International Journal of Marketing Management

ISSN 2454 - 5007

www.ijmm.net

Email ID: editor@ijmm.net , ijmm.editor9@gmail.com

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 9 Issuse. 2May 2021

Mining Stream Data using k-Means clustering Algorithm

1Medi Manishankar, 2Dr. K. Venkateswara Rao

Abstract- Time-stamped sequences of data comprise what is known as stream data. Sources with fluctuating update rates and high dimensionality may be many and varied. There are instances when it is not feasible to process all of the data in a stream, and storage is also a common problem. Random sampling, sliding windows, and histograms are just a few of the ways stream data may be processed. Analysis of traffic, telecommunications, and stock market data may all be done using stream datasets. Stream data analysis often makes use of data mining methods including association analysis, classification, and clustering. The k-means clustering technique is employed in this study to mine data from a city's road traffic stream. Sliding window method is used to manage the data in the stream. Python's visualisation tools are used to depict the clusters visually. People may follow the movement of traffic with the use of real-time updates to the clusters. This paper details the findings.

Keywords: Urban road traffic streaming data is clustered using the k-Means clustering algorithm.

1. INTRODUCTION

Massive amounts of information are being generated at breakneck speeds in the age of big data. Because data streams are so common these days, offline storage and querying are out of the question. Over 2.8ZB of data was produced and processed in 2012, according to the Digital Universe research [3], with a predicted rise of 15 times by 2020. The proliferation of sensors in our immediate surroundings is to blame for this increase in digital data output. The sensors collect data and transmit it in a variety of formats, which necessitates online analysis.There is no need to keep vast amounts of data in some type of memory when dealing with a data stream, which has a constant flow of data with a time stamp [9]. A data stream rather than a dataset is the most appropriate term for many current applications [3]. Telecommunications and networking, stock trading, and other similar fields all produce streams of data.in-store, on-line, and so on Consider telecommunications; it is necessary to evaluate phone records, blogs and page views as they come in on a regular basis. Stream data is challenging to manage because of its chronologically organised, rapidly changing, and theoretically limitless properties, such as the fact that it arrives continually.As a result, processing stream data before it expires [3], [4] and completing the necessary analysis are the major challenges.

1 PG Scholar in CSE, CVRCE, Hyderabad, India,
manishankarmedi@gmail.com
2 Professor of CSE, CVRCE, Hyderabad, India,
kvenkat.cse@gmail.com

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 9 Issuse. 2May 2021

Because many scans aren't feasible with stream data, we must choose the optimum output method for a single scan. The amount of features for each record in certain big, multi-purpose systems like satellite-mounted remote sensors, real-time robotic applications, may lead to high dimensionality. The high end of the permissible value range for each property may be present in certain records.

2. New data structures, methodologies, and algorithms are required for efficient processing of stream data. When it comes to stream data storage, we frequently have to make tradeoffs between accuracy and storage capacity. To put it another way, we're usually happy to make do with approximations rather than precise answers. In many data stream-based algorithms, the answer is computed within a factor-e of the real answer with a very high degree of confidence. The space needs often rise in direct proportion to the approximation factor. Sliding windows, random sampling, and histograms are prominent data structures and approaches for stream data processing [1], [10].

3. he sliding window is the most viable stream data processing strategy because it just uses the most recent data to make judgments, rather than doing calculations on all of the data that has been gathered so far.. And it uses main memory to carry out calculations. As a consequence of these factors, memory utilisation is efficient and reduced, hence minimising the amount of storage required. The most difficult part of data stream mining is figuring out how to use supervised and unsupervised learning methods to uncover patterns and trends. Using unsupervised learning, clustering may lead to the identification of previously unknown data.

5.It is possible to describe clustering as the grouping of comparable things based on their characteristics (attributes). Objects should be clustered in such a manner that they are similar to one another inside the cluster and distinct to those in other clusters. This is how clustering should be done.

9.In this work, urban road traffic stream data is clustered. Our goal is to visualise and measure traffic density. Traffic officials can monitor the flow of vehicles in real time and take quick action to reduce gridlock. This is beneficial to both drivers and the city's traffic authority. Making it easier for individuals to see traffic patterns by updating real-time graphic graphs.

13. STREAM DATA PROCESSING

Stream data processing methods, applications, problems, and concerns are covered in the following areas. Additionally, Data Stream Management Systems were examined. Processes for Handling Streamed Data

In order to process stream data, there are a variety of methods available, such as random sampling [1], [10], [11], which involves taking a sample of data from the entire data stream and then processing it. Another method is reservoir sampling, which involves taking a sample of data from the entire data stream and then performing calculations on recent data [2], [9].

