Available online at: https://ijact.in

Date of Submission	05/12/2019
Date of Acceptance	07/01/2020
Date of Publication	31/01/2020
Page numbers	3552-3559 (8 Pages)

Cite This Paper: Chitra K, Maheswari D. Map probabilistic density based subspace clustering for dimensionality reduction of big data analytics, 9(1), COMPUSOFT, An International Journal of Advanced Computer Technology. PP. 3552-3559.

ISSN:2320-0790

COMPUSOFT
An International Journal of Advanced Computer Technology

MAP PROBABILISTIC DENSITY BASED SUBSPACE CLUSTERING FOR DIMENSIONALITY REDUCTION OF BIG DATA ANALYTICS

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

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

[2]Head & Research Coordinator, School of Computer Studies, Rathnavel Subramaniam College of Arts and Science, Coimbatore, India

chitra.k@rvsgroup.com[1], maheswari@rvsgroup.com[2]

Abstract: Density based subspace clustering algorithms focus on finding dense clusters of random shape and size. Most of the existing density based subspace clustering algorithms in the literature is less effective and accuracy while taking big dataset as input. In order to overcome such limitations, a MAP Probabilistic Density based Subspace Clustering (MAPPD-SC) Technique is introduced. The MAPPD-SC technique is designed for high dimensional data to improve the clustering accuracy and dimensionality reduction. Initially MAPPD-SC technique designs Map Probabilistic Density Based Subspace Clustering (MPDSC) algorithm with aim of grouping the similar data with higher accuracy and minimum time utilization. During big data clustering, the MAPPD-SC technique applies the maximum a posteriori (MAP) calculation with the goal of clustering more related data together and thereby forming optimal number of clusters with high accuracy. After completing clustering process, the MAPPD-SC technique designs Fusion Tree Data Storage Structure (FTDSS) with objective of storing clustered big data with reduced space complexity. The FTDSS only stores bits values of clustered data in its memory by using fusion tree concepts. This generated bit values of input clustered data takes minimal amount of memory space. From that, proposed MAPPD-SC technique reduces the dimensionality of big data for effective big data analytics. Experimental evaluation of MAPPD-SC technique is carried out on factors such as clustering accuracy, clustering time and false positive rate and space complexity with respect to number of climate data using El Nino Data Set.

Keywords: Big data, Bit values, Fusion Tree Data Storage Structure, Maximum a posteriori (MAP) and Sketch Operation, subspace clustering.

I. INTRODUCTION

Subspace clustering is the method of discovering clusters with objects similar in different subsets of attributes defining subspaces. Recently, subspace clustering is considerable research area as it's applied in diverse applications such as face recognition, speech processing, social media mining, and habitat identification etc. The traditional subspace clustering methods predicts dense clusters in all the subspaces. But, clustering of big data is very complex owing to high probability of frequent clustering information existing in dissimilar subspaces. Hence, problems related to scalability, complexity and accuracy was not addressed. In order to resolve the above mentioned conventional issues, MAPPD-SC technique is developed. The main contributions of MAPPD-SC Technique is formulated as follows,

To get better clustering performance for big data analytics with a lower time complexity as compared to traditional works, Map Probabilistic Density Based Subspace Clustering (MPDSC) is designed in MAPPD-SC technique. MPDSC algorithm is designed based on the density of data points for grouping similar data together and also contains the benefit of generating clusters with random shapes and good scalability. Also, MPDSC algorithm is able to discover the noise data during the clustering process. This supports for MPDSC algorithm for efficient big data clustering process with minimal time complexity as compared to state-of-the-art works.

To minimize the space complexity of big data clustering as compared to conventional works, Fusion Tree Data Storage Structure (FTDSS) is employed in MAPPD-SC technique. FTDSS is a tree data structure which creates an associative array with integer keys through utilizing the constant-time machine word multiplication operation available on many real processors. The designed FTDSS is similar to a B-tree in many aspects. One differentiation between B-tree and FTDSS is the 'B' value. In a conventional B-tree, the 'B' value is a fixed while in a FTDSS the 'B' is a function of 'm'. Another distinction is the time to search a key in a node. The conventional B-tree data storage employ '$O(B)$' operations where FTDSS utilizes '$O(1)$' operations to search a key 'k' in a node. On the contrary to conventional B-tree, each node in FTDSS includes auxiliary information to speed up searching. On the contrary to traditional works, FTDSS utilize word-level parallelism to speed up searches by using individual operations to simulate parallel processing. Furthermore, sketching process is performed in FTDSS to decrease the number of bits while storing the clustered data. From that, MAPPD-SC technique attains minimal space complexity during big data clustering and analytics process as compared to existing works.

The rest of paper is planned as follows. The Section 2 explains the related works. In Section 3, proposed MAPPD-SC technique is explained with the assist of the architecture diagram. In Section 4, experimental settings are described and the experimental result of MAPPD-SC technique is discussed in Section 5. Section 6 shows the conclusion of the paper.

