ISSN: 0975-833X

Available online at http://www.journalcra.com

International Journal of Current Research
Vol. 9, Issue, 05, pp.50307-50309, May, 2017

INTERNATIONAL JOURNAL OF CURRENT RESEARCH

RESEARCH ARTICLE

COMPARISON OF CLUSTERING TECHNIQUES: A SURVEY

[1]Rajat Bhatia, [2]Amrutha, R., *[,3]Shubham Singh, [4]Shreya Mehta and [5]BhushanInje

[1,2,3,4]B.Tech. Department of Information Technology, MPSTME SVKM's NMIMS Shirpur, Maharashtra, India
[5]Assistant Professor, Department of Computer Engineering, MPSTME SVKM's NMIMS Shirpur, Maharashtra, India

ARTICLE INFO

Article History:
Received 27th February, 2017
Received in revised form
04th March, 2017
Accepted 08th April, 2017
Published online 23rd May, 2017

Key words:
KDD, Clustering, k-means,
Hierarchical, Density based, EM.

ABSTRACT

The Information Industry has a huge amount of data available with them. This data needs to be analyzed and converted into useful information. The method used for extracting pattern and getting insight from data is known as data mining. Our Survey is focused on huge amount of data generated by the telecommunication industry. The marketing strategy used in telecommunication industry has various step, the very beginning of step is to segment the customers on the basis of customer's usage of services and customer billing cycle. For segmenting this data, here we are comparing different clustering algorithms to find out the best fit for our data and perform the further actions.

Citation: Rajat Bhatia, Amrutha, R., Shubham Singh, Shreya Mehta and BhushanInje, 2017. "Comparison of clustering techniques: A survey", *International Journal of Current Research*, 9, (05), 50307-50309.

INTRODUCTION

The amount of data available in any industry is increasing day by day. This data is made of good use by the companies to find out useful customer patterns and improve their business. Data mining is the technique used to do this study. As proposed by Jiawei Han, data mining is known for another popularly used term, knowledge discovery from data (KDD),many a time people get confused data mining with KDD while data mining is just an essential step in the process of knowledge discovery. Association rule, classification, clustering, outlier analysis, etc are some of the known methods of data mining. One of the important data mining technique is segmentation. Segmentation is the process through which we try simplify the complexity faced with various individualcustomers, each with unlike needs and having potential value (Hugh Wilson *et al*., 2002). Traditional customer were segmented with help of method such as experiential classification methods or simple statistical methods. The segmentation process can help industry to determine the customer behaviour and buying patterns for a particular set of customers. An industry provides its customers with a lot of packages and offers. A customer can choose from a variety of them. We are focusing on the telecommunication industry. The various services provided by this industry are data plans, voice calling, message packs, night calling plans, international calling, roaming etc.

Corresponding author:* **Shubham Singh,
B.Tech. Department of Information Technology, MPSTME SVKM's NMIMS Shirpur, Maharashtra, India.

The goal of this survey paper is to compare the following clustering techniques:

- K-means Algorithm
- Hierarchical clustering Algorithm
- Density Based Algorithm
- Expectation Maximization Clustering Algorithm

Data mining techniques

A. *K-Means Algorithm*

It is the algorithms that can solve the clustering problem that is the k-means algorithm. K letter in k-means represents the number of clusters required as per user . Each value present in the data set is treated as an object that has some location in the space. K-means finds partitions such as objects in a cluster are nearer to each other (as close as possible) and away from objects belonging to other clusters (distance maximum) (Jayant Tikmani *et al*., 2015). K-Means clustering algorithm works on the idea that the initial centers are given. Looking for the final clusters or centers starts from these initial centers. K points from the dataset as the initial cluster centre, putting the sample to the class where the nearest cluster center in. Then, the distances of all data elements are calculated by distances formula, e.g., Euclidean distance, Manhattan, Cosine, Chebychev, Minkowski, Tanimoto, etc (Purnawansyah and Haviluddin, 2016).

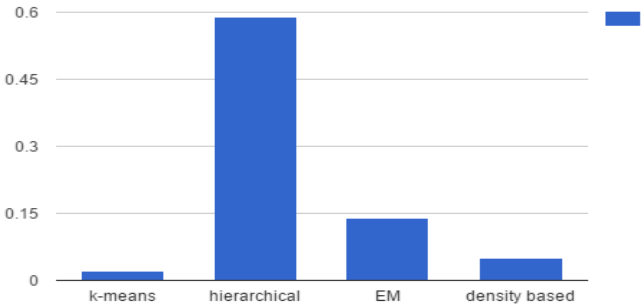
B. *Hierarchical Clustering Algorithm*

One of the important fact about H.C.Algo is the formation of tree like structure known as dendrograms, which are used by various application nowadays. Trees in hierarchical clustering algorithm represent data abstracted at various different levels. The consistency with which clustering solutions provide different levels of granularity allows partitions of various granularity that is being extracted as when data analysis process is carried out, that's the reason that they are used for interactive exploration and visualization (Ying Zhao and George Karypis, 2005). Agglomerative and divisive hierarchical clustering are two of the classified method for interactive exploration and visualization .Bottom-up or top-down fashion have prominent effect in classification done by hierarchical decomposition (YogitaRani and HarishRohil, 2013).

C. *Density Based Algorithm*

Density based algorithm is a type of partition clustering. Low density region to high density region are the two partitions created by density based clustering. Here a cluster is formed as when components that are densely connected hence can grow in random direction. One of the important reason that density based algorithms are able to discover clusters of custom shapes and provides protection to outliers. (Vivek S Ware and Bharathi, 2013) It has a need of only one input parameter and supports the user for determining an appropriate value for parameter. DBSCAN is useful even for large spatial databases (Martin Ester *et al*.,).

