IJCSE International Journal of Computer Sciences and Engineering Open Access
Research Paper Volume-5, Issue-12 E-ISSN: 2347-2693

A Survival Study on Data Structure Based Clustering Techniques for Multidimensional Data Stream Analysis

K.Chitra[1*], D. Maheswari[2]

[1*] School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, Coimbatore, India
[2] School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, Coimbatore, India

[*]Corresponding Author: chitra.k@rvsgroup.com

Available online at: www.ijcseonline.org

Received: 16/Nov/2017, Revised: 22/Nov/2017, Accepted: 18/Dec/2017, Published: 31/Dec/2017

Abstract— Data mining plays an effective role in the field of computer science to analysis the data objects. The data mining process is used to mine the knowledge from huge database. Then, the extracted information is modified into an understandable data structure for the future analysis. The data structure in a computer is an essential approach to categorize and manage the data which is utilized for efficient usage. The data stream is referred as a structured sequence of instances; the data stream mining discovers the knowledge structures from continuous and fast data records. The clustering is the process of creating the group by collecting the data of similar patterns and also describes the meaningful structure of data. The additional process of traditional clustering termed as Subspace Clustering which is utilized for detecting the clusters in various subspaces within dataset. Then, the subspace clustering algorithms are introduced to discover the cluster in multiple overlapping subspaces by searching the relevant dimensions. Many research works are developed for managing the high dimensional data with the objective of providing better improvement on minimizing the performance of dimensionality and enhancing the clustering accuracy. However, the existing works failed to reduce the space complexity. Therefore, the research work focuses on reducing the dimensionality with improved clustering accuracy by executing the clustering and subspace clustering for data stream with data structure techniques.

Keywords—Data stream, Multidimensional data, Data mining, Data structure, Subspace clustering.

I. INTRODUCTION

The main aim of the data mining is developed for detecting and directing the reliable data from large data sets on data base. The data structure is one of the aspect to categorize and storing data in the form of array, file, record, table and tree for the specified purposes. In data collection, identification of patterns and trends help to detect the unusual data records by implementing the different classification and clustering rules in mining process. Clustering is the process of collecting and organizing the similar data which are in same group or different groups.

Based on the predefined similarity measurement, cluster analysis techniques are used for clustering the set of data objects which are in same group are more similar to each other groups. Clustering is the one of effective technique to analyze the high dimensional data from a few dozen to many thousands of dimensions. While handling the multidimensional data for clustering, the data are grouped into categories such as data dimensions and measurements. In the data mining techniques, the analysis of multidimensional data stream with dimensionality reduction is a challenging task.

Several approaches such as dimensionality reduction, subspace clustering, projected clustering, hybrid clustering and correlation clustering are developed for clustering of multi dimensional data. The dimensionality reduction is the one of convenient process for grouping the objects in data mining by minimizing the number of random variables of a dataset. Then, the dimensionality reduction is involved with two processes such as feature selection and feature extraction. The subspace clustering is essential to cluster the data into different subspaces within a dataset. The subspace clustering is used to searches the relevant d dimensions by permitting them to group the data that are present in several or probably overlapping subspaces.

Many research works as clustering algorithms are developed with the aim of clustering the data objects. However, these works are failed to process the high-dimensional data. Therefore, some subspace clustering techniques are introduced to cluster the multidimensional data with dimensionality reduction. The performance of data structured based clustering techniques are analyzed

This paper is prepared as follows: Section II discusses reviews on data structured based clustering techniques, Section III describes the existing data stream clustering process with data structure techniques, Section IV identifies the feasible comparison among them, Section V explains the boundaries and Section VI concludes the paper, salient areas of research is specified as to achieve the dimensionality reduction and to improve the clustering accuracy by performing the data structure techniques based on clustering approach.

II. LITERATURE SURVEY

In [1], uncertain object was formed by implementing the Probability Density Function (PDF) of Gaussian distribution. Then, nested loop develops in the naive approach of distance-based outlier detection. The designed approach was valuable, due to the expensive distance function among two uncertain objects,. The populated-cells list (PC-list) approach was performed with the help of top-k outlier detection algorithm to classify the dataset objects and to detect the candidate objects. But, PDF of gaussian distribution failed to reduce the dimensionality issues and the time computation was high while using Naïve approach for identifying the top-k outlier.

