ISSN 22773061

Implementation and Analysis of Clustering Algorithms in Data Mining

Prabhjot Kaur*, Robin Parkash Mathur**
Saini_prabhjot@hotmail.com
*CSE, GES Polytechnic College Hoshiarpur
**CSE, Lovely Professional University Phagwara

Abstract:

Data mining plays a very important role in information industry and in society due to the presence of huge amount of data. Organizations in the whole world are already aware about data mining. Data mining is the process which uses various kinds of data analysis tools to obtain patterns which also referred to as knowledge discovery from data. Clustering is called unsupervised learning algorithm as groups are not predefined but defined by the data. There are so many research areas in data mining. This paper is focusing on performance and evaluation of clustering algorithm: K-means, SOM and HAC. Evaluations of these three algorithms are purely based on the survey based analysis. These algorithms are analyzed by applying on the data set of banking which is a very high dimensional data. Performances of these algorithms are also compared with each other. Our results indicate that SOM technique is better than k-means and as good as or better than the hierarchical clustering technique. We have also generated one code in Orange Python which is the enhanced algorithm based on the hybrid approach of SOM, K-means and HAC.

Council for Innovative Research
Peer Review Research Publishing System
Journal: INTERNATIONAL JOURNAL OF COMPUTERS & TECHNOLOGY
Vol 6, No 1

editor@cirworld.com
www.cirworld.com, member.cirworld.com

ISSN 22773061

Introduction:

Any clustering technique is [1] having the purpose of evolving a K*n partition matrix U(x) of a dataset. Clustering techniques broadly fallen into two main classes, partitioning and hierarchical. In any clustering system two fundamental questions arises: 1) How many clusters are actually present in the data and 2) how real or good is the clustering itself. Data mining algorithms [2] for processing large amount of data must be scalable. Algorithms of data mining which are used for processing data with changing patterns must have the capability of updating and learning data. One of the traditional data mining techniques is[3] clustering which is an unsupervised learning paradigm where clustering methods try to identify the inherent grouping of text document, such that a set of clusters formed which exhibit high intracluster similarity an low intercluster similarity.

K-means clustering algorithm is a very simple iterative method which is used for partitioning a dataset into k number of clusters which one is purely user specified.[4] This algorithm can easily adapt to dynamic P2P network where existing nodes drop out and new nodes join in during the execution of algorithm and the data in the network changes.

Objective: Following are the objective of our research:

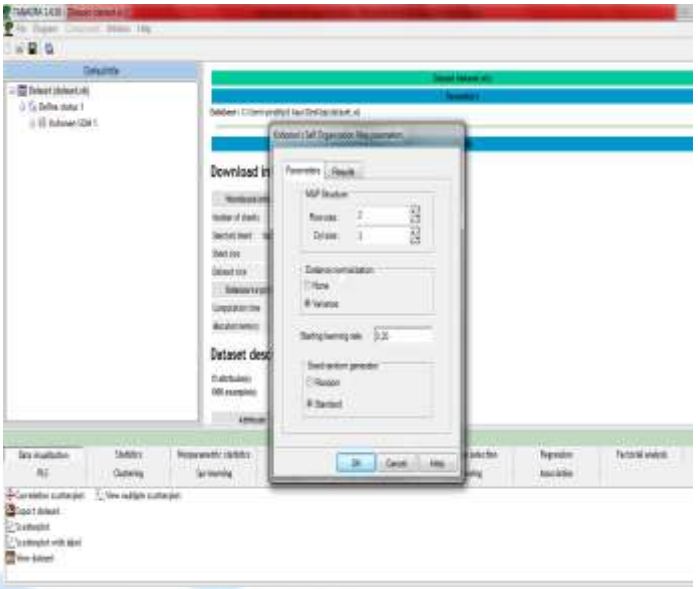
1. To evaluate the performance of clustering algorithm.

2. To analyze the banking data by applying clustering algorithms on it.

3. To find best possible solution for handling large amount of data.

Dataset: We analyze the Banking dataset. This is a real dataset. I have done a lot of surveys for finding the appropriate dataset according to my requirement. . In our research work, we will be focusing on performance and evaluation of clustering algorithms. There are many clustering algorithms in data mining but we will focus mainly on K-means, SOM and HAC. Data contain 1001 entries. We have adopted the hybrid approach of k-means, HAC and SOM.

