International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-9, July 2020

DEN-DIS: "May Get Life in Future" - Hybridized Data Stream Clustering Framework in Market Research Arena

Jasmine Natchial.F, Elango Parasuraman

Abstract*: Data streams pose several computational challenges due to their large volume of massive data arriving at a very fast rate. Data streams are gaining the attention of today's research community for their utility in almost all fields. In turn, organizing the data into groups enables the researchers to derive with many useful and valuable information and conclusions based on the categories that were discovered. Clustering makes this organization or grouping easier and plays an important role in exploratory data analysis. This paper focuses on the amalgamation of two very important algorithms namely Density Based clustering used to group the data and the dissimilarity matrix algorithm used to find the outlier among the data. Before feeding the data, the algorithm filters out the sparse data and a continuous monitoring system provides the frequent outlier and inlier checks on the live stream data using buffer timer. This approach provides an optimistic solution in recognizing the outlier data which may later get reverted as inlier based on certain criteria. The concept of DenDis approach will pave a new innovation world of considering every data which "May Get Life in Future".

Keywords: DenDis, monitoring, Clustering, Density.

I. INTRODUCTION

Data streams are defined as massive data generated at a very high speed. Streaming data poses challenges to today's computational world. Data Streams when properly mined and analyzed serve as an important tool to extract useful information that would assist researchers to derive valuable solutions in real time situations. This field of study fascinated many researchers over the last era to design innovative algorithms or adopt existing ones or amalgamate different algorithm catering to the needs of time. There are numerous number of techniques available to handle Data streams very effectively out of which four different categories are found to be of utmost importance namely ,

1. Two phase Mining
2. Hoeffding bound based Mining
3. Symbolic approximation-based mining
4. Granularity based mining

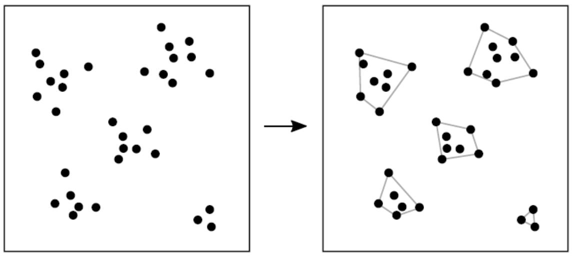
Revised Manuscript Received on July 30, 2020.
* Correspondence Author

F.Jasmine Natchial*, Research Scholar, Department of Computer Science, Bharathiar University, Coimbatore, INDIA. E-mail:jasmine_natchial@yahoo.com

Elango Parasuraman, Assistant Professor, Department of Information Technology, Perunthalaivar Kamarajar Institute of Engineering and Technology, Karaikal, INDIA. E-mail: elanalin_74@yahoo.com

Clustering is an important machine learning algorithm to group similar data and filter out the unwanted irrelevant data. In detail, classifies a huge massive real time data into a meaningful sub-group. The subgroups are selected with reference to the base point where the intra-cluster differences are reduced and the inter-cluster differences are increased to show a clear boundary layer among the clubbing of data. On the whole, Clustering (Fig 1) i.e. grouping the data objects according to similarity or dissimilarity, serves as a valuable and simple method of data mining technique. Below are some of the approaches made in the clustering arena such as,

- Partitioning based
- Hierarchy based,
- Density based and
- Grid based



I. Clustering of data objects

Outlier detection in data streams i.e., the process of deciding which data to reside in and out of the datasets, proves to be more useful in several areas such as fraud detection – Predominantly imposed in banking systems, computer network and cyber security, medical / public health anomaly detection, etc. The basic formulation to detect a data object as an outlier involves, checking the behavior / impact on the datasets and comparing with the expected behavior. In case of anomaly behavior, it's considered as an outlier. This research focuses on detecting the distance-based outliers by emphasizing the concept of identifying the data object in a generic metric space as an outlier by coining the condition, that the data objects should be bounded within distance R (Acceptable limit) from C (Mean value-Centroid). Many factors such as distance, density etc., can be used as variants to detect the anomaly data. As far as data streams are concerned, the dataset size has no boundary.