In [2], the high dimensional data was executed with the aid of Locality sensitive hashing (LSH) by solving the nearest-neighbor search issues. The high-dimensional data was collected by implementing the Two LSH techniques. The SimHash allows the cosine similarity and MinHash allows the similarity approximations. According to the single-pass process, low-dimensional Hamming embedding was developed to approximate a pair wise similarity matrix. During the single-pass process, the data storage needs not to perform, but it needs to preserve the low-dimensional embedding. Besides, the similarity matrix was formed with the help of bisection method by selecting the clustering solution. However, time complexity and computational cost remained unaddressed for clustering process. Then, the designed approach was not able to obtain the self-contained solution by combining the two LSH techniques.

In [3], the global and local clustering structure was conserved by introducing the dimension reduction algorithm modeled as unsupervised linear dimension reduction algorithm for data with cloud distribution. For all data in original space, the clustering labels were generated by using K-means clustering method. Then the local and global clustering structure was formed and protected with the help of obtained clustering labels. However, K-means clustering method able to search and detect the nearest neighbor but failed to enhance the clustering accuracy.

The distinctness of clustering structure was controlled by implementing the Distinctness Preserving Dimensionality Reduction (DPDR) method in [4]. Here, the structure was modeled as unknown. For the analysis purpose, the designed method was enclosed with one known cluster inside the data. This in turns, the space dimensionality gets reduced by protecting the data from Fisher's linear subspace. The designed method contains the reasonable assumptions without any knowledge of clusters. But, the clustering structure was not enhanced by using DPDR method. In addition, Fisher's linear subspace failed to reduce the space complexity.

In [5], an efficient new data-representation scheme was implemented to categorize the data and extract the categorical objects into Euclidean space. The space-structure based categorical clustering algorithm (SBC) was developed based on the data-representation. But, the time complexity was high while mapping the data objects by using Euclidean distance calculation. Besides, the clustering algorithm was not improved effectively for categorizing the data.

In [6], an effective data-driven similarity learning approach was implemented with the aim of generating the coupled attribute similitude measure to the nominal objects with attribute couplings for grouping the global representation of attribute similitude. The frequency-based intra-coupled similarity was estimated while considering the attributes. The determination of inter-coupled similarity was done based on the value of co-occurrences among attributes. According to the attributes relationship, similarity between two categorical values was estimated by developing the four measures for inter-coupled similarity. The theoretical analysis ensures better results in terms of accuracy and efficiency of measure while considering the intersection set for large-scale data sets. But, the data-driven similarity learning approach was not able to reduce the space complexity.

Based on the feature information and spatial structures, Tensor Low-Rank Representation (TLRR) and sparse coding-based (TLRRSC) subspace clustering method was presented by [7]. The lowest rank representation was detected in the original spatial structures with the spatial directions. Every sample was indicated as the atoms of learned dictionary by using the Sparse coding found dictionary along feature spaces. In the spatial and feature spaces, the spectral clustering was performed by affinity matrix. The TLRRSC constructs the global structure and intrinsic feature information of data. Besides, the TLRRSC carried outs the strong subspace segmentation from corrupted data. However, the performance of clustering accuracy was not enhanced through the TLRR and TLRRSC method.

The error effects from projection space than from input space. l_1-, l_2-,…, l_∞- was avoided by implementing the Novel error-removing method in [8]. The designed method was developed to share the property of intrasubspace projection dominance by the nuclear-norm-based linear projection spaces. The intersubspace data points were smaller than the coefficients over intrasubspace data points. The sparse similarity graph termed L2-graph was developed by subspace clustering and subspace learning algorithms.

However, time for the subspace clustering was able to reduce by using the Novel error-removing method.

The clustering dataset of repetitive data issues were solved in [9], by implementing the DCSTREAM method with k-Means divide and conquer approach and vector model. However, the STREAM failed to enhance the quality of cluster. In [10], the streaming of data in one-pass-thrown-away fashion was grouped by implementing an efficient versatile hyper-elliptic clustering algorithm termed VHEC. The elliptic micro-cluster factors such as boundary, direction, density, inter-distance and intra-distance were estimated by solving the issues in the one-pass-thrown-away clustering. But, the designed approach was not able to improve the clustering accuracy. If the number of data points is lesser than number of dimensions, then the designed approach produces wrong results.