II. RELATED WORKS

Adaptive Multi-view Subspace Clustering (AMSC) was accomplished in [1] to enhance accuracy of high-dimensional data. However, time complexity was very higher. A parallel hierarchical subspace clustering scheme called PAPU was presented in [2] to acquire high clustering efficiency for real-world large-scale datasets with a minimal amount of time complexity. But, false positive rate of big data clustering was not reduced.

An incremental semi-supervised clustering ensemble framework (ISSCE) was presented in [3] to carry out high dimensional data clustering with a minimal time complexity. However, clustering performance was poor. Robust and sparse k-means clustering was accomplished in [4] to increase accuracy. But, time and space complexity of high dimensional data using this method was lower.

Space structure based categorical clustering algorithms (SBC) was introduced in [5] for categorical data that maps categorical objects into Euclidean space. However, the Euclidean distance calculation for mapping the data objects consumed large amount of time. Data-driven similarity learning approach was presented in [6] to calculate the connection among categorical values. Though the clustering accuracy was improved, the space complexity remained unaddressed.

Sparse coding-based subspace clustering method was introduced in [7] with consideration of trait information and spatial structures. However, TLRR and TLRRSC method failed to improve the clustering accuracy. A new method was introduced in [8] to remove the errors effects from projection space than from input space. But, the subspace clustering time was not reduced using subspace clustering and subspace learning algorithms.

A rough set based subspace clustering technique was presented in [9] for finding non-redundant and interesting subspace clusters of better quality. However, computational complexity of this clustering algorithm was more. A novel algorithm was employed in [10] for fast and scalable subspace grouping of high-dimensional data. But, the ratio of number of data imperfectly clustered was very higher.

Fast and effective big data assessment was accomplished in [11] through clustering process with help of complex hierarchical clustering algorithm. A novel algorithm was presented in [12] for grouping related big data with different density with application of a Hadoop platform running MapReduce.

A novel Random forest implementation and optimization was performed in [13] for big data analytics with a lower time complexity. One-pass accelerated MapReduce-based k-prototypes clustering method was employed in [14] to get faster the clustering process.

An intelligent weighting k-means clustering (IWKM) algorithm was utilized in [15] for analysis of high-dimensional multi-view data in big data applications with higher accuracy. Clustering categorical data was carried out in [16] depends on the relational analysis approach and MapReduce with a lower false positive rate.

An efficient predictive analytics system was designed in [17] for big data by using scalable random forest (SRF) algorithm. A fragmented-periodogram approach was introduced in [18] for minimizing error rate of clustering big data time series.

III. MAP PROBABILISTIC DENSITY BASED SUBSPACE CLUSTERING TECHNIQUE

Clustering real world data often impacted with curse of dimensionality as real world data includes of many dimensions. Therefore, Multidimensional data clustering is performed through a density-based approach. The conventional density based clustering techniques does not provide higher clustering accuracy for big data analytics. In order to addresses this drawback, a MAP Probabilistic Density based Subspace Clustering (MAPPD-SC) Technique is proposed. The architecture diagram of proposed MAPPD-SC technique is depicted in below Figure 1.

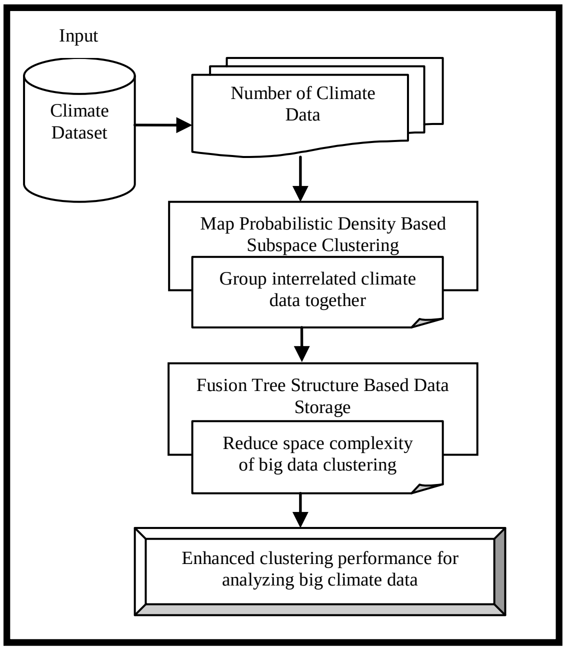
Input
Climate Dataset
Number of Climate Data
Map Probabilistic Density Based Subspace Clustering
Group interrelated climate data together
Fusion Tree Structure Based Data Storage
Reduce space complexity of big data clustering
Enhanced clustering performance for analyzing big climate data

Figure 1: Architecture Diagram of MAPPD-SC technique for Analytics of Climate Data

As demonstrated in above figure, MAPPD-SC technique initially obtains big climate dataset (i.e. El Nino Data Set) as input. After getting the input, MAPPD-SC technique applies Map Probabilistic Density Based Subspace Clustering (MPDSC) with objective of grouping the huge size of climate data into related clusters with lower false positive rate.