Rather of doing calculations on all of the past data, the sliding window approach relies on the most recent data to accomplish so [1]. There's always a fresh piece of info arriving at the same moment as t. The "expiration" time for this element is t + w, where w is the "size" of the window.

When only recent occurrences matter, the sliding window approach is an excellent choice for stocks or sensor networks. Because just a tiny portion of the data is kept, it uses less memory. Count-based and time-based sliding windows are two of the options in the sliding window. The amount of records to be processed in a single pass is fixed in the count-based sliding

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 9 Issuse. 2May 2021

window, and the time stamp sliding window specifies how long it will take to add new data to the sliding window. Because statistical summaries must be maintained in real-time as the stream develops, merging and dividing emerging clusters when clustering streaming data is extremely difficult [2].

Data Stream Processing Applications, Challenges, and Issues

In the recent days, many applications [4], [5] are producing massive amounts of data.

As an example, in the field of telecommunications and networks, a central database system keeps track of user calling records, web clicks, network monitoring, fraudulent calls, call drops, and so on. There is a lot of data that has to be examined in order to come up with a solution for the issue in real time.

There are a large number of buyers and sellers in the financial industry or stock market who trade shares for each other. As a result, real-time data analysis is necessary in order to educate the public about stock market patterns.

There are a large number of automobiles on the road in metropolitan areas. For example, vehicles equipped with high-tech sensors may help us gather stream data valuable for stream data analysis, which leads us to learn about dense places in real-time road traffic and how to modify signalling systems so that users can move more easily.

Dealing with data from streams, on the other hand, has its own set of difficulties and complications due to the fact that the data originates from surroundings that are both dynamic and high in terms of dimension. There may be blank spaces in certain records due to omissions. As a result, we're dealing with three major issues: quantity, velocity, and volatility. Preprocessing is essential in this case to prevent noisy data, redundant data, outlier reduction, and fill the missing values.

However, there are a few more issues [4] that need to be addressed.

- Uncertainty in the Data
- Treatment of data types

Validity of the clusters

Data has a lot of dimensionality.

Data Stream Management Systems (DSMS)

Data stream management is required to process stream data. Management solutions for continuous data streams are known as DataStream Management systems. It deals with the data in a data stream management system.

as a source of data, then uses a query processor to process it and store the results in databases.

Systems for managing data streams [2] [7] the DSMS process continues searches over enormous volumes of time-varying data streams, such as those generated by stack trade marketing, network traffic monitoring, sensor data analysis, real-time data warehouses, etc. Using the Continuous Query (CQ) technology, a user may register queries that indicate their specific interests in unbounded, live data sources.

Queries are continually evaluated by a query engine, which sends unrestricted streaming outputs to the right consumers. Sliding window connect between streams is a key operator in a CQ system.

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 9 Issuse. 2May 2021

The DSMS architecture is shown in the image below.
Fig 1 Data stream Management Systems (DSMS)

The following is a list of the DSMS's many components. Indexes and historical data are available in the local database. It is possible to store schema in a meta-database. Using main-memory indexes for efficient search, the local database keeps current data summaries and historical data in main memory.
The continuous query processor uses meta-data and the local database to optimise and execute queries over data streams. On-the-fly access to streaming and local data is possible with the use of continuous queries.

There are, however, significant scaling concerns with DSMS because of these constraints. That's what I mean.

The arrival of new components necessitates a rapid response. 2.

On the other hand, it is possible to process many data streams at once.

Low latency and high throughput are essential for a processing engine.

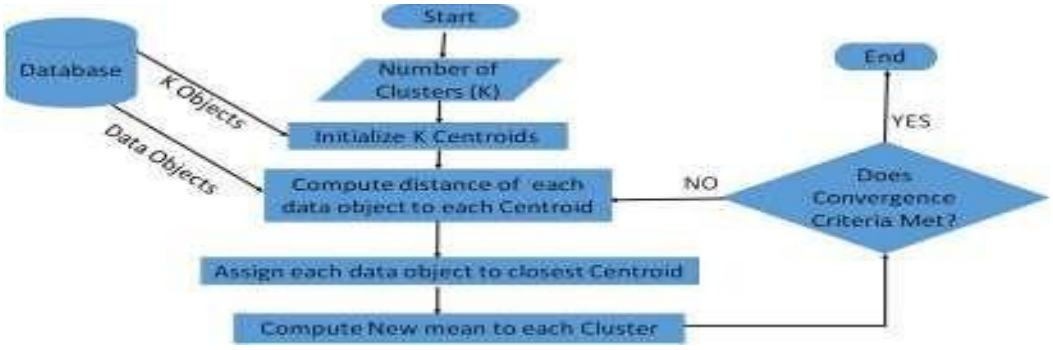
Sliding window joins may use a lot of memory to store tuples at high stream speeds and big window widths.