In [11], the asymmetric self-organizing map was clustered by means of the asymmetric approach. The asymmetric version of k-means algorithm was developed by performing the efficient clustering to obtain the two-stage fully asymmetric data analysis method. But, the designed approach failed to perform the dimensionality reduction and to reduce the space complexity. K-means is a clustering method for high dimensional data sets. But k-means is prone to local minima problem [12].

III. CLUSTERING OF MULTIDIMENSIONAL DATA

In the design of intelligent systems for real-time applications, clustering plays an essential role. Clustering is process of allowing the similar objects to be clustered into groups. The handling of large data set is one of most significant challenging task in the clustering process. While performing the clustering on multidimensional data, there is a need to attain the reduction on dimensionality with improved clustering accuracy. This in turns, the performance of clustering on multidimensional data is improved thereby the memory consumption for storing the data gets reduced.

Some subspace clustering techniques are introduced to cluster the multidimensional data with dimensionality reduction. The performance of data structured based clustering techniques are analyzed by comparing the three methods including Locality Sensitive Hashing (LSH) strategy, Tensor Low Rank Representation and Sparse Coding based subspace clustering (TLRRSC) method and Divide and Conquer Approach for data stream (DCSTREAM) method.

3.1 Handling the high dimensional data by using hashing-based clustering

The near-neighbor search issues are addressed in the high dimensional data by introducing the Locality Sensitive Hashing (LSH). The high dimensional data is clustered by implementing the two LSH strategies such as MinHash and SimHash. The Jaccard similarity approximations are enabled by MinHash and then the cosine similarity is estimated with SimHash.

The low dimensional Hamming embedding is used with single-pass procedure to estimate the pair wise similarity matrix while the computational costly data structure is formed for reacting with the queries from near neighbor. The single-pass procedure involves the low dimensional embedding maintenance but not the data storage. The similarity matrix is employed with the bisection method to find out the clustering solution. Besides, the conjunction with SimHash, the cosine similarity estimation is enhanced by developing penalty on the hamming distance.

In the designed approach, MinHash and SimHash are utilized on the clustered data with high dimensionality. Generally, the nearest neighbor is detected by storing the data from the multiple hash tables on memory. The designed approach estimates the similarities between objects with high dimensionality and avoids the requirement of data storage. This in turns, the memory required for finding the near neighbor is reduced. In memory, the closest pairs are controlled and the distant pairs or isolated data are removed by storing the distance in an index of neighbor depends on the adjacency lists.

The managing and neglecting process are done by altering the LSH parameters thereby the number of false positives to provide matches in buckets are limited. Then, the document-at-a-time method is developed to update the adjacency list. This in turns, the cost with estimating the similarity matrix is minimized. After that, the divisive clustering (bisection method) is carried out on lists for attaining the clustered documents which are similar with original feature space.

Besides, the SimHash variant also introduced to change the approximate computation of cosine similarity through providing the penalty factor which is proportional to the decimal representation of the hash code. According to the random hyperplane, the distance of projection of an object is indicated by the decimal magnitude of the hash value. The approximation of cosine similarity is enhanced by using SimHash variant.

3.2 Spectral clustering based on TLRRSC method

The TLRR and sparse coding-based subspace clustering method is implemented with the aim of addressing the issues while performing the spectral clustering. The redundant information and time complexity are solved by TLRRSC method according to the attribute information and spatial structures.

The TLRR is employed for presenting lowest rank representation among original spatial structures within all spatial directions. For indicating an each sample by few atoms of learned dictionary, the dictionary along feature spaces are implemented by using spare coding. In the spectral clustering, the affinity matrix is utilized which is created from the combined similarities in both feature and spatial spaces. Then, the inherent feature information and global structure of data are formed by TLRRSC.

TLRRSC method is presented to develop an original structure form (tensor) and to detect the similarities in all feature dimensions and spatial dimensions. The designed method as TLRR and SC is introduced to perform the subspace clustering by representing the input data in original structure form termed as tensor. For each input tensor, TLRRSC method identifies the lowest rank representation in the spatial mode. Then, the sparse representation for each individual sample in the feature mode is estimated based on the learned dictionary.

