IJARCCE

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

A Review on Clustering Techniques and Time Series Analysis

Nidhi Tiwari[1], Prof. Toran Verma[2]

M. Tech Scholar, Rungta College of Engineering & Technology, Bhilai (C.G.) [1]

Assistant Professor, Rungta College of Engineering & Technology, Bhilai (C.G.)[2]

Abstract: Data mining is the combination of data assembled by customary information mining philosophies and procedures with data accumulated over a period of time. Mining means extricating something helpful or important from an already existing data. This paper is intended to review time series analysis and clustering techniques. The clustering algorithms which are investigated are K-Means and Hierarchical Clustering. Also a review is also conducted on Time Series based clustering.

Keywords: Data Mining, K-Means Algorithm, Hierarchical clustering algorithm, Time Series analysis

I. INTRODUCTION

Data mining, the extraction of concealed prescient data from expansive databases, is a compelling new innovation with incredible potential to help organizations concentrate on the most essential data in their information stockrooms. Most organizations effectively gather and refine enormous amounts of information. Information mining strategies can be actualized quickly on existing programming and equipment stages to improve the benefit of existing data assets, and can be incorporated with new items and frameworks as they are brought on the web. With the dangerous development of data sources accessible on the World Wide Web, it has gotten to be progressively vital for clients to use mechanized instruments in discover the coveted data assets, and to track and dissect their use designs. These factors give rise to the necessity of creating server side and client side intelligent systems that can effectively mine for knowledge [1].

Cluster is a group of objects that belongs to the same class. In other words, similar objects are grouped in one cluster and dissimilar objects are grouped in another cluster [2]. Clustering is the process of making a group of abstract objects into classes of similar objects. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group .The group is called a cluster.

Major attributes of clustering are:-

- Scalability –Highly scalable clustering algorithms are needed to deal with large databases.
- Ability to deal with different kinds of attributes − Algorithms should be capable to be applied on any kind of data such as numerical data, categoricaldata etc.
- Discovery of clusters with attribute shape − Algorithm should be capable of detecting clusters of different shapes.
- Ability to deal with noisy data – Databases may contain noisy, missing or erroneous data. Algorithmshould not be sensitive to such data.
- Interpretability -Results of clustering should be interpretable, comprehensible, and usable.

Major applications of cluster analysis are:-

- Clustering analysis is broadly used in many applications such as market research, pattern recognition, data analysis, and image processing.
- Clustering can also help marketers discover distinct groups in their customer base. And they can characterize their customer groups based on the purchasing patterns.
- In the field of biology, it can be used to derive plant and animal taxonomies, categorize genes with similar functionalities and gain insight into structures inherent to populations.
- Clustering is also used in outlier detection applications such as detection of credit card fraud.

The agglomerative hierarchical clustering creates the clusters by considering each item or product as an individual cluster from its starting with which the retailers can easily identify sale trends. [3]. The knowledge of what a customer or a group of customers is going to purchase can be very useful for the retailers [4]. These results could also be helpful in determining which products appeal each other so that they can be put together in a market in order to increase the

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

sales. Time-series data occur naturally in many online applications, and the logging rate has increased greatly with the progress made on hardware and storage technology [5].A technique called AIRMA is proposedwhich is a time series based statistical model for predicting stock market behavior.[6].

II. K-MEANS ALGORITHM

The K-Means algorithm is a simple yet effective statistical clustering technique. Basic steps are:

- Choose a value for K, for determine no of clusters.
- Choose K data points) from dataset at random. These are the initial cluster centres.
- Use simple Euclidean distance to assign the remaining instances to their closest cluster centre.
- Use the instances in each cluster to calculate a new mean for each cluster.

If the new mean values are identical to the mean values of the previous iteration the process terminates. Otherwise, use the new means as cluster centres and repeat steps 3-5 [7].

III. HEIRARCHIAL CLUSTERING ALGORITHM

This method creates a hierarchical decomposition of the given set of data objects. We can classify hierarchical methods on the basis of how the hierarchical decomposition is formed [8]. There are two approaches are:-

- Agglomerative Approach
- Divisive Approach

Agglomerative Approach: This approach is also known as the bottom-up approach. It starts with each object forming a separate group. It keeps on merging the objects or groups that are close to one another. It keeps on doing so until all of the groups are merged into one or until the termination condition holds.
Divisive Approach: This approach is also known as the top-down approach. It starts with all of the objects in the same cluster. In the continuous iteration, a cluster is split up into smaller clusters. It is down until each object in one cluster or the termination condition holds. This method is rigid, i.e., once a merging or splitting is done, it can never be undone.

