The International Journal of Interdisciplinary Organizational Studies
ISSN: 2324-7649 (Print), ISSN: 2324-7657 (Online)
Volume 19, Issue 1, July-December, 2024

DEEP LEARNED MAPREDUCE MEAN SHIFT CLUSTERING FOR BIG E-COMMERCE DATA ANALYTICS

K.M.Padmapriya[*1] , Dr. K.Anandapadmanabhan [2]

[*1] Assistant professor in Computer Science, SSM College of Arts & Science, Tamilnadu ,India.

[2] DEAN, Sri Vasavi College (SF Wing), Erode, Tamilnadu. India.

ABSTRACT

Clustering of big data has attained greater significance recently. Few research works have been developed in existing for grouping similar data. But, clustering accuracy using conventional algorithm was not adequate when taking big e-commerce data as input. In addition, time complexity of big data clustering was also minimal. In order to resolve such limitations, Deep Learned MapReduce Mean Shift Clustering (DLMMSC) technique is introduced. The proposed DLMMSC technique contains three layers such as input, hidden and output layers to efficiently group the large volume of data in given dataset with higher accuracy and minimal time. The DLMMSC technique initially gets the big data as input in the input layer and sent it to the hidden layer. In DLMMSC technique, three hidden layers are used in order to deeply analyze the input big e-commerce data by applying the MapReduced Mean Shift Clustering concepts. Through a deep analysis, DLMMSC technique accurately groups the similar e-commerce data together into different clusters. At last, output layer produces the optimal clustering result of big e-commerce results. By the effectual clustering of data, DLMMSC technique increases the big e-commerce data analytics performance as compared to state-of-the-art works. The DLMMSC technique conducts the experimental evaluation using parameters such as clustering accuracy, time complexity and error rate and space complexity. The experimental result shows that the DLMMSC technique is able to improve the clustering accuracy and also minimizes the time complexity of big e-commerce data analytics when compared to state-of-the-art works.

Keywords: Big E-Commerce Data, Cluster Mean, Deep Neural Learning, Hidden Layer, Input Layer, MapReduced Mean Shift Clustering, Output Layer

1. INTRODUCTION

In data mining, clustering is a technique that groups the similar data points together. In theory, data that are in the same group contains related properties, while data in different groups includes highly dissimilar properties. Clustering method is employed in many fields. Besides, e-commerce data includes product details and transaction information, average order value, ecommerce conversion rate, time to purchase, and other data. The clustering of E-commerce data helps to analyze purchase activity on website. Several research works have been designed for clustering data. However, clustering accuracy of conventional algorithm was poor when considering big e-commerce data as input. Moreover, time complexity involved during clustering process of big data is also very higher. In order to solve these limitations, DLMMSC technique is proposed in this research work by using deep neural learning and MapReduce and Mean Shift Clustering.

A MapReduce-based hierarchical subspace-clustering algorithm called PAPU was employed in [1] to obtain enhanced clustering efficiency for real-world large-scale datasets. However, the time complexity using PAPU algorithm was very higher. A fast clustering algorithm called (MUFOLD-CL) was applied in [2] to computational complexity of big data clustering. But, number of data wrongly clustered was more.

A Fully Recurrent Deep Neural Learning based X-means Data Clustering (FRDNL-XDC) technique was designed in [3] for grouping same type of uncertain data with a lower time complexity. However, the clustering accuracy was minimal. A survey of different clustering techniques designed for big data using deep learning was presented in [4].

A double deep auto encoder structure was constructed in [5] for grouping the distributed and heterogeneous datasets. But, time complexity taken for big data clustering was not reduced. A robust multi-objective subspace clustering (MOSCA) algorithm was employed in [6] with aim of improving the accuracy of high-dimensional data. However, the error rate involved during clustering process was more.

A consensus fuzzy clustering algorithm was introduced in [7] in order to minimize the time of big data clustering. But, clustering performance was not sufficient. The fuzzy c-means approach was applied in [8] to cluster the same types of big data in the internet of things with minimal false positive rate. However, clustering accuracy was not enhanced.

Accelerated MapReduce-based k-prototypes (AMRKP) clustering was presented in [9] to obtain better clustering result for mixed large scale data. But, computational time required for clustering the big data was higher. A secure weighted possibilistic c-means algorithm (SWPCM) was employed in [10] to get higher scalability for grouping the big data. However, clustering time was taken for big data was very higher.