From that, the issues such as complexity and memory utilization involved in the classical subspace clustering methods are addressed. Thus, the affinity matrix is constructed to perform spectral clustering by combining the similarities in the spatial and feature modes. Then, the data is clustered into respective subspaces by utilizing the spatial correlation information. TLRRSC method provides better improvement on performing the highly corrupted data for the higher mode tensorial data sets, with the help of data structure.

3.3 Divide and conquer approach for data stream clustering

The new Divide and Conquer Approach for data stream (DCSTREAM) method was developed to address the issues while performing the data stream clustering. The DCSTREAM method provides better improvement on achieving accurate clustering results with improved speed and scalable fashion. By implementing the vector model, k-means divide and conquer approach; the big datasets were clustered in DCSTREAM method.

According to the Divide and Conquer k-Means algorithm, the huge data sets were clustered with the aid of data stream clustering. The designed data stream clustering approach was contained with online and offline components for the real time processing with the aim of grouping the high scale datasets with monotonous data and ordinal data types. For the online component, an effective data stream algorithm was developed with DCSTREAM method thereby the offline components were affected by the online components.

Then, the issues involved in reducing the size and complexities for clustering was achieved by using vector model and divide and conquer approach. In addition, the detection of novel micro-cluster and outliers in batch processing and expired micro-cluster rejection were processed through the designed approach by identifying the concept drift and simplified fading function.

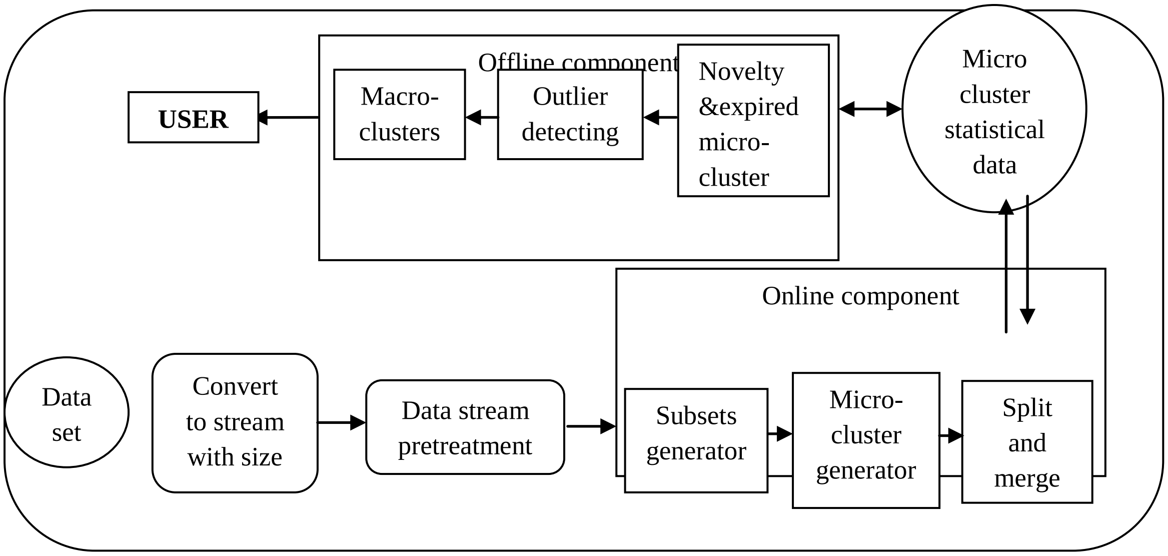

Figure 1 Architecture of DCSTREAM framework

The figure 1 shows the architecture of DCSTREAM framework. The online component contains three models such as micro-cluster generator, subsets generator and split and merge. The data stream pretreatment component is used for preprocessing the stream data from the raw data which is located outside from the online component. Then, the micro clusters are formed by applying the Divide and Conquer k-Means algorithm. After that, through the split and merge module, the concept drift is controlled. In addition, the offline component is implemented through the statistical database to ensure the structure for performing the clustering.