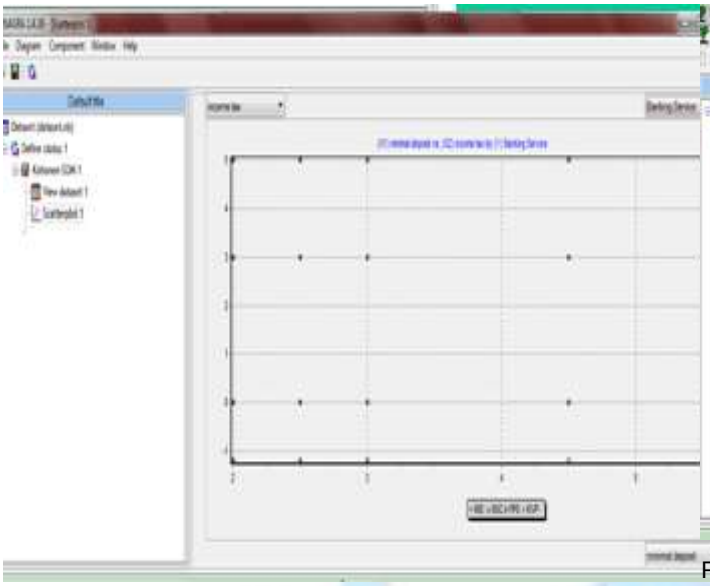
Tool Used: I have used the open source tool name Tanagra. Tanagra is very powerful tool which contain supervised learning as well as other paradigms like clustering, factorial analysis etc. In this project, I will apply K-Means, SOM and HAC using Tanagra tool and find the efficiency and performance of each algorithm and find out that using Tanagra tool which algorithm is able to handle with large amount of high dimensional data.



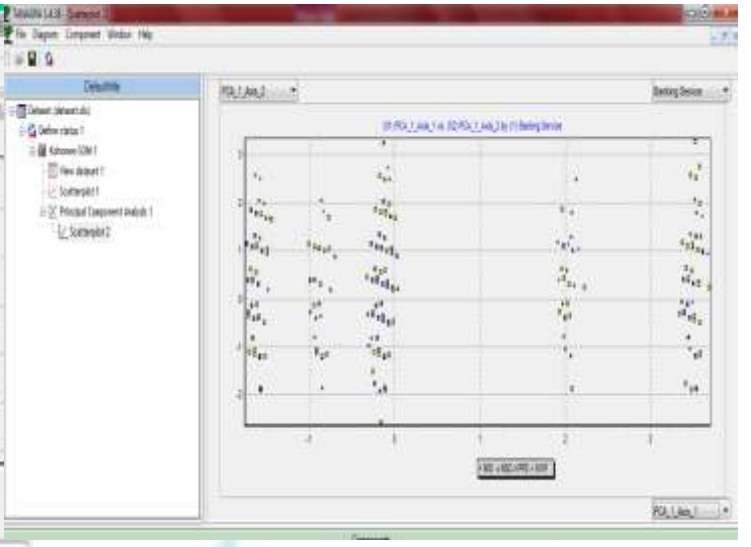
Firstly I have applied the Kohonen SOM algorithm on Dataset. I have also changed the parameters like rows are having size 2 and column are having size 3.

After applying SOM, we are getting the error ratio 0.6382.

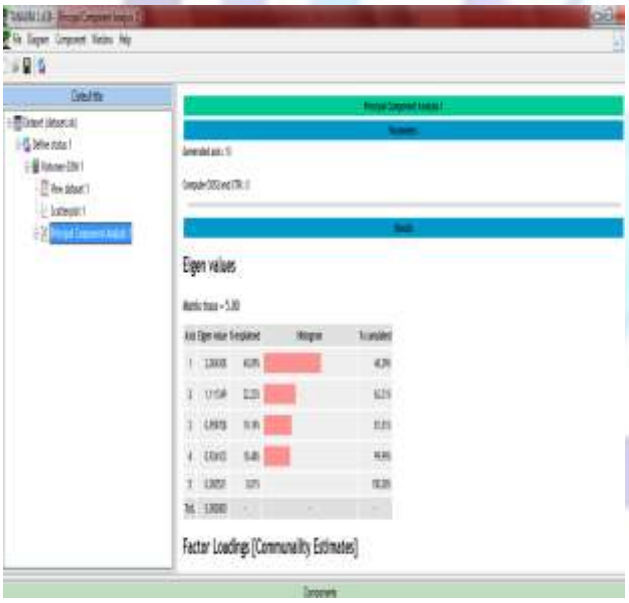
ISSN 22773061



Now I have clicked on the data visualization option and drag the scatter plot option and dropped on the Kohonen-SOM. When we click on this option, it will show the various services of banking in the form of dots, square, triangle having different colors, When we select the attribute say income tax on the y axis and attribute say minimal deposit on the y axis. As we see that dots which form clusters are not having clear view because we have high dimensional data.