Consequently, MAPPD-SC technique employs Fusion Tree Data Storage Structure (FTDSS) with aiming at storing the clustered big climate data with lower space complexity.

A. Map Probabilistic Density Based Subspace Clustering

The conventional density based subspace clustering method was suffered by density divergence problem that impacts the clustering accuracy of big data which results in higher false positive rate. In order to addresses this limitations, a novel clustering technique called Map Probabilistic Density Based Subspace Clustering (MPDSC) is proposed in MAPPD-SC technique. On the contrary to state-of-the-art works, MPDSC is introduced by combining maximum a posteriori (MAP) concept in density based subspace clustering algorithm to find out the optimal clusters that form on a different subspaces. The proposed MPDSC algorithm is better in handling huger volume of data than other existing works. The MAP concept is applied in MPDSC helps to determine maximization probability for each climate data in a given dataset to become a member of the cluster.

The MPDSC algorithm operates by discovering areas where data points are concentrated and where they are separated by areas that are empty or sparse. In MPDSC algorithm, Data Points that are not member of a cluster are considered as noise. The MPDSC algorithm identifies the dense region by grouping data points together that are closed to each other according to a distance calculation. The proposed MPDSC algorithm employs the concept of density reach-ability and density connectivity.

Density Reachability: A data point 'd_1' is said to be density reachable from a data point 'x' when data 'd_1' is within distance 'ω' from data point 'x' and 'x' contains sufficient number of data points in its neighbors which are within distance 'ω'.

Density Connectivity: A data point 'd_1' and 'x' are said to be density connected if there exist a data point 'y' which includes sufficient number of data points in its neighbors and both the data 'd_1' and 'x' are within the distance 'ω'.

Let us consider an input big climate dataset 'DS' contains a numbers of climate data denoted as '$d_{1,}d_2, d_3, \ldots\ldots d_m$'. Here, '$M$' refers to the total number of climate data in an input dataset. After taking big dataset as input, MPDSC algorithm arbitrarily choose the data point 'd_1' from big climate dataset. Followed by, MPDSC algorithm find outs the neighborhood of this data point 'd_1' by using the distance measurement 'ω'. All the data points that are within the distance 'ω' are considered as neighborhood in MPDSC algorithm. If data points 'd_1' contains sufficient neighborhood then clustering process is started where interrelated climate data are grouped into corresponding clusters by using maximum a posteriori (MAP) calculation and marked as visited. Otherwise, the chosen data point 'd_1' is labeled as noise. If a taken data point is selected to be a part of the cluster then its neighborhood is also the part of the cluster.

Let us assume the number of clusters '$c_1, c_2, c_3, \ldots . c_N$'. During the big data clustering process, MPDSC algorithm calculates the expected probability between each data point 'd_i' and cluster 'c_i' using below,

$$Exp\,\{P\ (c_i|d_i)\} = \log\left(\prod_{i=1}^{N} \frac{e^{-\frac{1}{2}\frac{(d_i-a)^2}{b^2}}}{\sqrt{2\pi b^2}}\right) \tag{1}$$

From above mathematical formula (1), '$Exp\ \{P\ (c_i|d_i)\}$' represents the expected probability of the climate data 'd_i' to be a member of cluster 'c_i'. Here, 'a' denotes a mean value of cluster and 'b' refers to a variance between the cluster and input climate data. Followed by, the MPDSC algorithm computes the maximization probability for each data point 'd_i' in an input dataset to become a part of the cluster 'c_i' using maximum a posteriori (MAP) calculation using below,

$$\vartheta_{MAP} = \arg\max\ Exp\,\{P\ (c_i|d_i)\} \tag{2}$$

$$\vartheta_{MAP} = \arg\max\ \log\left(\prod_{i=1}^{N} \frac{e^{-\frac{1}{2}\frac{(d_i-a)^2}{b^2}}}{\sqrt{2\pi b^2}}\right) \tag{3}$$

From the above mathematical expression (2) and (3), 'ϑ_{MAP}' point outs the maximum a posteriori function that enhance the expected probability between data point 'd_i' and cluster 'c_i' based on distance estimation to exactly group the more similar climate data together. The above processes of MPDSC is repeated until the all data in an input dataset is grouped into the different clusters and also marked as visited. As a result, the MPDSC algorithm improves clustering performance of big climate data with a minimal amount of time complexity.

