DOI: 10.21276/sjet

Scholars Journal of Engineering and Technology (SJET) **ISSN 2321-435X (Online)**
Sch. J. Eng. Tech., 2017; 5(7):355-358 **ISSN 2347-9523 (Print)**

(An International Publisher for Academic and Scientific Resources)
www.saspublisher.com

Research Article

Time Series Based Clustering Using K-Means and Hierarchical Techniques for Efficient Mining of Sales Data

Nidhi Tiwari[1], Prof. Toran Verma[2]
[1]M.Tech Scholar, Rungta College of Engineering & Technology, Bhilai (C.G.)
[2]Assistant Professor, Rungta College of Engineering & Technology, Bhilai (C.G.)

***Corresponding author**
Nidhi Tiwari
Email: nidhi.tiwari0109@gmail.com

Abstract: Data mining is the combination of data assembled by customary information mining philosophies and procedures with data accumulated over large no of data sources. Time-Series clustering is one of the important concepts of data mining that is used to gain insight into the mechanism that generate the time-series and predicting the future values of the given time-series. Time series clustering is a task that aims to assign each of the time series a class label from two or more classes having some training data. Classification is needed to distinguish between different types of time series. K-Means is one of the common and simplest unsupervised algorithms for clustering. It classifies data on the basis of Euclidian distance. In data mining and statistics, hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Time Series based clustering is used to classify time series data by K-Means or Hierarchical clustering. The proposed work aims to extract important information from Time Series sales data for analysing current sale trends for efficient Market Basket Analysis. The work also aims to provide better analysis of the proposed approach on the basis of obtained numerical results.
Keywords: Data Mining, K-Means Algorithm, Hierarchical Clustering, Market Basket Analysis.

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.

Clustering of Time-Series data is the unsupervised classification of a set of unlabeled time series into groups or clusters where all the sequences grouped in the same cluster should be coherent or homogeneous. It can be shape-level if it is performed on the many individual time-series or structure level if it works on single long-length time-series. The major issues related to time-series clustering are high dimensionality, temporal order and noise. Time-series clustering is Temporal-Proximity-Based Clustering if it works directly on raw data either in frequency or time domain; Representation-Based Clustering if it works indirectly with the features extracted from the raw data and Model-Based if it works with model built from raw data [1].Time series prediction (forecasting) is used to predict future values of a continuous time series based on the past observations. For example, prediction of stock prices or retail values of goods for the next day taking into account their values in the past weeks. For time series prediction it is necessary to build a model that would describe the behavior of the observed variable over time [2].

RELATED WORK

A time-series is essentially classified as dynamic data because its feature values change as a function of time, which means that the value(s) of each point of a time- series is/are one or more observations that are made chronologically. Time-series data is a type of temporal data which is naturally high dimensional and large in data size. Time-series data are of interest due to their ubiquity in various areas ranging from science, engineering, business, finance, economics, healthcare, to government. While each time-series is consisting of a large number of data points it can also be seen as a single object [3]. Clustering such complex objects is particularly advantageous because it leads to discovery of interesting patterns in time-series datasets. An approach presents the Market-Basket Analysis using

Agglomerative ("Bottom-up") hierarchical approach for clustering retail items [4]. Agglomerative hierarchical clustering creates a hierarchy of clusters which are represented in a tree structure called a Dendorogram. In agglomerative hierarchical clustering, dendrograms are developed based on the concept of „distance" between the entities or, groups of entities. The approach presents low level simulation of customers and products of a Supermarket [5]. The simulation is parameterized by results of analysis of real sales receipt data to achieve the results close to reality. The analysis includes building of probabilistic models based on clustering and time series analysis. Along with results of analysis, the simulation approach also incorporates results of empiric analysis of customer psychology. Further, the user can set other parameters like composition of customer types, prices of products and bargain offers, locations of shelves and products, allowing investigation of different influences and dedicated simulation of products of interest.