Approaches to Improve Quality of Hierarchical Clustering are:

- Perform careful analysis of object linkages at each hierarchical partitioning.
- Integrate hierarchical agglomeration by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro-clusters.

IV. TIME SERIES ANALYSIS

A time series is a series of data points indexed in time order. Most commonly, a time series is a sequence taken at successive equally spaced points in time. Thus it is a sequence of discrete-timedata [5]. Time series analysis comprises methods for analyzing time series data in order to extract meaningful statistics and other characteristics of the data. Time series forecasting is the use of a model to predict future values based on previously observed values. While regression analysis is often employed in such a way as to test theories that the current values of one or more independent time series affect the current value of another time series, this type of analysis of time series is not called "time series analysis"

V. CONCLUSION

Data mining is emerging technology dealing with major issues such as security and scalability and efficiency. This paper focused on studying various data mining algorithms like K-Means and Hierarchical Clustering. Also a brief overview was also discussed about time series analysis. Future work is intended to cluster data on basis of Time Series using K-Means and Hierarchical clustering methods and review its execution efficiency.

REFERENCES

[1] G.K. Gupta,Introduction to data mining with case studies:Prentics Hall of India, New Delhi, 2006
[2] Clustering and its Applications, L.V. Bijuraj, Proceedings of National Conference on New Horizons in IT - NCNHIT 2013.
[3] Market-Basket Analysis using Agglomerative Hierarchical approach for clustering a retail items ,RujataSaraf&SonalPatil,IJCSNS International Journal of Computer Science and Network Security,2016.
[4] Market Basket Analysis using Association Rule Learning ,NidhiMaheshwari, Nikhilendra K. Pandey&PankajAgarwal,International Journal of Computer Applications ,2016

ISSN (Online) 2278-1021
ISSN (Print) 2319 5940

International Journal of Advanced Research in Computer and Communication Engineering
ISO 3297:2007 Certified
Vol. 6, Issue 5, May 2017

[5] Mining Big Time-series Data on the Web,Y asushi Sakurai, Yasuko Matsubara &Christos Faloutsos, WWW 2016 Companion, April 11–15, 2016
[6] Time Series Data Analysis for Stock Market Prediction using Data Mining Techniques with R,Mahantesh C. Angadi, Amogh P. Kulkarni,International Journal of Advanced Research in Computer Science,2015
[7] On Clustering Time Series Using Euclidean Distance and Pearson Correlation, Michael R. Berthold & Frank H□oppner,arXiv:1601.02213v1 [cs.LG] , 2016
[8] Hierarchical Clustering Algorithms for Document Dataset, YING ZHAO & GEORGE KARYPIS, Data Mining and Knowledge Discovery, 10, 141–168, 2005
[9] Fast Algorithms for Mining Association Rules: RakeshAgrawalRamakrishnanSrikant VLDB Conference Santiago, Chile, 1994
[10] A Review of various k-Nearest Neighbor Query Processing Techniques :International Journal of Computer Applications (0975 – 8887) Volume 31–No.7, October 2011
[11] Fast Algorithms for Mining Association Rules: RakeshAgrawalRamakrishnanSrikant VLDB Conference Santiago, Chile, 1994
[12] High Performance Data Mining Using the Nearest Neighbor Join Christian Böhm Florian Krebs
[13] Top 10 algorithms in data mining, Xindong Wu, Springer-2007
[14] Mining Association Rules between Sets of Items in Large Databases:RakeshAgrawal ,Tomasz Imielinski,ArunSwamiACM SIGMOD ConferenceWashington DC, USA, May 1993
[15] Han, David, et al. Principles of Data Mining: MIT press. Cambridge, 2001.
[16] High Performance Data Mining Using the Nearest Neighbor Join Christian Böhm Florian Krebs.

DOI10.17148/IJARCCE.2017.6542