Then, the offline components comprises of three aspects such as macro- clusters, outlier detecting and novelty &expired micro-cluster. The macro-clusters are generated as same as micro-cluster generator in the online component. Then, the user introduces an outlier factor to identify the difference among the novel micro-cluster and abnormalities. The micro clusters are promoted to the real cluster. Then, the fading function with decay concept is used by DCSTREAM method thereby the expired micro clusters are identified. Due to this, the useful micro clusters are removed with memory constraint consideration. This in turns, time and space complexity for clustering the data on database is reduced.

The experimental evaluation using different clustering techniques are conducted on various factors such as clustering accuracy, space complexity and computational time.

IV. PERFORMANCE ANALYSIS OF DATA STRUCTURED BASED CLUSTERING TECHNIQUES

In order to compare the Subspace clustering using data structured techniques, number of data objects is taken as input to carry out the experiments. Various parameters are used for analyzing the performance of dimensionality reduction with improved clustering accuracy.

4.1 Performance analysis of clustering accuracy

The clustering accuracy is measured as the rate of the number of data objects are correctly clustered (grouped) from the total number of data objects using different clustering techniques. The clustering accuracy is mathematically represented as given below,

$$\mathrm{CA} = \frac{\text{Number of data objects grouped correclty}}{\text{Total number of data objects}} * 100.\text{Eqn (4.1)}$$

From the above Eqn (4.1), measure the experimental values for the clustering accuracy and it is denoted as ‘CA′. The clustering accuracy is measured in terms of percentage (%). Higher clustering accuracy ensures the better performance of the method.

Table 4.1 Tabulation of clustering accuracy

Number of data objects	Clustering accuracy (%)		
	LSH strategy	TLRRSC method	DCSTREAM method
10	26.64	20.50	18.65
20	30.92	24.47	20.34
30	33.72	28.76	23.53
40	34.82	32.85	25.67
50	37.64	36.76	29.76
60	41.86	40.99	32.74
70	43.96	45.84	35.94
80	47.22	48.33	41.62
90	54.61	52.87	45.83
100	60.87	55.95	50.66

The above table 4.1 shows the experiments results of clustering accuracy with respect to number of data objects is taken for performing the experiments. The number of data objects is considered from 10 to 100 which are considered as an input. For the simulation purposes, the three methods such as Locality Sensitive Hashing (LSH) strategy, TLRR and Sparse Coding-based subspace clustering (TLRRSC) method and Divide and Conquer Approach for data stream (DCSTREAM) method are compared.

As shown in table 4.1, Locality Sensitive Hashing (LSH) strategy ensures the better performance on improving the clustering accuracy than the other methods. Based on the table value, the graph is plotted in below figure 4.1.

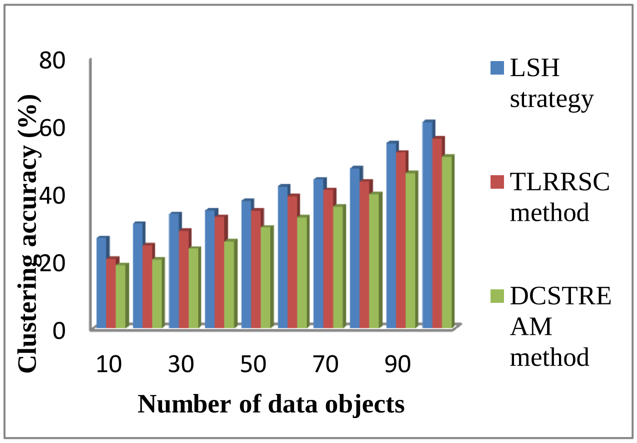

Figure 4.1 Measurement of clustering accuracy

Figure 4.1 illustrates the measurement of clustering accuracy by using three methods namely Locality Sensitive Hashing (LSH) strategy, TLRR and Sparse Coding-based subspace clustering (TLRRSC) method and Divide and Conquer Approach for data stream (DCSTREAM) method. From the figure 4.1, it is evident that, the Locality Sensitive Hashing

(LSH) strategy effectively improves the clustering accuracy when compared to other methods. This is because, Locality Sensitive Hashing (LSH) strategy employed with the bisection method for grouping the high dimensional data with MinHash and SimHash.

As a result, the clustering accuracy in Locality Sensitive Hashing (LSH) strategy is improved by 13% when compared to TLRR and Sparse Coding-based subspace clustering (TLRRSC) method and improves 31% when compared to Divide and Conquer Approach for data stream (DCSTREAM) method respectively.