Clustering of data streams

This is the process of putting together a collection of related things. All the data items in a single cluster are comparable to one other, yet they are distinct from those in other clusters. Partitioning, hierarchical, density-based, grid-based, and model-based clustering techniques are all forms of clustering. Partitioning algorithms based on the Divide-and-Conquer technique are the most appropriate since data processing must be performed in a single scan (One scan Divide-and- Conquer approaches have been widely used to cluster data streams [6]).
K-Means and K-Medoids [1] [6] are widely used in data mining as partitioning-based approaches. Alkermann has suggested an online K-Means method that uses the merge reduction approach to keep a short sketch of the input. K-means is the best technique for numerical data because of its simplicity. In most cases, K-Means is more computationally efficient than hierarchical clustering when the number of variables is minimal. Stream Data Processing using the k-Means Clustering Algorithm

The major phases of the K-Means clustering method are outlined in the following sections.
Algorithm: k-Means Clustering Algorithm
Input: D={r1, r2, r3, …,rn} and K value Output: K clusters
Method:
Initialize mean values for K-clusters randomly;
Say m1, m2, m3, …, mk;
Repeat
Assign each record in D to the cluster based on similarity;
Calculate new mean for each cluster;
Until convergence criteria are met;

Following flow chart in Fig 3 describes the algorithm.



International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 9 Issuse. 2May 2021

Fig 3 k-Means clustering algorithm flow chart
While performing clustering of data with the k-
First, the 'k' value must be determined for the clustering procedure. The term "number of clusters" refers to this concept. Initialize the cluster centroid values with a random value after determining the k value. The fact that the place contains two coordinates in two-dimensional views makes it necessary to use two-coordinate values for mining urban road traffic. C is the cluster centroid value for each cluster (x, y). The term "centroid" refers to a location in the centre of a cluster of data.

First, it clusters the items by picking k at random from a pool, each of which represents a cluster mean or centre. For the remaining items, the object's distance from the cluster mean is used to determine which cluster it belongs to. Equation may be used to compute the distance (1).
the d (x,

i) (x .a j

m i .a j

Where,
mi..ajis is the mean of aji's jth attribute, which has n characteristics in total.

The items have been assigned to their respective clusters, and the next step is

The Ci(X, Y) value should be updated.

computing the average of the cluster's values. It's easy to figure out by

If Cx = (x1, x2, x3,..., xn)/n, then Cy = (x1, x2,..., x3,..., yn)/n, therefore Ci = (Cx, Cy)

Then, as discussed before, go through the process of allocating items to clusters once again by calculating distances.
Fig 4 Sample dataset
For processing stream data, we need stream data set to be loaded as input to the programme. The sample dataset is shown in Fig 4.

This process will be continues convergence criteria.

MID	LONGITTUDE	LATTITUDE	TIME	DATE
492	25.89693108	70.65074809	11:28:09	29-05-2018
435	23.10282438	70.63469947	18:06:02	12-03-2018
140	15.98019247	73.11980458	10:01:36	29-04-2018
7	18.23621351	72.23749498	19:32:36	26-05-2018
55	24.44035766	71.95261788	19:48:13	26-04-2018
59	10.24952979	68.19566075	13:48:16	22-04-2018
285	20.57568212	73.16153941	13:08:30	23-02-2018
307	14.88187968	72.57758794	11:50:33	12-02-2018
294	20.1282351	67.43114083	16:07:20	30-03-2018
149	27.41309995	71.30015784	18:02:49	19-03-2018
201	15.38221217	74.96741749	15:35:37	03-05-2018
153	16.90214131	73.73925305	16:56:28	15-05-2018
136	29.40657863	72.35874156	17:19:42	01-04-2018
174	24.03338616	69.70961455	14:29:58	21-04-2018
277	17.74858437	73.1857714	19:19:15	19-04-2018
443	16.2988262	74.66218119	15:13:34	05-02-2018
492	19.59172693	71.26790394	12:30:52	08-03-2018
330	26.79924097	65.38709842	10:18:56	25-03-2018

There are three methods to define convergence criteria.