So anomaly removal is performed over a sliding window, i.e., by counting the active objects. This is to assure and ascertain the efficiency in computation and anomaly computation in a local arena. This is a type of market research which helps the car manufacturer to decide the factors that are valued by the customers while purchasing a car.

This paper has the following subtopics: In Section 2, a survey of Literature is presented. Section 3 discusses briefly the problem definitions and area of research. In Section 4, we provide our performance metrics considered. In Section 5, proposed methodologies and algorithms are explained. In Section 6, Performance Evaluation of this research and final section 7 lists down the reference papers used in this research.

II. LITERATURE SURVEY

Chi Wing and Ada. W. Fu et al [2] designed three algorithms namely Chernoff bound, BOMO and Lossy counting algorithms to prune and process the top K datasets. They addressed many problems associated with mining top k itemsets in their research. The itemsets were categorized into batches for easy processing into local and global pools. They were successful in utilizing memory to a greater extent. Claudio et al [3] focused his research in solving the problem of identifying the top K patterns in the presence of disturbances and issues with the data itself. He used PANDA algorithm to process the dataset. Behera-et-al [10] proposed an algorithm to mine the outlier using the technique of clustering. He combined the clustering algorithm and some outlier detection techniques to deal with the data of both lower and higher dimensions. Ming-jian Zhou-et-al [9] proposed anomaly finding algorithm using Dissimilarity principle. The extent of dissimilarity called as dissimilarity degree is found and compared with Threshold to identify the outliers. This algorithm failed to consider the non-numerical attributes.

III. PROBLEM DEFINITION

The complexities involved in the disembarkation and desertion of data objects in a streaming environment introduces new challenges in outlier detection in terms of time and space efficiency. If we look back, several studies have been performed adopting unsupervised definition and ignoring the distributional assumptions on data values in the case of distance-based outlier detection in data streams (DBODDS). We systematically evaluate the most recent algorithms for DBODDS under various stream settings and outlier rates.

A. Data Stream

A data stream is a possible incessant series of data points ..., o_n-2, o_n-1, o_n... where data point o_n is received at time o_n.t. In this definition, a data point o is associated with a time stamp o.t at which it arrives and the stream is ordered by the arrival time. As new data points arrive continuously, data streams are typically processed in a sliding window, i.e., a set of active data points.

B. Inlier Data Set

In business considerations, the boundary value is considered as threshold T (T > 0) – which is a maximum and minimum bounded value, a data point x is a neighbor of data point x ′ if the distance between x and x ′ is not greater than T.

C. Outlier Data Set

Given a dataset d_t, a count threshold k (k > 0) and a distance threshold T (T> 0), a distance-based outlier in d_t is a data point that has less than k neighbors in d_t.

A data point that has at least k neighbors is called an inlier. Figure 2, explains the evaluation of data in both static datasets and Data Streams.

D. Dendis Methodology

In this methodology (Fig 3), the high stream data is filtered using Sparse Combo filter and then clustered using density clustering algorithm.

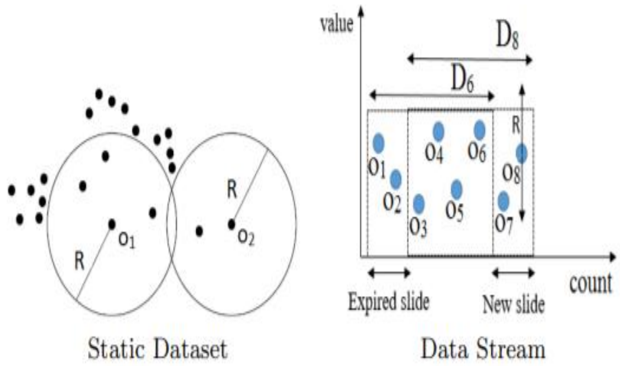


Fig.2.Outlier/Inlier Data Set Identification

It is then taken to the next level of scrutinization using dissimilarity matrix technique where the final outlier detection is processed and the data is validated.

The core problem in the field of Data streams is with evaluating the data which is dynamic and changing frequently and essentially care should be taken in validating the data, for the data which is outlier at present may become inlier in due course.