In order to addresses the above mentioned existing problems in big data clustering and thereby enhancing the e-commerce data analytics performance, DLMMSC technique is developed. The main contributions of DLMMSC technique are described as,

- ✓ To achieve better clustering accuracy for clustering the big e-commerce data as compared to conventional works, DLMMSC technique is introduced using deep neural learning and MapReduce and Mean Shift Clustering.
- ✓ To minimize the time needed for clustering the big -commerce data as compared to state-of-the-art works, deep neural learning concepts is applied in DLMMSC technique where it contains input, hidden and output layers in order to deep analyze big data and thereby precisely cluster the data with a minimal time.
- ✓ To reduce the number of data inaccurately clustered as compared to existing works, Mean Shift Clustering is designed in DLMMSC technique as it appropriate for real data analysis. In addition to that, Mean Shift Clustering does not assume any prior shape on data clusters. Also, Mean Shift Clustering is suitable for processing large-scale datasets.

The rest of paper structure is constructed as follows. In Section 2, proposed methodology is explained with the assist of the architecture diagram. In Section 3, Simulation settings are

described and the comparative result of DLMMSC technique is discussed in Section 4. Section 5 shows the literature survey. Section 6 depicts the conclusion of the paper.

2. METHODOLOGY

The Deep Learned MapReduce Mean Shift Clustering (DLMMSC) technique is developed in order to increase big e-commerce data analytics performance via clustering with higher accuracy and minimal time. The DLMMSC technique is introduced by combining the deep neural learning and MapReduce and Mean Shift Clustering on the contrary to conventional works. The DLMMSC technique takes big e-commerce dataset as input. This dataset comprises of huge volume of e-commerce data such as product purchase details and transaction information, average order value, ecommerce conversion rate, time to purchase, etc. To deeply analyze input big e-commerce data with a lower amount of time complexity on the contrary to state-of-the-art works, DLMMSC technique is proposed using deep neural learning. Therefore, proposed DLMMSC technique is inspired by the structure and function of the brain called artificial neural networks. In DLMMSC technique, MapReduced Mean Shift Clustering is designed for accurately clustering the large volume of input e-commerce data with a minimal error rate. The MapReduced Mean Shift Clustering is developed by integrating MapReduce function in conventional mean shift clustering algorithm.

On the contrary to existing techniques, MapReduced Mean Shift Clustering is utilized in DLMMSC technique as it suitable for processing of huge e-commerce data. Besides to that, MapReduced Mean Shift Clustering is used in DLMMSC technique does not need prior knowledge of the number of clusters, and does not constrain the shape of the clusters on two-dimensional space on the contrary to state-of-the-art works. Also, MapReduced Mean Shift Clustering is employed in DLMMSC technique is a centroid-based algorithm that works by mapping the each data to the mean of clusters in two-dimensional space. These mapped data are then filtered in a reduce phase and thereby eliminates near-duplicate data and constructs final set of optimal clusters with higher accuracy. As a result, DLMMSC technique significantly groups the relevant e-commerce data together into dissimilar clusters with a minimum amount of time. From that, DLMMSC technique improves the clustering performance of big e-commerce data analytics as compared to conventional works. The architecture diagram of DLMMSC technique is shown in below Figure 1.