The algorithmic processes of MPDSC is presented in below,

```
// Map Probabilistic Density Based Subspace Clustering Algorithm
Input: Big Climate Dataset '$DS$: $d_1, d_2, d_3, \ldots . d_m$';
Output: Achieve higher clustering accuracy for big climate data
Begin
Step 1:    For each input dataset 'DS'
Step 2:    Consider number of clusters '$c_1, c_2, c_3, \ldots . c_N$'
Step 3:       Randomly select data point from 'DS'
Step 4:       Find the neighborhood of this data point '$d_i$' using distance measurement '$\omega$'
Step 5:       If there is sufficient neighborhood around this data point then
Step 6:          Clustering process is begins
Step 7:          Compute maximum a posteriori using (3)
Step 8:          Group data point '$d_i$' into related cluster '$c_i$' using MAP
Step 9:          Data point '$d_i$' is marked as visited
Step 10:      Else
Step 11:         Data point '$d_i$' is considered as noise
Step 12:      End If
Step 13:      This process continues until all data points are marked as visited
Step 14:End For
End
```

Algorithm 1: Map Probabilistic Density Based Subspace Clustering

B. Fusion Tree Data Storage Structure

After clustering process, MAPPD-SC technique designs Fusion Tree Data Storage Structure (FTDSS) for efficiently storing clustered climate data with a lower space complexity. The FTDSS used associative array in a known universe size and also which is suitable while universe size (input data) is large. The proposed FTDSS is similar to a B-tree with degree 'w^c'. Here, 'w' refers the word size and 'c' is constant smaller than '1'. This represents a height of tree will be '$\log_w m$' in which 'm' denotes a number of data stored in the tree. A key operation performed in FTDSS is sketch which is employed to compress 'w' bit keys. This helps for FTDSS to take a minimal amount of memory space for effective big data storage as compared to existing works. The Data Storage Structure of FTDSS is depicted in below Figure 2.

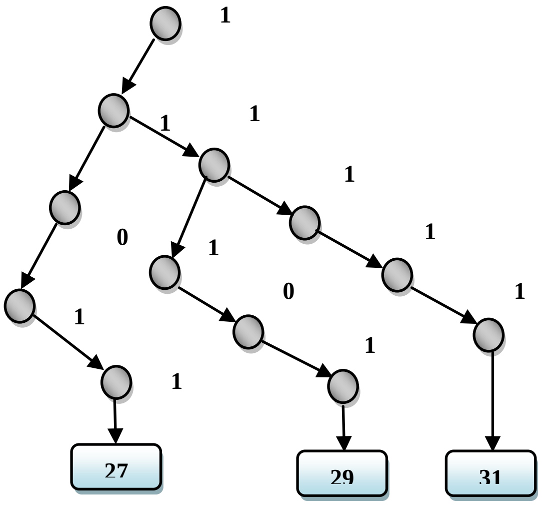

Figure 2: Data Storage Structure of Fusion Tree

As presented in the above figure 2, FTDSS combines two or more bits value of input data together to form space efficient data storage for big climate data analytics. For example, let consider a three numbers 27, 29, 31 to be stored on fusion tree. Then, sketch operation gives the bits value for these number data as follows, Sketch (27) = sketch (11011) = 11011, Sketch (29) = sketch (11101) = 11101, Sketch (31) = sketch (11111) = 11111. Subsequently, bit values are stored on fusion tree as shown in above figure 2.

Let us consider FTDSS takes number of clustered data as input which is denoted as '$D_i = D_1, D_2, \ldots D_m$'. The FTDSS is designed to improve storage efficiency and thereby minimizing the space complexity. The FTDSS initially creates the fusion tree with a number of nodes to store clustered climate data which is mathematically represented as follows,

$$FT \rightarrow \{n_1, n_2, .. n_n\} \tag{4}$$

From the above equation (4), 'n_n' signifies the number of nodes designed in fusion tree. After constructing the fusion tree structure, FTDSS stores clustered data in the form of bits through sketch operation with aiming at reducing space complexity during data storage process. For each input clustered climate data 'D_i', then FTDSS produce bits value. From that, the sketch operation 'SO' is mathematically carried out as follows,

$$SO \rightarrow Bits(D_i) \tag{5}$$

From the above mathematical expression (5), '$Bits(D_i)$' indicates the generated bits of input clustered data 'D_i'. The sketch operation generates unique bits value for each input climate data. After that, FTDSS stores bit values of data in it fusion tree using below equation,

$$Insert\ (Bits(D_i)) \rightarrow FT(n_i) \tag{6}$$

From the above mathematical representation (6), FTDSS stores input clustered data with a lower amount of memory consumption. Here, '$insert$' operation assists for FTDSS to store bit values of data where '$FT(n_i)$' denotes nodes in fusion tree. The designed FTDSS also allow delete operation to remove the stored data on it storage. The data deletion operation is carried out using below formula,

$$Delete\ (Bits(D_i)) \rightarrow FT(n_i) \tag{7}$$

From the above mathematical formula (7), '$delete$' supports for FTDSS to significantly eliminate bit value of input data '$Bits(D_i)$' that stored on fusion tree '$FT(n_i)$'. By using the above two operations i.e. insertion and deletion, the proposed FTDSS improves the storage efficiency of big climate data analytics with a minimal space and time complexity as compared to state-of-the-art works.