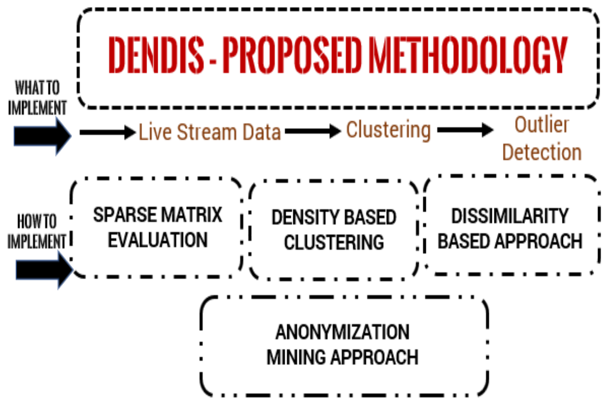


Fig.3.Dendis Methodology

This is due to the threshold value change and the data will be reconsidered for future level of accuracy metric evaluation (Fig 4). In addition, we need to analyze on the source of data which may be Distributed or Centralized.

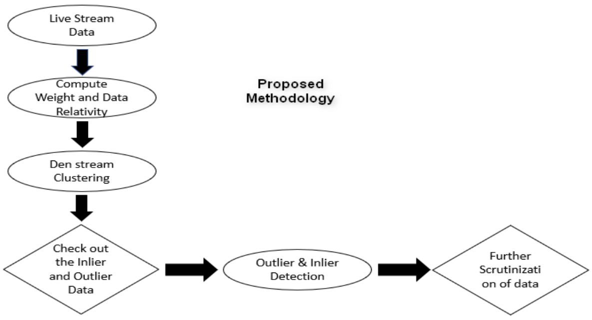


Fig.4.Proposed Methodology Algorithmic Steps

Data points get generated at multiple nodes in a distributed environment where the nodes do some computations locally and the aggregate results are sent by a sink node to find the outliers globally.

IV. EVALUATION METRICS

CPU time and peak memory requirement are the most important utility metrics for streaming algorithms. The time needed for processing new slide, the expired slide and the time needed for manipulation and estimation of outlier comes under the CPU time. The peak memory consumption measures the highest memory used by a DBODDS manipulation for each window which includes the data storage as well as the algorithm specific structures to incrementally maintain neighborhood information.

V. PROPOSED METHODOLOGY [DEN STREAM ALGORITHM + DIS-SIMILARITY MATRIX]

Existing methodologies involves outlier detection on the similar data and the removal of outliers as a whole. Here there are possibilities for an outlier to carry potential information that would affect the existing scenario and they don't have an algorithm to work on the dissimilarity matrix data.

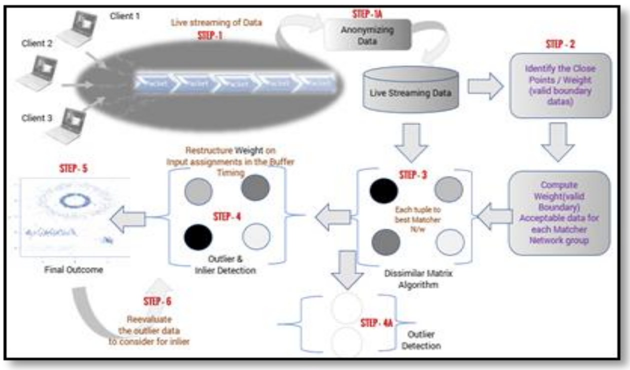


Fig.5. Proposed Methodology Architecture

The proposed approach (Fig 5) involves a live streaming system which gathers data from multiple locations. This data will get loaded at regular intervals and fed to the centralized server (Fig 6) using pull subscription methodology (Fig 7).