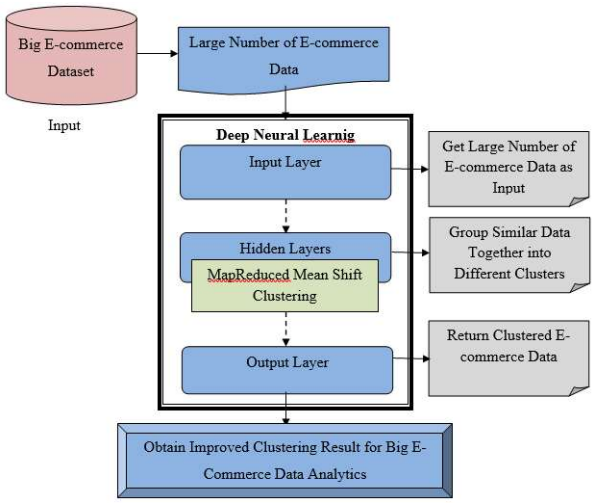

Figure 1 Architecture Diagram of DLMMSC Technique for Big Data Analytics

Figure 1 demonstrates the overall flow processes of DLMMSC technique for efficient big e-commerce data analytics through clustering. As depicted in the above architecture diagram, a DLMMSC technique initially acquires the big e-commerce dataset i.e. Amazon-E-commerce-Data-set as input where it contains an enormous amount of data denoted as '$D_1, D_2, D_3, \ldots . D_N$'. Then, DLMMSC technique employs deep neural learning to group the same kind of data in input big dataset with a lower amount of time. In deep neural learning, input layer obtains input huge volume of e-commerce data and consequently forward it to the hidden layers.

In hidden layers, DLMMSC technique applies MapReduced Mean Shift Clustering concept in order to exactly cluster the similar e-commerce data together in given dataset. On the contrary to conventional clustering algorithm, the MapReduced Mean Shift Clustering does not define the number of clusters in advance. Finally, the output layer generates the clustered E-commerce data results. By the effective clustering of E-commerce data, proposed DLMMSC technique significantly carry outs big data analytics process as compared to state-of-the-art works. The exhaustive explanation about the DLMMSC technique is described in below.

The DLMMSC technique is devloped based on the structure and function of the brain called artificial neural networks with several layers between the input and output layers. Besides, DLMMSC structure is a feedforward neural network. In DLMMSC technique, data flows from the input layer to the output layer as shown in below Figure 2.

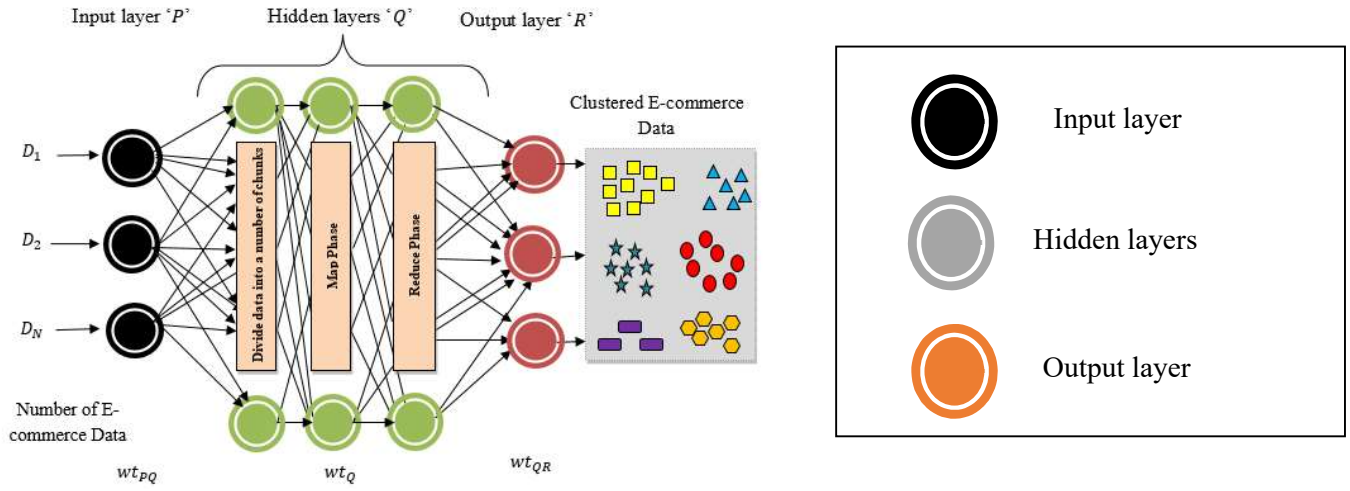

Figure 2 Structure of DLMMSC technique for Big Data Clustering

Figure 2 presents the structure of DLMMSC technique to increases the clustering perfrmance of big data analytics. The DLMMSC structure includes of three main parts namely input, hidden, output layer where these layers are connected with each other. In DLMMSC structure, input layer comprises of neurons which obtains an huge size of data as input and forward taken input to the second layer called as the hidden layers. In proposed DLMMSC technique, three hidden layers are utilized to deep analyze the input e-ccommerce data using MapReduced Mean Shift Clustering for performing big data clustering process with higher accuracy. At last, the output layer gives the clustered result.