PROBLEM IDENTIFICATION

The sales of regular products can be predicted by application of time series analysis or just by the experience of store managers. Causal interdependencies between the observed objects, namely the supermarket and customers, remain uninvest gated. Therefore, closer examination of the irregular behavior of offered bargain products along with simulations on a lower level of single objects (customers, products, shopping carts, and sales receipts) may provide more accuracy and the possibility of experimenting with changing the circumstances. As a result, simulations on a detailed enough level could allow insight into the reasons for sale changes and interdependencies between customers and the supermarket which can be evaluated by analyzing and forecasting of simulated event logs and sales receipt data. In addition, selected bargain offer products can be analyzed in greater detail.

Lots of human resources are required to make such arrangement with which the customers can easily get whatever they want more efficiently. The products are arranged in such a way that the items which are purchased together are placed in one shelf beside to each other and by providing the different loyalty schemes on such a items, the total sales of the product have been increased. But while doing so it's very difficult to predict which products should be kept beside to each other and in which product the customer will show their interest. For this purpose it is necessary to find out which products are frequently purchased by customers from the total sale and by using this we can easily achieve the market-basket analysis by placing the most frequently purchased items besides to those items which are necessary along with the purchased item but not compulsory to purchase. By considering this scenario, it is necessary to group all such a items which can ultimately increase the total sale of the product.

PROPOSED METHODOLOGY

K-Means Algorithm

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

1. Choose a value for K, for determining no of clusters.
2. Choose K data points) from dataset at random. These are the initial cluster centers.
3. Use simple Euclidean distance to assign the remaining instances to their closest cluster center.
4. Use the instances in each cluster to calculate a new mean for each cluster.
5. 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 centers and repeat steps 3-5.

Hierarchical Clustering Algorithm

1. Calculate the distance between all objects. Store the results in a distance matrix.
2. Search through the distance matrix and find the two most similar clusters /objects.
3. Join the two clusters/objects to produce a cluster that now has at least 2 objects.
4. Update the matrix by calculating the distances between this new cluster and all other clusters.
5. Repeat step 2 until all cases are in one cluster.

Proposed Approach

A time series is a sequence of real numbers that represent the measurements of a real variable at equal time intervals, whereas a time series database is a collection of time series. Time sequences appear in many applications, to be more precise, in any applications that involve a value that changes over time. There are several important aspects of mining time series that include trend analysis, similarity search and

mining of sequential and periodic patterns in time related data. The clustering performed is whole clustering where clustering is similar to that of conventional clustering of discrete objects. Provided there is a set of individual time series data, the purpose is to classify similar time series into the same cluster. Basic steps are:-

1. Filter the sales dataset into Time series Based data.
2. Work out an appropriate distance/similarity metric.
3. Use existing clustering techniques, such as k-means, hierarchical clustering, density-based clustering or subspace clustering, to find clustering structures.
4. Analyze cluster structures identify the sale patterns for particular structure for identifying sales trends.

RESULTS AND DISCUSSION

The goal of experiment conducted is to efficiently analyze the sale trends by filtering out clusters on the basis of Time series data. The sale dataset for the Experiment is Sample data that appears in the December Tableau User Group presentation.