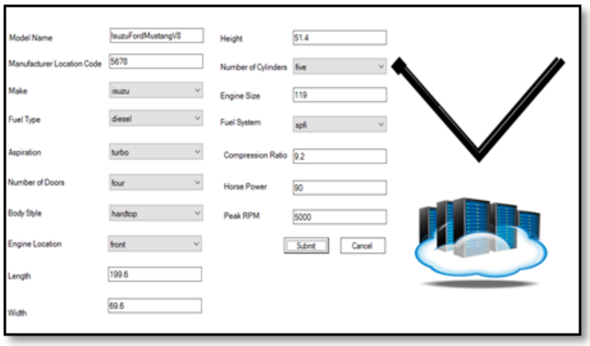
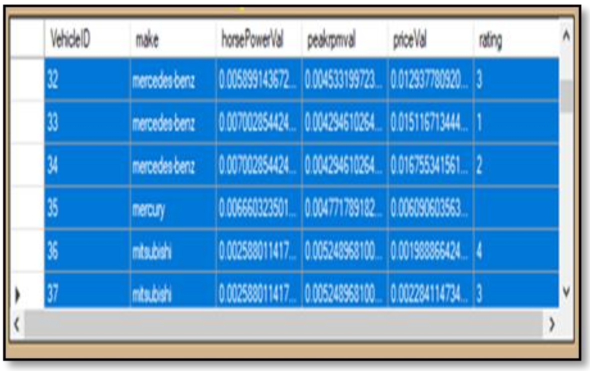
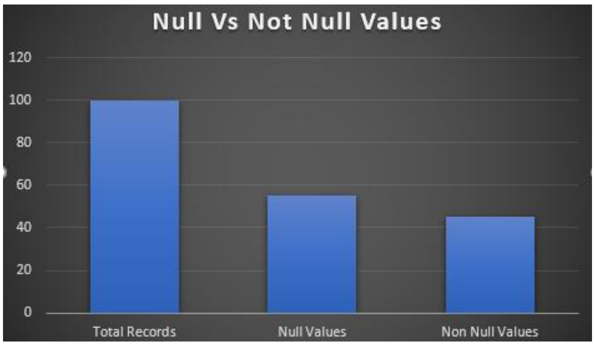


Fig.6.Data Input Feed System



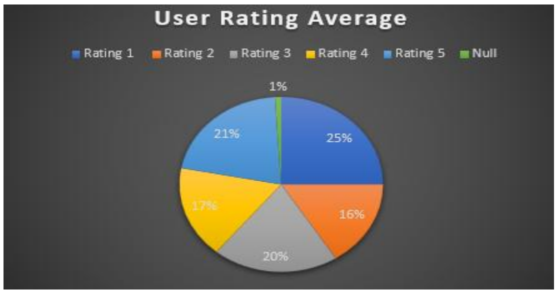
VehicleID	make	horsePowerVal	peakrpmval	priceVal	rating
32	mercedes-benz	0.00589143672...	0.004533199723...	0.012937780920...	3
33	mercedes-benz	0.007002854424...	0.004294610264...	0.015116713444...	1
34	mercedes-benz	0.007002854424...	0.004294610264...	0.016755341561...	2
35	mercury	0.006660323501...	0.004771789182...	0.006090603563...	
36	mitsubishi	0.002588011417...	0.005248968100...	0.001988866424...	4
37	mitsubishi	0.002588011417...	0.005248968100...	0.002284114734...	3

Fig.7.Input Data –Car feedback datasets



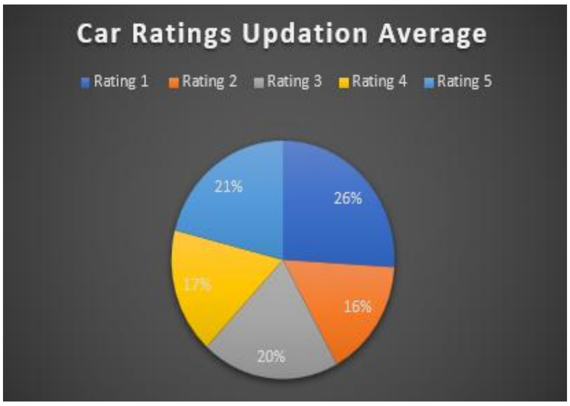
Data Fed into the system for the car records

Total Records	Null Values	NonNull Values
100	55	45



Rating Fed into the system for the car records

Rating 1	Rating 2	Rating 3	Rating 4	Rating 5	Null
595	877	854	877	887	1



Car Ratings Updation Fed into the system for the car records

Rating 1	Rating 2	Rating 3	Rating 4	Rating 5
26	16	20	17	21

Once the data is pushed into the centralized server, the sparse matrix algorithm (Fig 8) is placed to filter the data and appropriate null proximation technique is used to remove the unwanted data.