Let us consider an input big e-commerce dataset with massive amount of data is representd as '$DS = D_1, D_2, D_3, \ldots . D_N$'. Here, '$N$' indicates the total number of data presented in given dataset. The DLMMSC technique takes number of data points 'D_i' as input. After getting the input, the input layer 'P' in DLMMSC technique combine the data 'D_i'with weights ' wt_{PQ}' and bias term 'b_j'. Consequently, the neurons activity in the input layer '(P)' is mathematically represented using below,

$$P(t) = \sum D_i \, wt_{PQ} + b_j \qquad (1)$$

From the above equation (1), '$P\ (t)$' indicates a neuron process in input layer at a time 't' in which 'D_i' signifys an input data points 'D_i' and 'μ_{AB}' denotes the weight between the input and hidden layer and 'b_j' is a bias term. After receiving the input, the input layer sent the data to hidden layers. The first hidden layer in DLMMSC Technique partions the input whole data into a number of chunks randomly in two dimensional space. Thus, neurons process in first hidden layer '$Q_1(t)$' is mathematically obtained using below,

$$Q_1(t) = \{C_1, C_2, .., C_n\} \qquad (2)$$

From equation (2), '$Q_1(t)$' represents the first hidden layer output at the time 't' whereas 'C_n' indicates the total number of chunks. Then, the outcome of first hidden layer is forwarded to second hidden layer. The DLMMSC technique carried out mapping process in second hidden layer by calculating the mean for each chunks (i.e. cluster) in two dimensional space. For all cluster in two dimensional spaces, then mean is determined using below mathematical formula,

$$\alpha = \frac{1}{N}\sum_{i=1}^{N} D_i \quad (3)$$

From the above equation (5), 'α' represents a mean of the cluster and 'D_i' represents the data. In DLMMSC technique, the mean is measured as the weighted average of the data points. During the mapping process, each data point is mapped to mean of cluster. For each mean (i.e. centroid), the nearby data are grouped into the cluster using Gaussian kernel function using below expression,

$$Map\ Phase \rightarrow\ GKF\ (\alpha, D_i) = \exp\left(-\frac{\|D_i-\alpha\|^2}{2\,\beta^2}\right) \quad (4)$$

From the above equation (6), 'GKF' represent a Gaussian kernel function. Here, '$\|D_i - \alpha\|^2$' indicates a squared distance between the data point and cluster mean (i.e. centroid) in two dimensional space and 'β' denotes a deviation from its mean. Thus, neurons action in second hidden layer '$Q_1(t)$' is mathematically formulated as,

$$Q_2(t) = Map(key_1/value_1) \rightarrow list(key_2/value_2) \quad (5)$$

From the above formula (7), '$Q_2(t)$' represents output of second hidden layer at the time 't'. Here, 'key_i' denotes a cluster mean and '$value_i$' indicates a value of data. During the each iteration, the data point is shifts to the nearest cluster mean. This process is repeated until the data points are not moved into the clusters. After mapping process, reduce phase in DLMMSC technique generates the optimal e-commerce data clustering result and also removes the irrelevant data in given dataset. Consequently, neurons process in third hidden layer '$Q_1(t)$' is mathematically expressed as,

$$Reduce\ Phase \rightarrow\ Q_3(t) = OCR \quad (6)$$

From above mathematical representation (2), '$Q_3(t)$' signifies result of third hidden layer output at the time 't' whereas 'OCR' indicates optimal clustering result of e-commerce data. This obtained optimal clustering result is then sent to output layer. Therefore, process in output layer at time 't' i.e. '$R(t)$' is mathematically represented as,

$$R(t) = Awt_{QR}Q_3(t) \quad (7)$$

From the above expression (13), 'wt_{QR}' refers a weight between the hidden and output layer. Here, 'A' is an activation function and '$Q_3(t)$' denotes the output get from the third hidden layer. With the help of above mathematical formula, the output layer in DLMMSC technique generates optimal clustering result for an input big e-commerce dataset. By using the above processes, the DLMMSC technique accurately group the same type of e-commerce data together in diverse clusters with a minimal lower amount of time complexity.

The algorithmic process of DLMMSC technique is described as follows.