The algorithmic processes of FTDSS are presented in below.

```
// Fusion Tree Data Storage Structure Algorithm
Input: Number Of Clustered Data '$D_i = D_1, D_2, \ldots D_m$'
Output: Efficient storage of Clustered Big Data with minimal space complexity
Step 1: Begin
Step 2:   Take number of clustered data as input.
Step 3:   Create a structure FusionTree ‘$FT$’
Step 4:   Construct a function init() for creating the nodes using (4)
Step 5:   For each clustered data ‘$D_i$’
Step 6:       Employ sketch operation
Step 7:       Generate bit values using (5)
Step 8:       Design a function insert() to insert the bit values of data into the tree using (6)
Step 9:   End for
Step 10:  Construct a function delete() to delete the bit values of data from the tree using (7)
Step 11:End
```

Algorithm 2: Fusion Tree Data Storage Structure

IV. EXPERIMENTAL SETTINGS

In order to estimate the clustering performance, both the proposed MAPPD-SC technique and conventional Adaptive Multi-view Subspace Clustering (AMSC) [1] and a parallel hierarchical subspace clustering scheme called (PAPU) [2] are implemented in Java language using big El Nino Data Set. This is a large volume of climate data get from UCI machine learning repository [21] which contains oceanographic and surface meteorological data with 178080 numbers of instances and 12 attributes. The proposed MAPPD-SC technique takes diverse number of big climate data in the range of 1000 to 10000 from El Nino Data Set to conduct experimental process. The performance of MAPPD-SC technique is evaluated in terms of clustering accuracy, clustering time and false positive rate and space complexity with respect to various number of input climate data. The experimental result of MAPPD-SC technique is compared against existing AMSC [1] and PAPU [2].

V. RESULTS

The experimental result of MAPPD-SC technique is discussed in this section. The effectiveness of MAPPD-SC technique is compared against with traditional Adaptive Multi-view Subspace Clustering (AMSC) [1] and a parallel hierarchical subspace clustering scheme called (PAPU) [2] with helps of the tables and graphs using below metrics.

Case 1: Impact of Clustering Accuracy

In MAPPD-SC technique, Clustering accuracy ‘CA’ estimates the ratio of number of climate data that exactly grouped to the total number of climate data considered for experimental process. The clustering accuracy is calculated mathematically as follows,

$$CA = \frac{\tau_{EC}}{\tau} * 100 \qquad (8)$$

From the above mathematical equation (8), ‘τ_{EC}’ refers to number of exactly clustered climate data in which ‘τ’ denotes a total number of climate data. The clustering accuracy is computed in terms of percentage (%).

Sample Calculation:

Proposed MAPPD-SC: Number of climate data correctly clustered is 900 and the total number of climate data is 1000. Then the clustering accuracy is obtained as follows,

$$CA = \frac{900}{1000} * 100 = 90\ \%$$

Existing AMSC: Number of climate data accurately clustered is 840 and the total number of climate data is 1000. Then the clustering accuracy is estimated as follows,

$$CA = \frac{840}{1000} * 100 = 84\ \%$$

Existing PAPU: Number of climate data precisely clustered is 870 and the total number of climate data is 1000. Then the clustering accuracy is acquired as follows,

$$CA = \frac{870}{1000} * 100 = 87\ \%$$

The experimental result analysis of clustering accuracy for big data analytics with respect to various numbers of climate data in the range of 1000-10000 using three methods namely proposed MAPPD-SC technique and existing AMSC [1] and PAPU [2] is depicted in Figure 3.

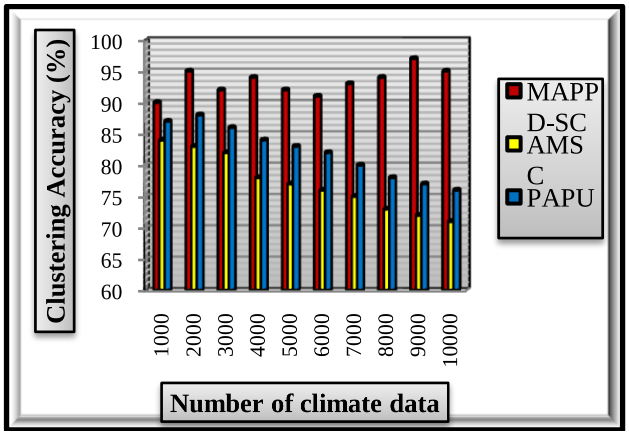

Figure 3: Experimental result of clustering accuracy versus different number of climate data

The proposed MAPPD-SC technique renders improved clustering accuracy for big climate data while increasing the number of data as input when compared to traditional AMSC [1] and PAPU [2]. This is owing to process of Map Probabilistic Density Based Subspace Clustering (MPDSC) in proposed MAPPD-SC technique on the contrary to state-of-the-art works as where it applied maximum a posteriori (MAP) calculation to find out maximization probability for each input climate data to become a part of the cluster. Based on the measured value of maximum a posteriori, finally proposed MAPPD-SC technique precisely clusters all the input climate data into consequent clusters with higher accuracy. This assists for proposed MAPPD-SC technique to enhance the ratio of number of climate data that are perfectly clustered when compared to other existing AMSC [1] and PAPU [2]. Hence, proposed MAPPD-SC technique increases the clustering accuracy of big climate data analytics by 21 %

and 14 % as compared to conventional AMSC [1] and PAPU [2] respectively.