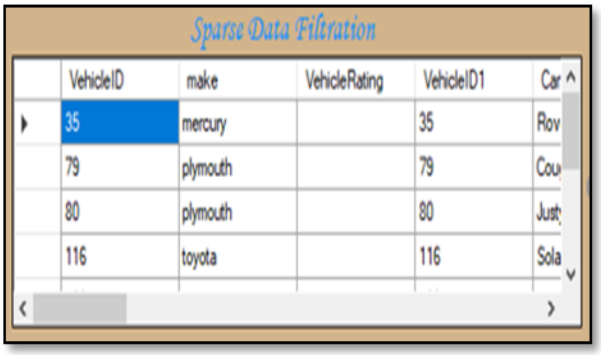
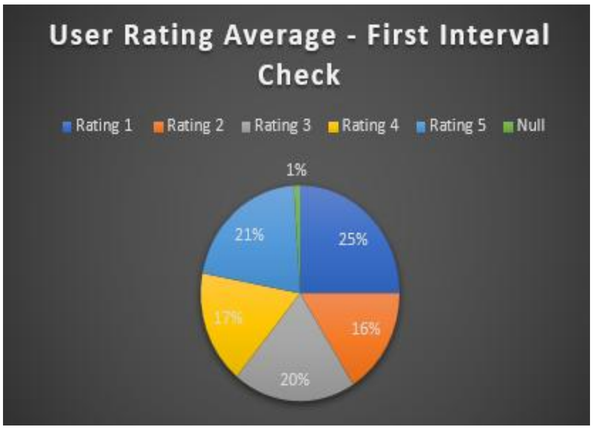
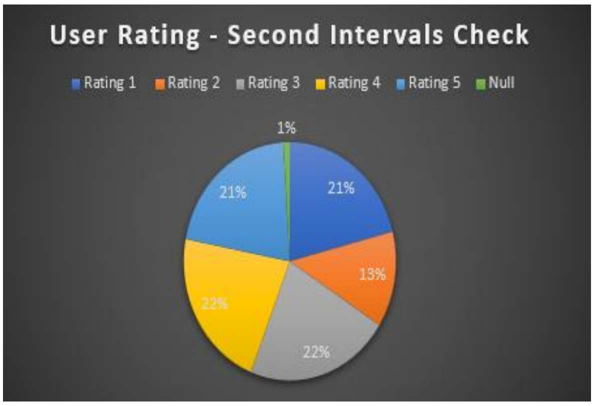


Fig.8.Input Data – Car feedback datasets

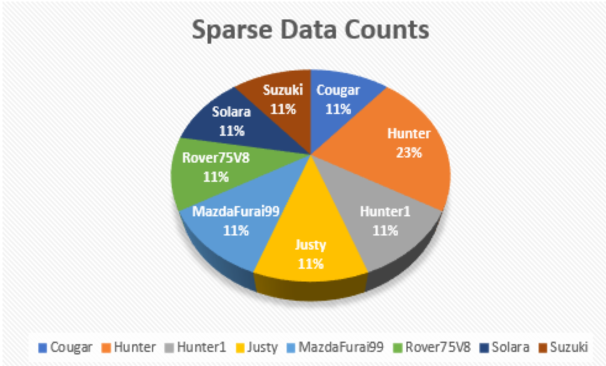


Rating 1	Rating 2	Rating 3	Rating 4	Rating 5	Null
25	16	20	17	21	1

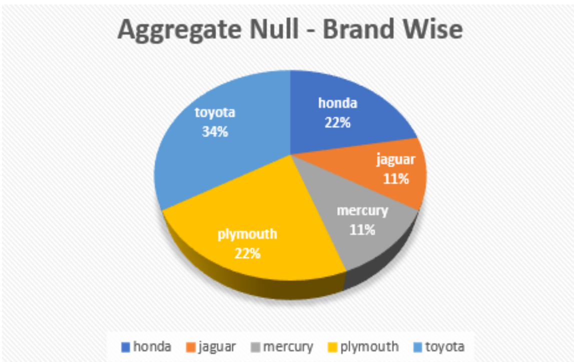


Rating 1	Rating 2	Rating 3	Rating 4	Rating 5	Null
21	13	22	22	21	1

Sparse Filtering



Car Name	Aggregate Null Count
Cougar	1
Hunter	2
Hunter1	1
Justy	1
MazdaFurai99	1
Rover75V8	1
Solara	1
Suzuki	1