Running a programme requires input from a dataset.

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 9 Issuse. 2May 2021

The mean clusters have not changed. The clusters have stabilised and are no longer moving.

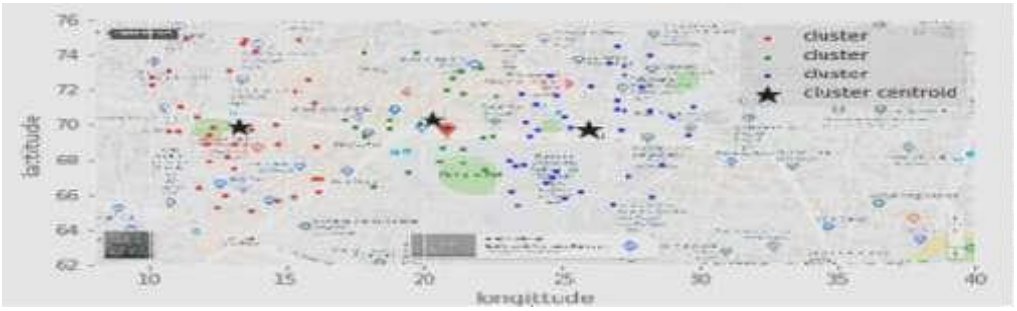
standards by whichThe following figures shows the results obtained.

- Stop Clustering after the fixed number of iterations
- The sum of the squared distance of each record to its "representative mean" in each cluster is less than some threshold value. The threshold value can be calculated by using below mathematical equation (2).

E Kd2 (x, m)

Where,

```
number of vehicles in clusters are:

cluster-1: 8

cluster-2: 30

cluster-3: 12
100
150
Random clusters centroids are:
[[17. 72.]
 [29. 78.]
 [24. 65.]]


cluster centroids are:
[[16.697195 68.50744...
 [25.556175 70.16902 ]
 [13.643834 73.458405]]


cluster lables are
[0. 2. 2. 0. 0. 2. 0. 1. 2. 2. 1. 1. 2. 1. 1. 2. 2. 2. 1. 1. 1. 1. 1. 0.
 1. 1. 1. 1. 2. 1. 1. 1. 2. 1. 1. 1. 1. 0. 1. 1. 1. 2. 1. 0. 0. 0. 0. 1.
 1. 1.]
```



CONCLUSION

A stream data clustering approach called k-Means is examined in this research. Python is used to programme the system. Graphics may be created with the aid of the matplotlib requirement. Data about clusters is shown as text. There are graphic graphs that display the cluster densities. We're pleased with the outcomes

REFERENCES

[1] Jiawei Han and Michelin Kamber ,"Data mining Concepts and techniques", second edition, Pg No. 383-531, 2006.

[2] 2 Sobhan Badiozamany, "Real-time data stream clusters on sliding windows,

International Journal of Marketing Management

ISSN 2454-5007, www.ijmm.net

Vol. 9 Issuse. 2May 2021

Digital Comprehensive Summaries of Research Papers."
[3] The findings are shown in a database table.
[4] Dissertations from the University of Uppsala.
[5] the University of Toronto's Faculty of Science, and
[6] After real-time clustering has been completed.
[7] ISBN 1651-6214, pp. 2016: "Technology 1431"
[8] Georg Krempel, Ammar Shaker, "Open Challenges for a Sustainable World"
[9] The IEEE Transactions on Big Data, vol. 16, no. 1, pp. 1-9, 2013, Data Stream Mining Research.
[10] Noorwati Mustapha Norwati, and Norwati Mustapha "IMECS: Information Management and Communications Standard
[11] Challenges and issues in stream clustering "Volume 1 of the journal, p. 2010.
[12] An overview of stream data algorithms may be found in T. SoniMadulata's "ACIJ:Overview of stream data algorithms" (pp. 151-160).
[13] shows a live graph with multiple color-coded groupings.
[14]
[15] [6]Xiangliang Zhang, Cyril Furtlehner, Cécile Germain- Renaud, and MichèleSebag, "IEEE transactions on knowledge and data engineering: Data Stream Clustering with Affinity Propagation", vol. 26, 1644-1656.
[16] Abhirup Chakraborty , School of Informatics and Computer Science, Ajit Singh Department of Electrical and Computer Engineering, “IEEE transactions, Parallelizing Windowed Stream Joins In A Shared-Nothing Cluster", pp. 2013.

[17] Springer Link, “Evolution of real-time traffic applications