Case 2: Impact of Clustering Time

In MAPPD-SC technique, Clustering Time 'CT' calculates the time required to group similar climate data together. The clustering time is mathematically obtained as follows,

$$CT = \tau * T(CSD) \tag{9}$$

From the above mathematical formula (9), '$T(CSD)$' signifies a time utilized to cluster a single climate data and 'τ' denotes a total number of climate data considered. The clustering time is estimated in terms of milliseconds (ms).

Sample Calculation for Clustering Time:

Proposed MAPPD-SC: time used to cluster one climate data is 0.031 ms and the total number of climate data is 1000. Then the clustering time is computed as follows,

$$CT = 1000 * 0.031 = 31\ ms$$

Existing AMSC: time employed to cluster one climate data is 0.04 ms and the total number of climate data is 1000. Then the clustering time is calculated as follows,

$$CT = 1000 * 0.04 = 40\ ms$$

Existing PAPU: time taken to cluster one climate data is 0.042 ms and the total number of climate data is 1000. Then the clustering time is acquired as follows,

$$CT = 1000 * 0.042 = 42\ ms$$

The comparative result analysis of clustering time to analyze the large size of data with respect to different numbers of climate data in the range of 1000-10000 using three methods namely proposed MAPPD-SC technique and conventional AMSC [1] and PAPU [2] is demonstrated in Figure 4.

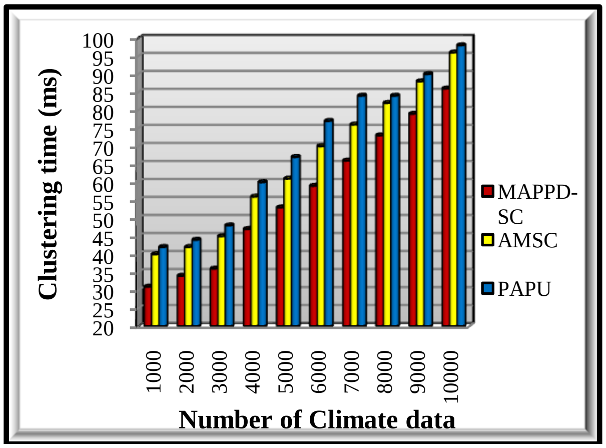

Figure 4: Experimental result of clustering time versus different number of climate data

As presented in above graphical representation, proposed MAPPD-SC technique gives minimal amount of clustering time in order to effectively analyze large climate data with increasing the number of data as input when compared to conventional AMSC [1] and PAPU [2]. This is because of process of Map Probabilistic Density Based Subspace Clustering (MPDSC) in proposed MAPPD-SC technique on the contrary to existing works. On contrary to existing works, MPDSC employs maximum a posteriori (MAP) computation to identify maximization probability of each input climate data to become a cluster member. By using this maximum a posteriori (MAP) computation concept, proposed MAPPD-SC technique accurately groups all the interrelated climate data together with a minimal amount of time complexity. This supports for proposed MAPPD-SC technique to diminish the time used to cluster similar climate data together when compared to other traditional AMSC [1] and PAPU [2]. Therefore, proposed MAPPD-SC technique minimizes clustering time of big climate data examination by 15 % and 20 % as compared to traditional AMSC [1] and PAPU [2] respectively.

Case 3: Impact of False Positive Rate

In MAPPD-SC technique, False Positive Rate 'FPR' measured as ratio of number of climate data incorrectly grouped to the total number of climate data. The false positive rate is mathematically estimated as follows,

$$FPR = \frac{\tau_{WC}}{\tau} * 100 \tag{10}$$

From the above mathematical expression (10), 'τ_{WC}' signifies a number of climate data wrongly clustered and 'τ' point outs a total number of climate data. The false positive rate is calculated in terms of percentage (%).

Sample Calculation for False Positive Rate:

Proposed MAPPD-SC: number of climate data inaccurately grouped is 100 and the total number of climate data is 1000. Then the false positive rate is computed as follows,

$$FPR = \frac{100}{1000} * 100 = 10\ \%$$

Existing AMSC: number of climate data imperfectly clustered is 160 and the total number of climate data is 1000. Then the false positive rate is obtained as follows,

$$FPR = \frac{160}{1000} * 100 = 16\ \%$$

Existing PAPU: number of climate data mistakenly clustered is 130 and the total number of climate data is 1000. Then the false positive rate is determined as follows,

$$FPR = \frac{130}{1000} * 100 = 13\ \%$$

The performance result of false positive rate is obtained during clustering process to examine the huge size of data based on varied numbers of climate data in the range of 1000-10000 using three methods namely proposed MAPPD-SC technique and traditional AMSC [1] and PAPU [2] is presented in Figure 5.