International Journal of Innovative Technology and Exploring Engineering (IJITEE)
ISSN: 2278-3075 (Online), Volume-9 Issue-9, July 2020

A. Density Clustering Implementation

Based on the data relativity, the live stream data will be clustered. This can be done based on the weighing process on certain deciding factors.

Relative Weight, $w = \sum_{j=1}^{n} f(t - t_j)$
Weight should be above the threshold $w \geq \mu$.

Where $f(t - t_j)$ – Function to manipulate the relative weight at a particular interval period.
μ.-Threshold to find outlier or inlier

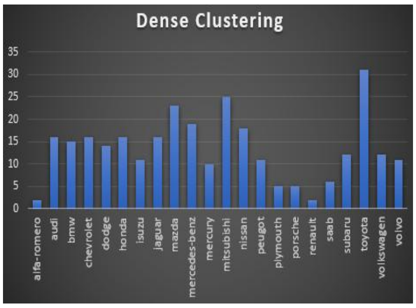
In this research, the clustering process is done with respect to product make.

```
Algorithm : Density Clustering Algorithm
Select n points as Initial centroids  //Productmake is considered in this research
Repeat the process
  Form n clusters by assigning each point to its closest centroid.
  Recompute the centroid of each cluster by summing up the relative weight
until the stream data is clustered
```

The above process clusters the data which is further scrutinized to get inlier and outlier data using dissimilarity matrix algorithm. This can be achieved with the manipulation of centroid value in a cluster.

$$\textbf{Centroid , c} = \frac{\sum_{j=1}^{n} f(t - t_j) \, . \, p_j}{w}$$

Where p_j – Average Change due to dynamic inputs



Brand Name	Dense Clustering
alfa-romero	02
audi	16
bmw	15
chevrolet	16
dodge	14
honda	16
isuzu	11
jaguar	16
mazda	23
merc-benz	19
mercury	10
mitsubishi	25
nissan	18
peugot	11
plymouth	05
porsche	05
renault	02
saab	06
subaru	12
toyota	31
volkswagen	12
volvo	11

B. Dissimilarity Algorithm Implementation

Brand Name	Aggregate Null Count
honda	2
jaguar	1
mercury	1
plymouth	2
toyota	3

Dissimilarity matrix involves the following algorithm

```
Algorithm : Dissimilar Matrix Algorithm
//////// Inlier and outlier Segregation ////////
Initializing the list of clusters – ( The clusters are associated with related points )
repeat
  Validate a cluster from the list of clusters
  { Perform several "trial / validations"   on the chosen cluster. }
 for i = 1 to number of trials( Based on the no. of elements in the cluster)
    do
            Check for the outlier and inlier data based on the centroid value of the cluster
            Outlier data will be moved out from the cluster
   end for
//////// Outlier Buffer manipulation ////////
Initializing the list of outlier data
repeat
  Validate the outlier data with the cluster data
  { Perform several "trial / validations"   on the chosen cluster. }
 for i = 1 to Buffer timer
    do
            Check for the outlier and inlier data based on the centroid value of the cluster
            Outlier data will be retained out from the cluster. In case of inlier data, its moved into the cluster
   end for
```

The next level of process involves the outlier detection technique using DenDis framework which involves the self-evaluation of data using top k evaluation to finalize the outlier data and the relative inlier data.