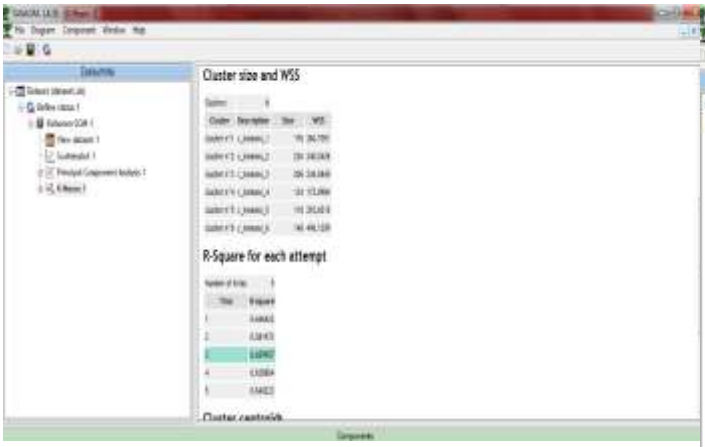
For the proper visualization of services, we have used the concept of PCA i.e. Principle Component Analysis. Above figure shows the clear view of various services present in the form of dots, square, triangle etc. Since we are having 6 no of clusters. When we choose axis 2 i.e. qualified for rebate on the y axis and axis 1 i.e. income tax on x axis, we can see the various services like MIS, NSC, PPS, KVP etc. lies in cluster1, cluster 2, cluster3, cluster 6, cluster 5 and some part of services lies in cluster 4.Since maximum number of KVP clusters are built in cluster no 3 and 1.In rest of the clusters, maximum number of PPS lies.PPS service is having the majority.

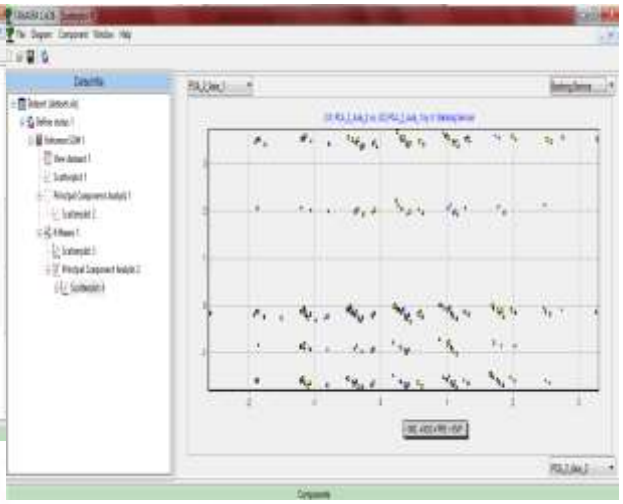


Above figure shows the Eigen values for each axis. Since we have 5 axis.

When we will drag K-means on Kohonen-SOM, then we will see the above results. In this case R-Square is 0.6600 which seems to be very high as compare to Kohonen-SOM. Within sum of square is 1700.2125 and total sum of square is 5000.000

ISSN 22773061



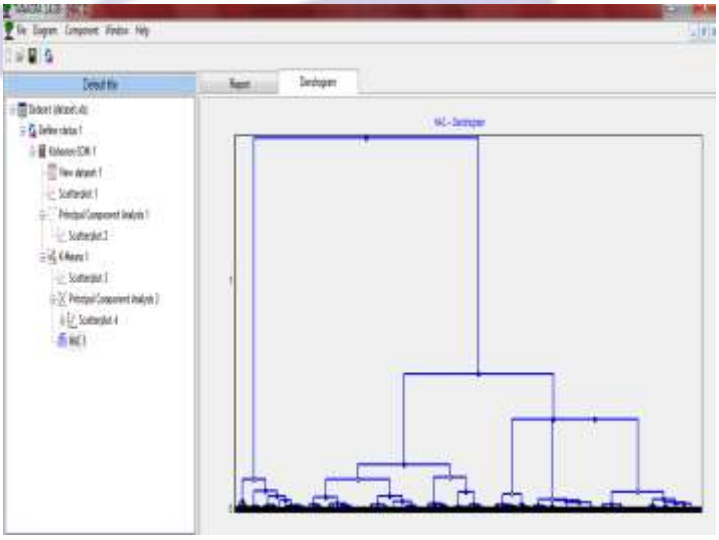


After applying PCA, now we have shown the scatterplot between various axis. When we choose axis 1 i.e. income tax on the y axis and axis 2 i.e. qualified for rebate on the x axis, we see that above results.

Above figure shows the cluster centroids for each attribute in each cluster.

In above figure we have dragged the HAC and drop on K-means. Above shows the report of HAC and user defined number of clusters i.e. 6.