// Deep Learned MapReduce Mean Shift Clustering Algorithm
Input: Number of e-commerce data in Dataset '$DS = D_{1,}, D_{2,}, D_{3,} \ldots, D_N$'
Output: Improved clustering accuracy with minimal time complexity
Step 1:Begin
Step 2: **For** an input big e-commerce dataset 'DS'
Step 3: Input layer 'P'get e-commerce data '$D_{1,}, D_{2,}, D_{3,} \ldots, D_N$' as input using (1)
Step 4: 'P'forward received e-commerce data to first hidden layer

Step 5: First hidden layer 'Q_1' divide dataset into a number of chunks using (2)
Step 6: 'Q_1' sent partitioned data result to second hidden layer
Step 7: **For** each chunk (i.e. cluster)
Step 8: Second hidden layer 'Q_2' determine cluster mean 'α' using (3)
Step 9: **End For**
Step 9: 'Q_2' perform mapping using (4) and (5) and sent result to third hidden layer 'Q_3'
Step 10: 'Q_3' remove the unrelated data and generate optimal clustering result using (6)
Step 11: This optimal clustering result is forward to output layer 'R'
Step 12: Output layer 'R' give clustered e-commerce data using (7)
Step 13: End for
Step 14:End

Algorithm 1 Deep Learned MapReduce Mean Shift Clustering

Algorithm 1 explains the step by step process of DLMMSC technique to achieve better big e-commerce data analytics performance through clustering. As demonstrated in the above algorithmic process, the number of e-commerce data is obtained as input in the input layer. Then, the DLMMSC technique forwards the acquired input e-commerce data to first hidden layer. Followed by, DLMMSC technique partitions the taken e-commerce data into a number of chunks (i.e. cluster) in two dimensional spaces. This partitioned data result is then transmitted to second hidden layer. The DLMMSC technique determines cluster mean for each chunks in two dimensional spaces at the second hidden layer. After that, DLMMSC technique map each data point in two dimensional spaces to nearby cluster mean using Gaussian kernel function. The mapped data result to corresponding cluster mean is then sent to the third hidden layer. The DLMMSC technique eliminates the irrelevant data and produce optimal e-commerce data clustering result at the third hidden layer. Consequently, DLMMSC technique transmits the obtained optimal clustering result to output layer. Finally, the output layer provides the clustered data result.

With the help of above algorithmic process, DLMMSC technique correctly clusters the related e-commerce data together in various clusters with a lower amount of time consumption as compared to conventional works. This supports for DLMMSC technique to decreases the inaccurate clustering of data (i.e. error rate) as compared to state-of-the-art works. With the obtained clustering result of big e-commerce data, DLMMSC technique determines which products sell well and which products are best suited for customer base and which products are supported by best marketing efforts. Thus, DLMMSC technique enhances the big e-commerce data analytics performance as compared to existing works.

3. EXPERIMENTAL SETTINGS

In order to measure the performance of the proposed, DLMMSC technique is implemented in Java Language using Amazon-E-commerce-Data-set [21]. This dataset contains a Real e-commerce product data that were available on-sale at Amazon on-line market place on November 17-19, 2014. The dataset comprises of products from 6 main categories such as Automotive, Books, Electronics, Movies, Phones and Home including 1529 sub-categories. In order to conduct

the experimental evaluation, DLMMSC technique takes 1000-10000 data from Amazon-E-commerce-Data-set. The performance of DLMMSC technique is determined in terms of clustering accuracy, time complexity, and error rate and space complexity. The experimental results of DLMMSC technique are compared against two conventional methods namely A MapReduce-based hierarchical subspace-clustering algorithm called PAPU [1] and fast clustering algorithm called (MUFOLD-CL) [2].

4. RESULT AND DISCUSSIONS

In this section, the result of DLMMSC technique is discussed and compared with two existing methods. The experimental result of DLMMSC technique is compared with existing MapReduce-based hierarchical subspace-clustering algorithm called PAPU [1] and fast clustering algorithm called (MUFOLD-CL) [2] using below parameters with the help of below tables and graphical representation.

4.1 Measurement of Clustering Accuracy

Clustering accuracy 'CA' determines ratio of number of e-commerce data exactly grouped to the total number of data. The clustering accuracy is calculated using below mathematical formula,

$$CA = \frac{\gamma_{EC}}{N} * 100 \qquad (8)$$

From the above equation (8), 'γ_{EC}' denotes number of exactly clustered e-commerce data and 'N' indicates a total number of data taken for performing experimental process. The clustering accuracy is estimated in terms of percentage (%).