C. Dendis Framework Implementation Methodology:

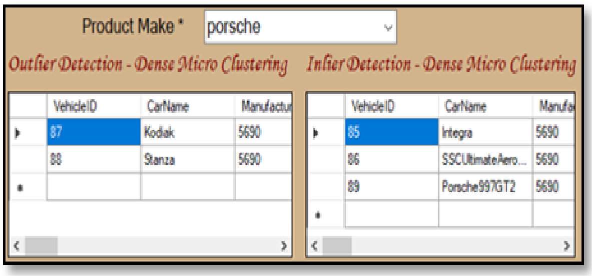
Step 1: Identify the evaluation metrics and set the metrics as the base for the evaluation. This research comes up with some of the mandatory fields like horsepower of the engine, price, and peakrpm and engine size.

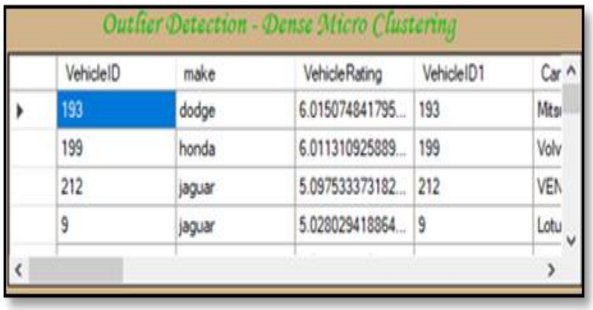
Step 2: Check for the outlier

Step 3: Condition manipulated in identifying the outlier is,

if (Not exists)
h1.horsepower <= h.horsepower
AND h1.price <= h.price
AND h1.peakrpm <= h.peakrpm
AND h1.enginesize >= h.enginesize
and
(h1.horsepower < h.horsepower OR
h1.price < h.price OR h1.peakrpm < h.peakrpm OR h1.enginesize > h.enginesize))

Step 4: Fetch the data sets which doesn't match Step 3. This is considered as inlier data sets (Fig 9) which used for evaluation and final prediction of pricing.





VehicleID	make	VehicleRating	VehicleID1	Car
193	dodge	6.015074841795...	193	Mits
199	honda	6.011310925889...	199	Volv
212	jaguar	5.097533373182...	212	VEN
9	jaguar	5.028029418864...	9	Lotu

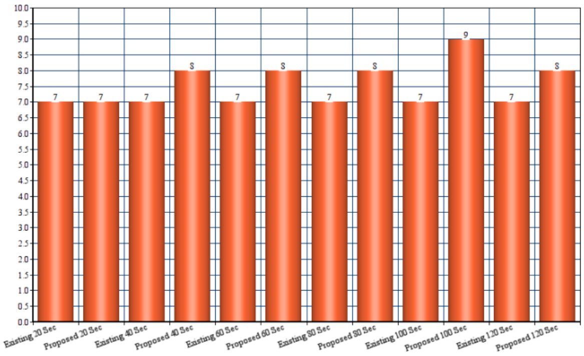
Fig.9.Inlier vs Outlier data sets segregation

Fig.10. Outlier data objects

Once the data is filtered, the research provides an option of filtering and analyzing the data as outlier (Fig 10) and inlier data based on custom filter.

VI. RESULT ANALYSIS

A comparative analysis has been done among the existing and proposed system where the buffer timers option were considered as the base difference due to the swirling of input data moving on towards outlier and inlier detection. The proposed system provides an optimized solution for the problem entified.



Existing System	Proposed System	Performance Improved Due to Data reconsideration
42	48	**14%**

Existing System - No Buffer Timers				
Time (t)	No Of input feeds	Product Make Cluster	Outlier Detection Count	Inlier Detection Count
20	10	Mazda	3	7
40	15	Mazda	3	7
60	22	Mazda	3	7
80	31	Mazda	3	7
100	47	Mazda	3	7
120	64	Mazda	3	7
Proposed System - With Buffer Timers				
Time (t)	No Of input feeds	Product Make Cluster	Outlier Detection Count	Inlier Detection Count
20	10	Mazda	3	7
40	15	Mazda	2	8
60	22	Mazda	2	8
80	31	Mazda	2	8
100	47	Mazda	1	9
120	64	Mazda	2	8

Fig.11. Existing Vs Proposed Performance Evaluation

In the existing system (1), the buffer timers were not considered as a factor in restructuring the outlier data. The data which is coined as outlier resides out of focus from the systems. In the above table, the outlier detection and inlier detection count remains constant irrespective of time. Even the new data couldn't impact the counts due to the reason for not reconsidered after a specific period of time. Proposed systems (Fig 11) resolve this issue with reconsidering the data within a buffer time. The data will reorganize based on the Dendis Algorithm.