4.2 Performance analysis of space complexity

The space complexity is defined as the difference between the total memory space and the unused memory. The space complexity is measured as the amount of memory consumed to execute the algorithm based on the size of data objects. The Space complexity is mathematically expressed as given below.

SC = Total memory space – unused memory -- Eqn (4.2)

From the above Eqn (4.2), measure the experimental values for the space complexity and it is represented by 'SC'. The space complexity is measured in terms of Kilo Bytes (KB). When the space complexity is less, then the method is said to be more efficient.

Table 4.2 Tabulation of space complexity

Data object Size (KB)	Space complexity (KB)		
	LSH strategy	TLRRSC method	DCSTREAM method
10	7	6	8
20	17	15	18
30	26	24	28
40	35	33	37
50	46	42	48
60	54	51	56
70	64	62	68
80	76	74	78
90	84	81	86
100	89	86	93

The above table 4.2 shows the experiments results of space complexity with respect to number of data objects is taken for performing the experiments. The size of data objects is considered from 10 to 100 KB which are considered as an input. For the simulation purposes, the three methods such as Locality Sensitive Hashing (LSH) strategy, TLRRSC method and DCSTREAM method are compared.

As shown in table 4.2, TLRRSC method ensures the better performance on reducing the space complexity than the other methods. Based on the table value, the graph is plotted in below figure 4.2.

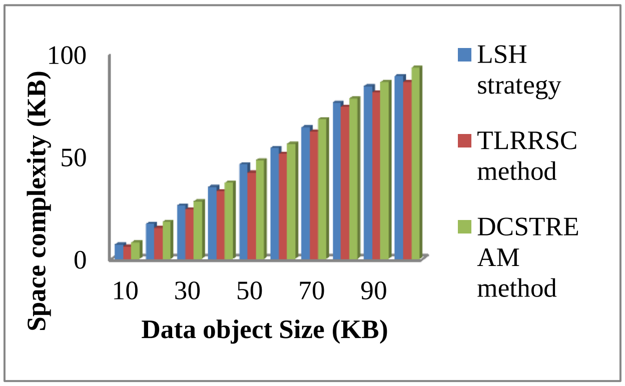

Figure 4.2 Measurement of space complexity

Figure 4.2 illustrates the measurement of space complexity by using three methods namely Locality Sensitive Hashing (LSH) strategy, TLRRSC method and DCSTREAM method. From the figure 4.2, it is evident that, the Tensor Low-Rank Representation and Sparse Coding-based subspace clustering (TLRRSC) method effectively reduces the space complexity when compared to other methods. The TLRRSC method effectively performs the corrupted data by sparse coding representation thereby the space complexity is reduced.

As a result, the space complexity in TLRRSC method is reduced by 7% when compared to Locality Sensitive Hashing (LSH) strategy and 12% when compared to Divide and Conquer Approach for data stream (DCSTREAM) method respectively.

4.3 Performance analysis of computational time

The computational time is measured as the time required for performing the clustering process by the algorithm based on the number of data objects. The computational time is also defined as the difference of ending time and starting time for clustering the objects. The computational time is mathematically formulated as given below,

CT = Ending time – starting time for clustering the objects …Eqn (4.3)

From Eqn (4.3), computational time represented by 'CT' and it is measured in terms of milliseconds (ms). If computational period is less, then the technique is said to be more efficient

Table 4.3 Tabulation of computational time

Number	Computational time (ms)

of data objects	LSH strategy	TLRRSC method	DCSTREAM method
10	24.2	30.6	20.38
20	29.7	33.8	27.85
30	36.2	38.6	34.14
40	42.1	45.4	42.32
50	56.9	58.2	55.89
60	63.8	65.1	69.32
70	73.3	75.2	72.17
80	86.4	88.5	84.45
90	89.4	90.6	88.23
100	93.8	94.3	90.32

The above table 4.3 shows the experiments results of Computational time with respect to number of data objects is taken for performing the experiments. The number of data objects is considered from 10 to 100 which are considered as an input. For the simulation purposes, the three methods such as Locality Sensitive Hashing (LSH) strategy, TLRRSC method and DCSTREAM method are compared.