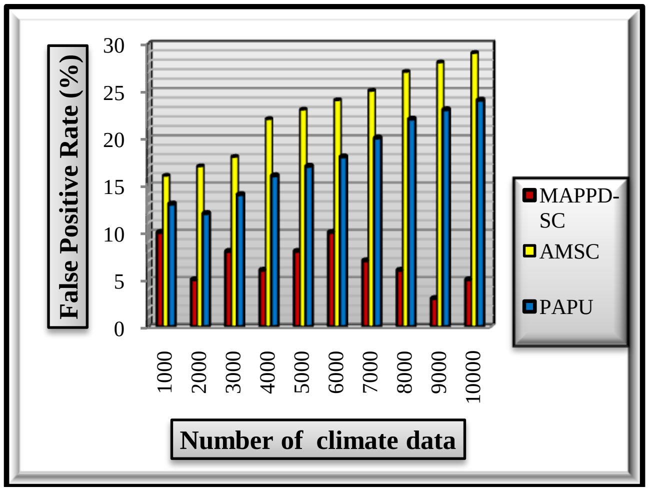

Figure 5: Experimental result of false positive rate versus different number of climate data

Figure 5 demonstrates impact of false positive rate with respect to dissimilar numbers of climate data using three proposed MAPPD-SC technique and conventional AMSC [1] and PAPU [2]. As shown in above graphical demonstration, proposed MAPPD-SC technique presents lower false positive rate in order to accurately examine huge size of climate data with increasing the number of data as input as compared to state-of-the-art AMSC [1] and PAPU [2]. This is owing to process of Map Probabilistic Density Based Subspace Clustering (MPDSC) in proposed MAPPD-SC technique on the contrary to traditional works. The proposed MAPPD-SC technique enhances the clustering accuracy of big data based on the density of data points and also generating clusters with random shapes and good scalability. From that, proposed MAPPD-SC technique improves accuracy of big data clustering. This aid for proposed MAPPD-SC technique to lessen the ratio of number of climate data mistakenly grouped when compared to other existing AMSC [1] and PAPU [2]. Therefore, proposed MAPPD-SC technique reduces the false positive rate of big climate data analysis by 68 % and 59 % as compared to conventional AMSC [1] and PAPU [2] respectively.

Case 4: Impact of Space Complexity

In MAPPD-SC technique, Space Complexity (SC) determines memory space taken for storing clustered climate data. The space complexity is mathematically computed as follows,

$$SC = \tau * M(SSD) \qquad (11)$$

From the above mathematical representation (11), '$M(SSD)$' designates memory space needed for storing a single climate data and 'τ' signifies a total number of climate data. The space complexity is evaluated in terms of Megabytes (MB).

Sample Calculation for Space Complexity

Proposed MAPPD-SC: total number of climate data are 1000 and the amount of memory utilized to store a single climate data is 0.038 MB, then space complexity is estimated as follows,

$$SC = 1000 * 0.038\ MB = 38\ MB$$

Existing AMSC: total number of climate data are 1000 and the memory space required to store the single climate data is 0.045 MB, then space complexity is computed as follows,

$$SC = 1000 * 0.045\ MB = 45MB$$

Existing PAPU: total number of climate data are 1000 and the memory space taken to store the single climate data is 0.049 MB, then space complexity is evaluated as follows,

$$SC = 1000 * 0.049\ MB = 49\ MB$$

The space complexity result is acquired during the big data clustering process along with diverse numbers of climate data in the range of 1000-10000 using three methods namely proposed MAPPD-SC technique and traditional AMSC [1] and PAPU [2] is depicted in Figure 6.

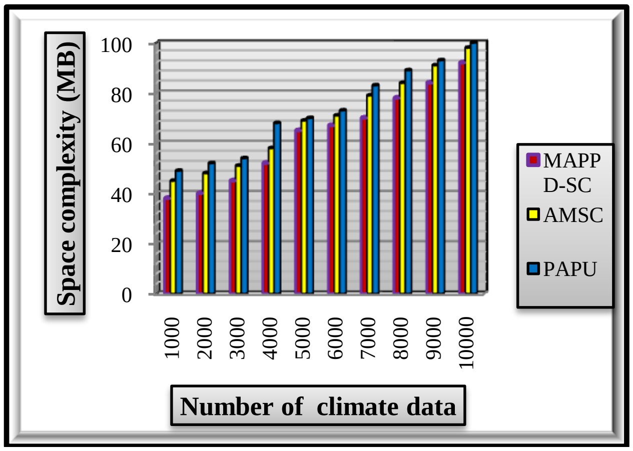

Figure: 6 Experimental result of space complexity versus different number of climate data

As demonstrated in above graphical depiction, proposed MAPPD-SC technique gives minimal amount of memory space for effective analytics of big data with increasing the number of data as input as compared to conventional AMSC [1] and PAPU [2]. This is owing to process of Fusion Tree Data Storage Structure (FTDSS) in proposed MAPPD-SC technique on the contrary to state-of-the-art works. With the concepts of FTDSS, proposed MAPPD-SC technique unites two or more bits value of input data together to design space efficient data storage during big climate data analytics process. This help for proposed MAPPD-SC technique to minimize the memory space employed to store clustered climate data when compared to other existing AMSC [1] and PAPU [2]. For that reason, proposed MAPPD-SC technique decreases the space complexity of big climate data assessment by 10 % and 15 % as compared to conventional AMSC [1] and PAPU [2] respectively.