VII. CONCLUSION

There are many methods available for identifying the outliers in data stream. But all these methods eliminate the identified outliers and do not consider them in the future. This may lead to some potential information to get lost or the valuable information carrying datasets may be driven out of focus as outliers. These factors are taken into consideration in this paper .The data sets identified as outliers are maintained in the buffer for certain time so that they get opportunity to get reorganized into inlier. Thus this algorithm ensures that no information get lost.

This paper speaks about amalgamation of two different techniques which are already proven to be efficient. Thus utilizing the usefulness of various algorithms to prune the datasets and filter them very cautiously seems to be the highlight of this DENDIS framework.

REFERENCES

1. Feng Cao - Department of Computer Science and Engineering, Fudan University , Martin Estery - School of Computing Science, Simon Fraser University, Weining Qian-Department of Computer Science and Engineering, Fudan University, Aoying Zhou-Department of Computer Science and Engineering, Fudan University - ***Density-Based Clustering over an Evolving Data Stream with Noise***
2. Chi-Wing Wong, Ada W.Fu. Sep 2006 "***Mining top-K frequent itemsets from data streams***" in Data Mining and Knowledge Discovery 13(2):193-217 · September 2006.
3. Claudio Lucchese, Salvatore Orlando, Raffaele Perego Published in SDM 2010 - "***Mining Top-K Patterns from Binary Datasets in presence of Noise***" DOI:10.1137/1.9781611972801.15
4. N. Archak, A. Ghose, and P.G. Ipeirotis, "***Show Me the Money!: Deriving the Pricing Power of Product Features by Mining Consumer Reviews***," Proc. 13th ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 56- 65, 2007.
5. X. Lin, Y. Yuan, Q. Zhang, and Y. Zhang, "***Selecting Stars: The k Most spokesperson Skyline Operator***," Proc. Int'l Conf. Data Eng. (ICDE), 2007.
6. Luh Yen ; Marco Saerens ; Francois Fouss - "***A Link Analysis Extension of Correspondence Analysis for Mining Relational Databases***" in IEEE Transactions on Knowledge and Data Engineering (Volume: 23 , Issue: 4 , April 2011)
7. S.Vijayarani, P. Jothi, ***"An Efficient Clustering Algorithm for Outlier Detection in Data Streams"*** in International Journal of Advanced Research in Computer and Communication Engineering, Vol. 2, Issue 9, September 2013.
8. Yu Zhongqing, Fang Yi, Pan Zhenkuan, Shao Fengjing, ***"OLPA Architecture,"*** Qingdao-Hong Kong international Computer Conf., 1999, 10.
9. Ming-jian Zhou, Xue-jiao Chen, ***An Outlier Mining Algorithm Based on Dissimilarity, in the Procedia Environmental Sciences*** 12 (2012) 810–814, 2011 International Conference on Environmental Science and Engineering.
10. Behera H.S., Abishek Ghosh, Sipak ku. Mishra, ***A New Hybridized K-Means Clustering Based Outlier Detection Technique For Effective Data Mining.***
11. Surekha V Peshatwar, Snehlata Dongre, ***Outlier Detection Over Data Stream Using Cluster Based Approach And Distance Based Approach,*** International Conference On Electrical Engineering and Computer Science.

AUTHORS PROFILE

F.Jasmine Natchial is a research scholar of Bharathiar University, Coimbatore. She completed MCA at Pondicherry University and M.Phil. at Bharathidasan University, Thiruchirapalli, India. Her area of interests includes Cloud Computing, Data mining and Outlier mining.

Elango Parasuraman is working as an Assistant Professor in Department of Information Technology at Perunthalaivar Kamarajar Institute of Engineering and Technology, Karaikal, India. His area of interests includes image processing, data mining, and web mining. He completed his Ph.D., at National Institute of Technology Tiruchirappalli, India, in 2011, and his M.Tech, at National Institute of Technology Karnataka, India, in 2005.