As shown in table 4.3, Divide and Conquer Approach for data stream (DCSTREAM) method ensures the better performance on reducing the Computational time than the other methods. Based on the table value, the graph is plotted in below figure 4.3.

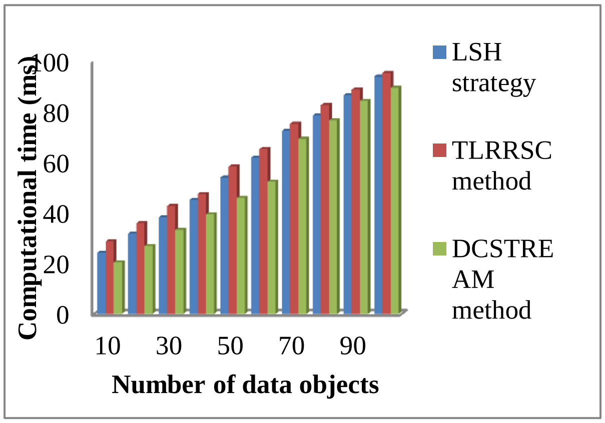

Figure 4.3 Measurement of computational time

Figure 4.3 illustrates the measurement of computational time by using three methods namely Locality Sensitive Hashing (LSH) strategy, TLRRSC method and DCSTREAM method. From the figure 4.3, it is evident that, the DCSTREAM method effectively reduces the computational time when compared to other methods. This is because, use of k-Means divide-and-conquer approach in the DCSTREAM method. By identifying the monotonous data, the extra clustering process are avoided thereby time for executing the data gets preserved.

As a result, the computational time in DCSTREAM method is reduced by 10% when compared to Locality Sensitive Hashing (LSH) strategy and 16% when compared to TLRRSC method respectively.

V. DISCUSSION ON LIMITATIONS TO PERFORM CLUSTERING FOR MULITDIMENSIONAL DATA

The designed approach is employed with two Locality-Sensitive Hashing (LSH) techniques as MinHash and SimHash to manage the high dimensional data by solving near-neighbor search issues. In addition, bisection method is implemented to attain the clustering solution. However, the time complexity remained unaddressed during the clustering process. Besides, the designed approach failed to reduce the computational cost.

The TLRRSC subspace clustering method is developed with the objective of improving the performance on highly corrupted data based on the characteristic information and spatial structures. The TLRR identifies the lowest rank representation through implementing global structure. In addition, the sparse coding is developed with feature spaces for spectral clustering. But, the clustering accuracy was not improved through the TLRR and TLRRSC method.

The DCSTREAM method is developed with the combination of vector model and k-Means divide and conquer approach for big data analysis. The DCSTREAM method is used with offline and online components to cluster high scale datasets with ordinal data types and repetitive data. This in turns, the size and complexity are reduced for performing the clustering. However, STREAM failed to improve the quality of cluster.

5.1 FUTURE DIRECTION

The future direction of the proposed scheme is to enhance the performance of clustering process on multidimensional data. Besides, the subspace clustering techniques are developed to attain the dimensionality reduction with minimized time complexity while grouping the data objects. Another future work is carried out to improve the clustering accuracy with the relevant subspace is identified by considering the data structure of multidimensional data stream.

VI. CONCLUSION

The survival study is carried out to reduce the dimensionality with improved clustering accuracy for multidimensional data. The data structured based clustering techniques are tested with the metrics such as clustering accuracy, space complexity and computational time. The experiments are conducted on existing methods as Locality Sensitive Hashing (LSH) strategy, TLRRSC method and DCSTREAM method. From the simulation results, it is clear that, the Locality Sensitive Hashing (LSH) strategy improves the performance of clustering accuracy for high dimensional data by applying the bisection method to the similarity matrix. Then, the TLRRSC method provides better results on reducing the space complexity through performing the subspace clustering by spatial structures and feature spaces. In addition, the computational time for clustering the high dimensional data is reduced by implementing the DCSTREAM method based on the vector model. Finally, from the result, each method has its own merit and demerit and based on them new methods can be devised as future research work. The further research work is concentrated on performing the subspace clustering of multidimensional data stream with improved clustering accuracy and minimal space complexity.