VI. CONCLUSION

The MAPPD-SC technique is proposed with the purpose of increasing the clustering accuracy of big climate data with a minimal time and space complexity. The aim of MAPPD-SC technique is obtained with the application of Map Probabilistic Density Based Subspace Clustering (MPDSC) and Fusion Tree Data Storage Structure (FTDSS) on the contrary to state-of-the-art works. The proposed MAPPD-SC technique improves the ratio of number of climate data that are correctly grouped when compared to conventional works. As well as, proposed MAPPD-SC technique lessen the amount of time desired to cluster same type of climate data when compared to other traditional works. Furthermore, proposed MAPPD-SC technique decreases the memory space needed to store clustered climate data. The experimental result shows that proposed MAPPD-SC technique presents better big data clustering performance in terms of accuracy, time and false positive rate and space complexity for analyzing big climate data as compared to state-of-the-art works.

VII. REFERENCES

[1] Fei Yan, Xiao-dong Wang, Zhi-qiang Zeng, Chao-qun Hong, “Adaptive Multi-view Subspace Clustering for High-dimensional Data”, Pattern Recognition Letters, Elsevier, Pages 1-10, 2019

[2] Ning Pang, Jifu Zhang, Chaowei Zhang, and Xiao Qin, “Parallel Hierarchical Subspace Clustering of Categorical Data”, IEEE Transactions on Computers, Volume 68, Issue 4, Pages 542 – 555, April 2019

[3] hiwen Yu ; Peinan Luo ; Jane You ; Hau-San Wong ; Hareton Leung ; Si Wu ; Jun Zhang ; Guoqiang Han “Incremental Semi-Supervised Clustering Ensemble for High Dimensional Data Clustering”, IEEE Transactions on Knowledge and Data Engineering, Volume 28, Issue 3, Pages 701 – 714, March 2016

[4] Šárka Brodinová, Peter Filzmoser, Thomas Ortner, Christian Breiteneder, Maia Rohm, “Robust and sparse k-means clustering for high-dimensional data”, Advances in Data Analysis and Classification, Springer, Pages 1–28, 2019

[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] B. JayaLakshmi, M.Shashi, K.B.Madhuri, “A rough set based subspace clustering technique for high dimensional data”, Journal of King Saud University - Computer and Information Sciences, Elsevier, Pages 1-6, 2017

[10] Amardeep Kaur & Amitava Datta, “A novel algorithm for fast and scalable subspace clustering of high-dimensional data”, Journal of Big Data, Springer, Volume 2, Issue 17, Pages 1-24, 2015

[11] Michele Ianni' Elio Masciari'Giuseppe M.Mazzeo' Mario Mezzanzanica' Carlo Zaniolo, “Fast and effective Big Data exploration by clustering”, Future Generation Computer Systems, Elsevier, Volume 102, Pages 84-94, 2019

[12] Safanaz Heidari, Mahmood Alborzi, Reza Radfar, Mohammad Ali Afsharkazemi, Ali Rajabzadeh Ghatari, “Big data clustering with varied density based on MapReduce”, Journal of Big Data, Springer, Volume 6, Issue 7, December 2019

[13] Victor M. Herrera, Taghi M. Khoshgoftaar, Flavio Villanustre, Borko Furht, “Random forest implementation and optimization for Big Data analytics on LexisNexis’s high performance computing cluster platform”, Springer, Journal of Big Data, Volume 6, Issue 68, Pages 1-36, December 2019

[14] Mohamed Aymen Ben HajKacem, Chiheb-Eddine Ben N’cir, Nadia Essoussi, “One-pass MapReduce-based clustering method for mixed large scale data”, Journal of Intelligent Information Systems, Springer, Volume 52, Issue 3, Pages 619–636, June 2019

[15] Qian Tao, Chunqin Gu, Zhenyu Wang, Daoning Jiang, “An intelligent clustering algorithm for high-dimensional multi-view data in big data applications”, Neurocomputing, Elsevier, Pages July 2019

[16] Yasmine Lamari, Said Chah Slaoui, “Clustering categorical data based on the relational analysis approach and MapReduce”, Journal of Big Data, Springer, Volume 4, Issue 28, Pages December 2017

[17] Myat Cho MonOo, ThandarThein, “An efficient predictive analytics system for high dimensional big data”, Journal of King Saud University - Computer and Information Sciences, Elsevier, September 2019

[18] Jorge Caiado, Nuno Crato, Pilar Poncela, “A fragmented-periodogram approach for clustering big data time series”, Advances in Data Analysis and Classification, Springer, Pages 1–30, June 2019