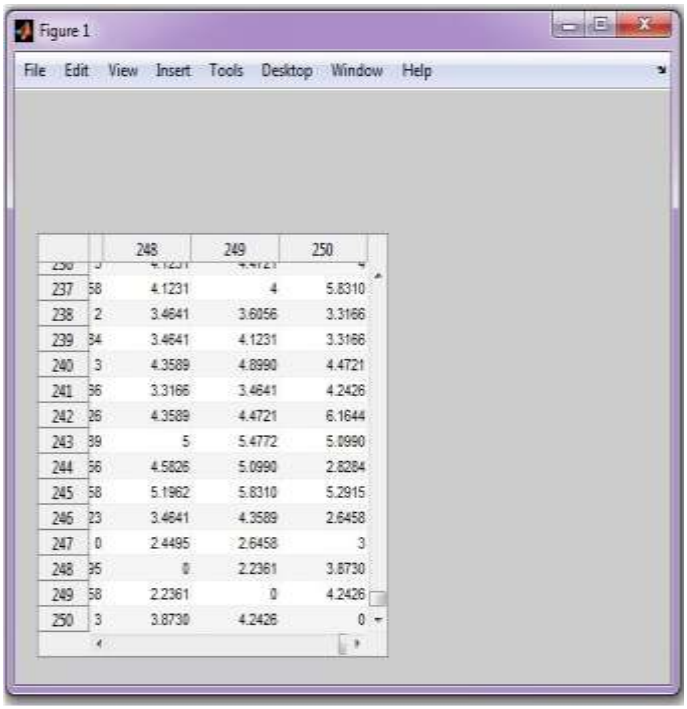


Figure: 5 Output of object clustering

The results show that, after applying K-Means algorithm, two clusters are formed. Out of 250 students, 103 students not taking admissions and 147 students converting their enquires into admission.



		248	249	250
237	58	4.1231	4	5.8310
238	2	3.4641	3.6056	3.3166
239	34	3.4641	4.1231	3.3166
240	3	4.3589	4.8990	4.4721
241	66	3.3166	3.4641	4.2426
242	26	4.3589	4.4721	6.1644
243	39	5	5.4772	5.0990
244	66	4.5826	5.0990	2.6284
245	58	5.1962	5.8310	5.2915
246	23	3.4641	4.3589	2.6458
247	0	2.4495	2.6458	3
248	95	0	2.2361	3.8730
249	58	2.2361	0	4.2426
250	3	3.8730	4.2426	0

Figure: 6 Data Generated for Relational Clustering

The matrix is generated for relational clustering of 250*250. Where the relation is calculated on many to many basis. On this relational matrix clustering is applied i.e. K-Means algorithm.

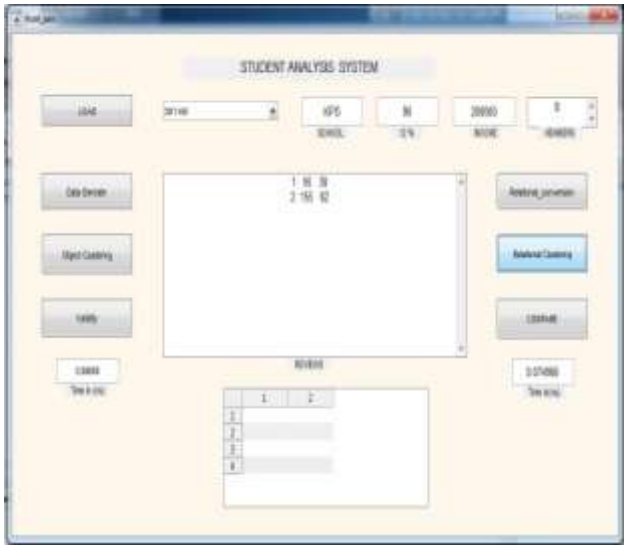


Figure: 7 Output of relational clustering

After calculations, the relational clustering converts the data with better efficiency than object clustering. Out of 250 students, 96 students not taking admissions and 155 students converting their enquires into admission.