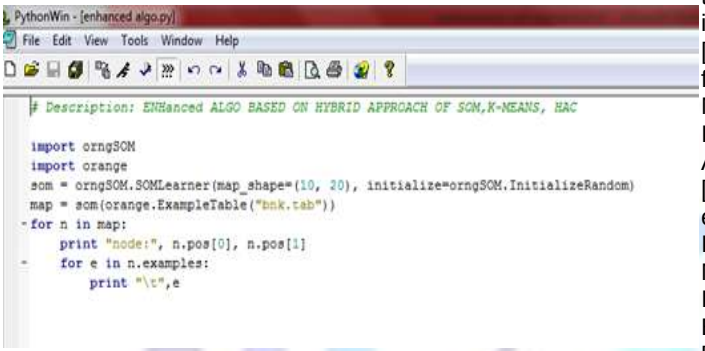
Above view shows the Eigen values for 5 axis and its histogram is also shown in the view



Above figure shows the dendrogram of HAC.

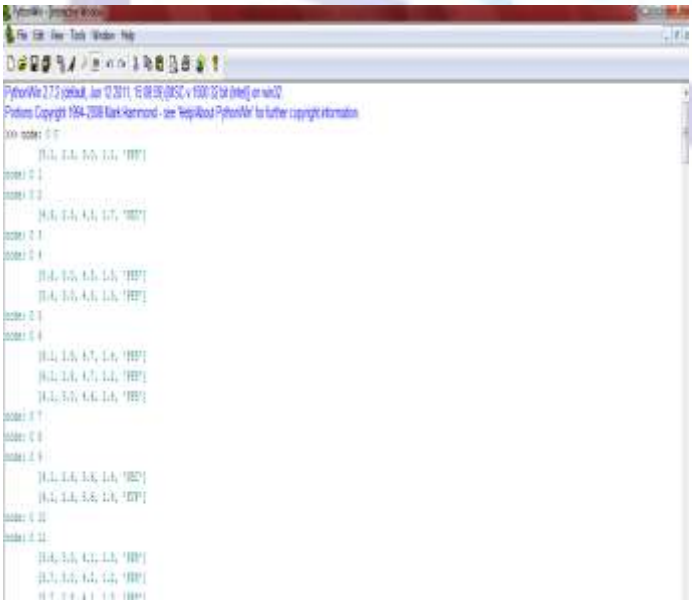
Conclusion:-From above work, we can clearly see that the results Coming using K-means are not able to handle large amount of data. Its error ratio is also very large, as in this case we mainly deal with high dimension data. SOM is helpful for handling with large data and also used for pattern recognition, image processing etc.

Proposed work: I have proposed one enhanced algorithm which will give the better results and visualization of nodes as results in Tanagra was not clear as the end of hybrid approach.



I have generated one code in Orange Python based on the hybrid approach of SOM, HAC & K-means.

We have browsed the script file name and press ok button



Now the code is running, we are seeing the nodes having different attribute values and services.

Future work: I will make one optimal algorithm which will overcome the drawbacks of SOM, HAC and K-means. I will generate one program in Python which will give the outputs in the form of clusters.

References:

[1].Performance evaluation of some clustering algorithms and validity indices, Ujjwal Maulik, Member, IEEE,Sanghamitra Bandyopadhyay, Member,IEEE,IEEE transactions on pattern analysis and machine intelligence.VOL.24,NO.12 DECEMBER 2002.

[2].A supervised clustering and classification algorithm for mining data with mixed variables, Xian yang Li, Member, IEEE, and Nong Ye, Senior Member, IEEE, IEEE transactions on system, Man and cybernetics-Part A: System and Humans, Vol. 36, No 2, MARCH 2006.

[3].An efficient concept based mining model for enhancing text clustering, Shady Shehata, Member, IEEE, Fakhri Karray, Senior Member, IEEE, and Mohamed S. Kamel, Fellow, IEEE,
IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 22, NO. 10, OCTOBER 2010

[4]. Approximate Distributed K-Means Clustering over a Peer-to-Peer Network Souptik Datta, Chris R. Giannella, and Hillol Kargupta, Senior Member, IEEE, IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 21, NO. 10, OCTOBER 2009.

[5]. A Fast Clustering Algorithm to Cluster Very Large Categorical Data Sets in Data Mining, Zhexue Huang, Cooperative Research Centre for Advanced Computational Systems, CSIRO Mathematical and Information Sciences, GPO Box 664, Canberra 2601, AUSTRALIA.

[6]. Clustering of the Self-Organizing Map, Juha Vesanto and Esa Alhoniemi, Student Member, IEEE, IEEE TRANSACTIONS ON NEURAL NETWORKS, VOL. 11, NO. 3, MAY 2000

[7]. Data Mining on DNA Sequences of Hepatitis B Virus, Kwong-Sak Leung, Kin Hong Lee, Jin-Feng Wang, Eddie Y.T. Ng, Henry L.Y. Chan, Stephen K.W. Tsui, Tony S.K. Mok, Pete Chi-Hang Tse, and Joseph Jao-Yiu Sung, IEEE/ACM TRANSACTIONS ONCOMPUTATIONAL BIOLOGY AND BIOINFORMATICS, VOL. 8, NO. 2, MARCH/APRIL 2011.