Table 1 Comparative Result of Clustering Accuracy for Three Techniques

Number of data (N)	Clustering Accuracy (%)		
	DLMMSC	PAPU	MUFOLD-CL
1000	94	90	81
2000	95	89	83
3000	97	86	78
4000	92	87	79
5000	95	88	81
6000	96	86	79
7000	97	83	75
8000	95	86	78
9000	96	87	81
10000	96	84	80

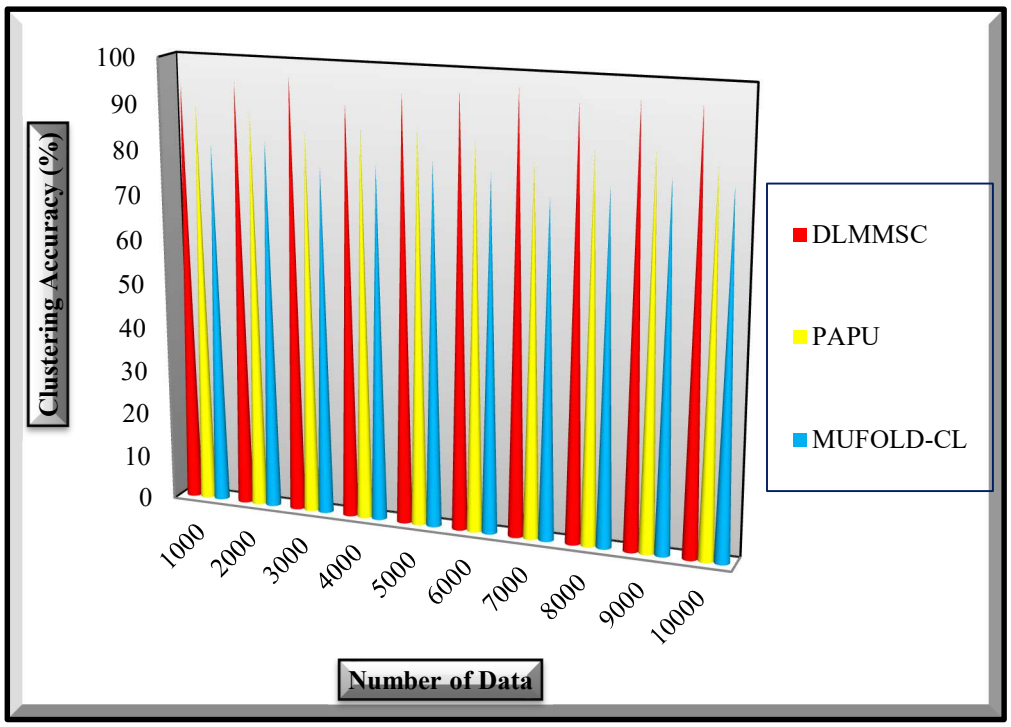

Figure 3 Experimental Result of Clustering Accuracy versus Number of Data

Table 1 and Figure 3 demonstrates the tabulation and graphical representation of clustering accuracy result obtained for three techniques namely proposed DLMMSC technique and existing PAPU [1] and MUFOLD-CL [2]. As demonstrated in the above performance result, proposed DLMMSC technique provides better clustering accuracy for grouping large volume of e-commerce data when compared to conventional PAPU [1] and MUFOLD-CL [2]. This is because of application of MapReduced Mean Shift Clustering in DLMMSC technique.

On the contrary to existing works, the proposed DLMMSC technique performs clustering process by mapping the each data to the mean of clusters in two-dimensional space. After that, proposed DLMMSC technique separates mapped data in a reduce phase and also removes near-duplicate data. From that, proposed DLMMSC technique generates final set of optimal clusters with a minimal error rate. Accordingly, DLMMSC technique effectively clusters the relevant e-commerce data together into different clusters with a higher accuracy. Thus, DLMMSC technique enhances the ratio of number of e-commerce data precisely clustered as compared to other works [1] and [2]. Hence, proposed DLMMSC technique enhances clustering accuracy by 10 % as compared to PAPU [1] and 19 % as compared to MUFOLD-CL [2].