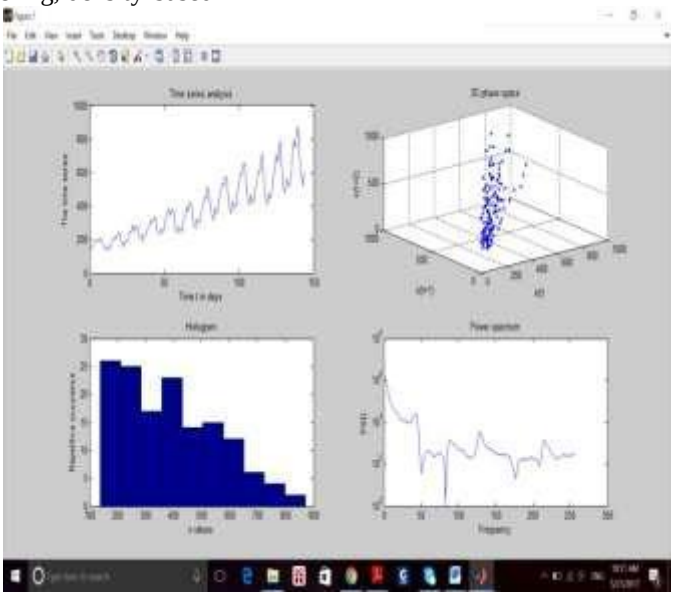

Fig 1: Time Analysis

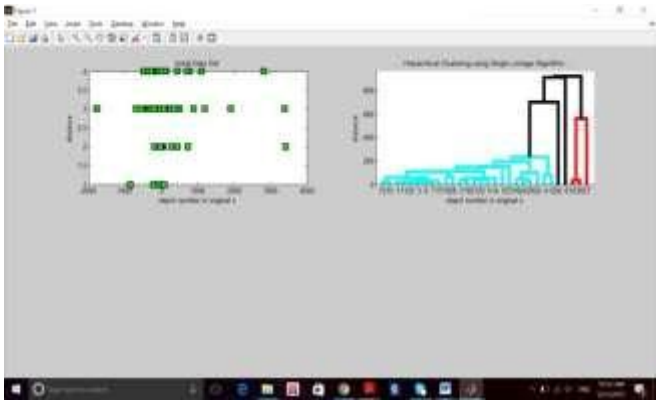

Fig 2: Initial Dataset and Hierarchal Clustering

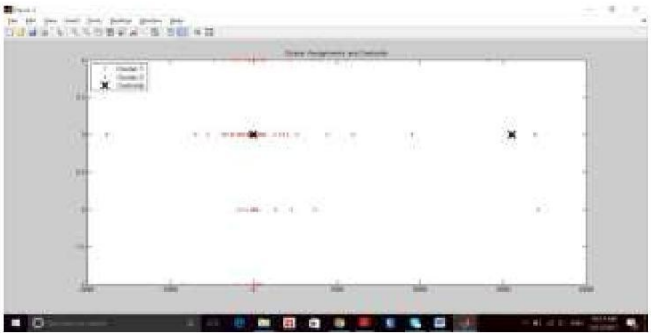

Fig 3: K-Means Clustering

The experiment conduct on sample super store data shows the behavior patterns overtime according to Time Series analysis. The clustering performed on the data is used effectively for monitoring sale trends overtime.

CONCLUSION

Data mining is emerging technologies dealing with major efficiency issues. Time series analysis is a well-known task; however, currently research activities are being carried out with the purpose to try to use clustering for the intentions of time series analysis. The main motivation for representing a time series in the form of clusters is to better represent the main characteristics of the data. Future work aims to increase the efficiency by using better parameters of other algorithms also. It is not possible to develop a system that makes all the requirements of the user. User requirements keep changing as the system is being used. Some of the future enhancements that can be done to this system are:

- As the technology emerges, it is possible to upgrade the system and can be adaptable to desired environment.
- Because it is based on object-oriented design, any further changes can be easily adaptable.
- The efficiency of algorithm can be further increased by applying more efficient data mining algorithms in near future. More work is possible on security of data in cloud servers.

REFERENCES

1. Rani S, Sikka G. Recent techniques of clustering of time series data: a survey. International Journal of Computer Applications. 2012 Jan 1;52(15).
2. http://thescipub.com/abstract/10.3844/jcssp.2014.23 58.2359.
3. Time-seriesclustering – A decade review, Saeed Aghabozorgi, Ali Seyed Shirkhorshidi & TehYingWah, ELSEVIER, Information Systems 53 (2015) 16–38
4. Saraf R, Patil S. Market-Basket Analysis using Agglomerative Hierarchical approach for clustering a retail items. International Journal of Computer Science and Network Security (IJCSNS). 2016 Mar 1; 16(3):47.
5. Schwenke C, Ziegenbalg J, Kabitzsch K, Vasyutynskyy V. Simulation based forecast of supermarket sales. InEmerging Technologies & Factory Automation (ETFA), 2012 IEEE 17th Conference on 2012 Sep 17 (pp. 1-8). IEEE.
6. Kusrini K. Grouping of Retail Items by Using K-Means Clustering. Procedia Computer Science. 2015 Jan 1; 72:495-502.