Algorithm	Number of Clusters	Cluster instance	Number of Iteration	Time	Accuracy
K-means	2	210(35%) 390(65%)	6	0.02s	55.20%
Hierarchical Algorithm	2	599(100%) 1(%)	4	0.59s	54.16%
EM Algorithm	2	365(61%) 235(39%)	4	0.14s	53.83%
Density based Algorithm	2	213(35%) 387(65%)	6	0.05s	50.83%



X-axis : algorithms
Y-axis : time (s)

D. *Expectation Maximization Clustering Algorithm*

EM is a clustering method that is model based. It's very first task is optimizes coordination between the data and functional mathematical model. They are very much rely upon the fact that the data generated is combination of defined probability distributions. It can be seen as a modified version of the Lloyd's algorithm. It is very much repetitive in nature and it's pretty much sure to get maximized solution (BhagyashreeUmale and Nilav, 2014).

Following are the reasons to choose EM based Algorithm (Osama Abu Abbas, 2008):

- It consist of strong statistical basis.
- Database size is linear.
- It shows robustness towards noisy data.
- It could take n number of cluster as input.
- Capable of handling high dimensional.
- If initialization is good it's handling is fast.

Comparision

Raj Bala *et al*. in the paper "A Comparative Analysis of Clustering Algorithms" uses a 'Bank' dataset which was downloaded from web.thecomparison of algorithms is as follows (RajBala *et al*., 2014):

Conclusion

The parameters accuracy and time are used to compare the algorithms. It is observed that the hierarchical algorithms takes more time to form clusters and is less accurate than k-means but more accurate than the other two algorithms. K-means on other hand takes the least time and is the most accurate algorithm to perform data clustering.

REFERENCES

Abbas, W.F.., N.D. Ahmad, NurlinaBintiZaini, 2013. "Discovering Purchasing Pattern of Sport Items Using Market Basket Analysis," International Conference on Advanced Computer Science Applications and Technologies.

BhagyashreeUmale and NilavM., 2014. "Overview of K-means and Expectation Maximization" International Conference on Quality Up-gradation in Engineering, Science and Technology, Algorithm for Document Clustering

Chan, Chu-Chai Henry. 2005. Online auction customer segmentation using a neural network model. *International Journal of Applied Science and Engineering*, pp. 101–109.

Divya D. Nimbalkar, 2013. "Data mining using RFM Analysis"*International Journal of Scientific & Engineering Research*, December, Volume 4, Issue 12

Hugh Wilson 1, Elizabeth Daniel and Malcolm McDonald, 2002. Factors for Success in Customer Relationship Management (CRM) Systems, *Journal of Marketing Management*, 18, 193- 219

Insani, R. and H. Laksmiwati, 2016. "Business Intelligence for Profiling of Telecommunication Customer," 2nd Asia Pacific Conference on Advanced Research, *APJCECT*.

Insani, R. and H. Laksmiwati, 2016. "Data Mining for Marketing in Telecommunication Industry," IEEE Region 10 Symposium (TENSYMP), Bali, Indonesia.

Jayant Tikmani, Sudhanshu Tiwari, Sujata Khedkar, 2015. "An Approach to Customer Classification using k-means", International Journal of Innovative Research in Computer and Communication Engineering.

Liu Jiale and Du Huiying, 2010. "Study on Airline Customer Value Evaluation Based on RFM Model," *International Conference On Computer Design And Applications*, Vol.4

Martin Ester, Hans-Peter Kriegel, Jiirg Sander, Xiaowei Xu, "A Density-Based Algorithm for Discovering Clusters" AAAI

MohsinNadaf and VidyaKadam, 2013. "DataMiningin Telecommunication", *International Journal on Advanced Computer Theory and Engineering*, Volume-2, Issue-3

Osama Abu Abbas, 2008. "Comparison between data clustering Algorithms", *The International Arab Journal of Information Technology*, Vol. 5, No.3

Purnawansyah, Haviluddin, 2016. "K-Means Clustering Implementation in Network Traffic Activities", *International Conference on Computational Intelligence and Cybernetics*.

RajBala, Sunil Sikka, Juhi Singh, 2014. "A Comparative Analysis of Clustering Algorithms", *International Journal of Computer Applications*.

VasilisAggelis, Dimitris Christodoulakis, 2005. "Customer clustering using RFM analysis", ICCOMP'05 Proceedings of the 9th WSEAS International Conference on Computers.

Vivek S Ware and Bharathi H N. 2013. "Study of Density based Algorithms", *International Journal of Computer Applications*.

XiongWeiwen, Chen Liang, Zhang Zhiyong, QiuZhuqiang, 2008. "RFM Value and Grey Relation Based Customer Segmentation Model in the Logistics Market Segmentation," International Conference on Computer Science and Software Engineering.

Ying Zhao and George Karypis, 2005. "Hierarchical Clustering Algorithms for Document Datasets", Springer Science + Business Media, Inc. Manufactured in The Netherlands.

YogitaRani and HarishRohil, 2013. "A Study of Hierarchical Clustering Algorithm", *International Journal of Information and Computation Technology*, Vol. 3

Yun Chen, Guozheng Zhang, Dengfeng Hu, Shanshan Wang, 2006. "Customer Segmentation In Customer Relationship Management Based On Data Mining," Intemational Federation for Information Processing, Boston: Springer

Zhang Yihua, 2010. "Vip Customer Segmentation Based on Data Mining in Mobile-communication Industry". The 5th International Conference on Computer Science & Education.

****