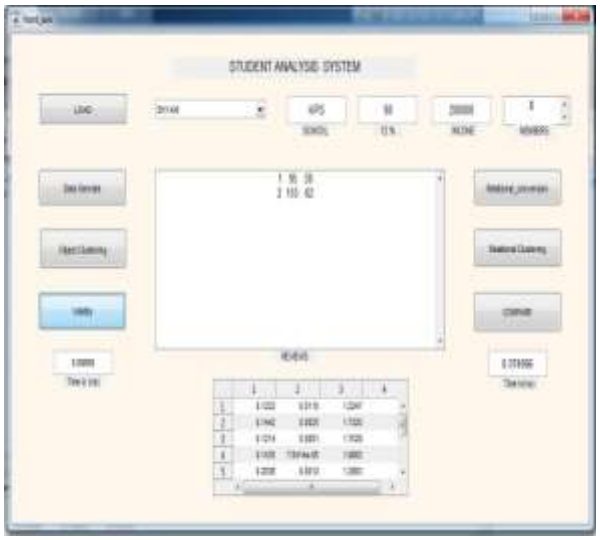


Figure: 8 Indexes value

From the indexes value shown in this figure, it is clear that two cluster are required for the data set. And it helps to identify the number of cluster required for huge data set.

IV. RESULTS AND CONCLUSION

- On the basis of these outcomes, we found that K-Means clustering is the easiest algorithm with relational data and it comes with good results as compared to the object clustering.
- The value of cluster-1 stands 41.2 and cluster-2 58.8 in object clustering. While the values of cluster-1 stands 38 and cluster-2 stands 62 in relational clustering.
- The validity of the clusters depends on the value of the indexes. Smaller the index values, larger the cluster approximation value.

This will help in identifying the set of students that need to be focused to actually convert the enquiry into admission. Management will get beneficial inputs.

V. FUTURE WORK

We wish to develop the algorithm with the concept of clustering with relational data, and to deploy the same in other management procedure to fulfill the requirement. Each work has some limitations and improvement area. The methodology adopted here will discover fact of non converting admissions. This methodology is based on clustering for managerial aspect and system improvement.

Thus the future scope is use different techniques of visualization and association rules can be applied. Another area for improvement can be towards the web deployment. and to compare the result with existing algorithm and we try to observe in terms of accuracy and throughput.

VI. REFERENCES

[1] Rakesh Kumar Arora, Dr. Dharmendra Badal "Admission Management through Data Mining using WEKA", International Journal, Volume 3, Issue 10, October 2013. Page No. [674-678].

[2] Dr. Sudhir B. Jagtap and Dr. Kodge B. G., "Census Data Mining and Data Analysis using WEKA", International Conference in "Emerging Trends in Science, Technology and Management-2013, Singapore. Page No. [35-40].

[3] Narendra Sharma, Aman Bajpai, Mr. Ratnesh Litoriya "Comparison the various clustering algorithms of WEKA tools' , International Journal of Emerging Technology and Advanced Engineering (ISSN 2250-2459, Volume 2, Issue 5, May 2012) Page No-[73-80].

[4] R Robu and C. Hora, "Medical data mining with extended WEKA " , IEEE 16TH International conference on intelligent Engineering Systems , June 13-15, 2012 page No. [347-350].

[5] Suhem Parack, Zain Zahid Fatima Merchant., "Application of data mining in Educational databases for predicting academics trends and patterns. ", IEEE-2012.

[6] Anand V Saurkar, Vaibhav Bhujade, Priti Bhagat, Amit Khaparde "International Journal of Advanced Research in Computer Science and Software Engineering, Volume-4, Issue 4, April 2014 ISSN: 2277 128X . Page No-[98-101].

[7] M Bhoomi, "Enhanced K-Means Clustering algorithm to reduce time complexity for numeric values ", International Journal of Computer Science and Information Technologies, Volume 5(1), 2014, ISSN: 0975-9646 page No. [876-879].

[8] http://www.ijcsit.com/docs/Volume%205/vol5issue01/ijcsit20140501189.pdf

[9] http://www.iaeng.org/publication/WCE2009/WCE2009_pp308-312.pdf

[10] http://documents.software.dell.com/Statistics/Textbook/Data-Mining-Techniques

[11] http://esatjournals.net/ijret/2013v02/i11/IJRET20130211019.pdf

[12] https://en.wikipedia.org/wiki/K-means_clustering

[13] https://www.google.co.in/search?q=flowchart+of+k-means+clustering+algorithm&biw=1024&bih=639&tbm=isch&tbo=u&source=univ&sa=X&sqi=2&ved=0ahukewiy2fyh2kdjahubh5qkhxvgbkiqsaqiia