REFERENCES

[1] Salman Ahmed Shaikh and Hiroyuki Kitagawa, "*Top-k Outlier Detection from Uncertain Data*", International Journal of Automation and Computing, Volume 11, Issue 2, Pages 128-142, April 2014.

[2] Juan Zamora, Marcelo Mendoza and Hector Allende, "*Hashing-based clustering in high dimensional data",* Expert Systems with Applications, Elsevier, Volume 62, Pages 202-211, November 2016,

[3] Weiling Cai, "*A dimension reduction algorithm preserving both global and local clustering structure",* Knowledge-Based Systems, Volume 118, Pages 191–203, February 2017

[4] Ewa Nowakowska, Jacek Koronacki and Stan Lipovetsky, "Dimensionality reduction for data of unknown cluster structure", Information Sciences, Elsevier, Volume 330, Pages 74-87, February 2016,

[5] Yuhua Qian, Feijiang Li, Jiye Liang, Bing Liu, and Chuangyin Dang, "*Space Structure and Clustering of Categorical Data*", IEEE Transactions on Neural Networks and Learning Systems, Volume 27, Issue 10, Pages 2047-2059, October 2016

[6] Can Wang, Xiangjun Dong, Fei Zhou, Longbing Cao and Chi-Hung Chi, "*Coupled Attribute Similarity Learning on Categorical Data*", IEEE Transactions on Neural Networks and Learning Systems, Volume 26, Issue 4, Pages 781-797, April 2015

[7] Yifan Fu, Junbin Gao, David Tien, Zhouchen Lin, and Xia Hong, "*Tensor LRR and Sparse Coding-Based Subspace Clustering*", IEEE Transactions on Neural Networks and Learning Systems, Volume 27, Issue 10, Pages 2120 – 2133, October 2016

[8] Xi Peng, Zhiding Yu, Zhang Yi, and Huajin Tang, "*Constructing the L2-Graph for Robust Subspace Learning and Subspace Clustering*", IEEE Transactions on Cybernetics Volume 47, Issue 4, Pages 1053 – 1066, April 2017

[9] Madjid Khalilian, Norwati Mustapha and Nasir Sulaiman, "*Data stream clustering by divide and conquer approach based on vector model",* Journal of Big Data, Springer, Volume 3, Issue 1, Pages 1-21, 2016

[10] Niwan Wattanakitrungroj, Saranya Maneeroj and Chidchanok Lursinsap "*Versatile Hyper-Elliptic Clustering Approach for Streaming Data Based on One-Pass-Thrown-Away Learning",* Journal of Classification, Springer, Volume 34, Issue 1, Pages 108–147, April 2017

[11] Dominik Olszewski, "*Asymmetric k-Means Clustering of the Asymmetric Self-Organizing Map*", Neural Processing Letters, Springer, Volume 43, Pages 231–253, 2016

[12] Ghatage Trupti B, Patil Deepali E, Takmare Sachin B and Patil Sushama *A "Joint Feature Learning and Clustering Techniques for Clustering High Dimensional Data: A Review*", International Journal of Computer Sciences and Engineering, Volume 4, Issue 3, Pages 54 – 58, 2016

Authors Profile

K. Chitra received her B.Sc Computer Technology from Coimbatore Institute of Technology, Coimbatore, India. She had her M.Sc Computer Communication from Bharathiar University, Coimbatore, India. She holds M.Phil in Computer Science from Bharathiar University, Coimbatore, India. She has 7 years of experience in teaching. She is presently working as an Assistant Professor in Rathnavel Subramaniam College of Arts and Science, Coimbatore. Her research interest includes Data Structures, Data Mining, Big Data Analytics. Now she is pursuing her Ph.D Computer Science in Rathnavel Subramaniam College of Arts and Science, Coimbatore.

Dr. D. Maheswari received her M.Sc Computer Science from Avinashilingam University for Women, Coimbatore, India. She completed her M.Phil and Ph.D degree in Computer Science in Avinashilingam University for Women, Coimbatore, India. She has been working as Head & Research Coordinator in School of Computer Studies in Rathnavel Subramaniam College of Arts and Science, Coimbatore, India. She has published more than 35 papers in International / National Journal and Conferences. Her research work focuses on Image processing and Data Mining. She has 7 years of teaching experience and 7 years of research experience.