4.2 Measurement of Time Complexity

Time Complexity 'TC' estimates the time required for clustering similar e-commerce data. The time complexity is determined using below mathematical formula,

$$TC = N * T_{CS} \quad (9)$$

From equation (9), 'TC' represents a time employed to group a single e-commerce data and 'N' designates a total number of data. The time complexity is evaluated in terms of milliseconds (ms).

Table 2 Comparative Result of Time Complexity for Three Techniques

Number of data (N)	Time Complexity (ms)		
	DLMMSC	PAPU	MUFOLD-CL
1000	17	21	24
2000	20	26	32
3000	21	30	33
4000	24	32	36
5000	30	38	40
6000	36	44	47
7000	41	48	53
8000	46	53	57
9000	53	59	61
10000	58	64	67

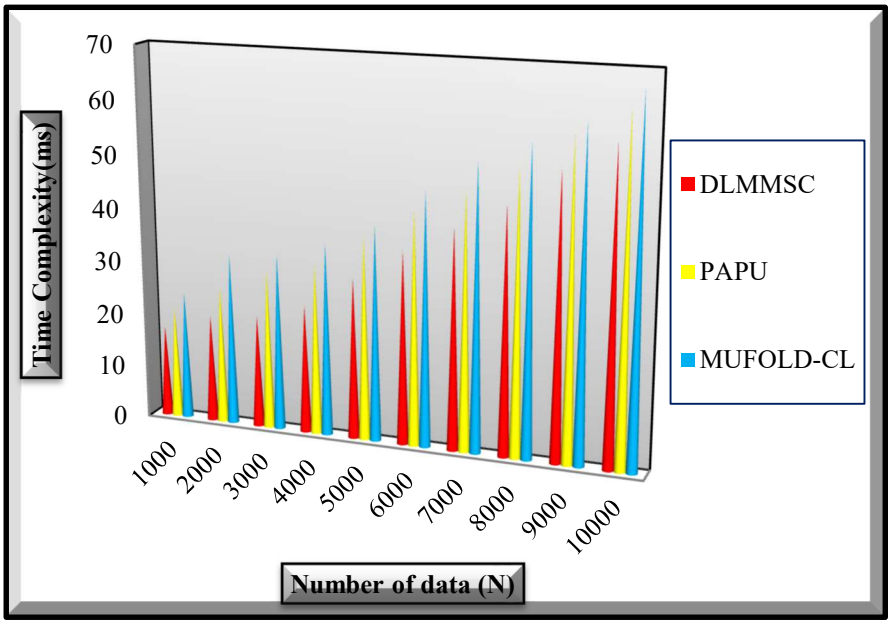

Figure 4 Experimental Result of Time Complexity versus Number of Data

Table 2 and Figure 4 shows the tabulation and graphical illustration of time complexity result acquired during the big data clustering process for three techniques namely proposed DLMMSC technique and existing PAPU [1] and MUFOLD-CL [2]. As presented in the above comparative result, proposed DLMMSC technique gets minimal amount of time complexity to accurately cluster the massive amount of e-commerce data in input dataset as compared to conventional PAPU [1] and MUFOLD-CL [2]. This is owing to application of deep neural learning concepts in DLMMSC technique.

On the contrary to state-of-the-art works, the proposed DLMMSC technique is designed based on the structure and function of the brain called artificial neural networks. This structure helps for proposed DLMMSC technique to deeply examine input big e-commerce data with a

minimum amount of time during clustering process. In addition to that, proposed DLMMSC technique includes three main layers namely input, hidden and output layers for deep analyze big data and thereby exactly cluster the data in given dataset with lower time consumption. Therefore, DLMMSC technique reduces the time required for grouping similar e-commerce data when compared to other works [1] and [2]. Thus, proposed DLMMSC technique decreases the time complexity by 16 % as compared to PAPU [1] and 25 % as compared to MUFOLD-CL [2].

4.3 Measurement of Error Rate

Error Rate 'ER' measures ratio of number of e-commerce data inaccurately grouped to the total number of data. The error rate is computed using below mathematical representation,

$$ER = \frac{N_{IC}}{N} * 100 \quad (10)$$

From above equation (10), 'N_{IC}' point outs a number of e-commerce data incorrectly clustered and 'N' indicates a total number of data. The error rate is determined in terms of percentage (%).

Table 3 Comparative Result of Error Rate for Three Techniques

Number of data (N)	Error Rate (%)		
	DLMMSC	PAPU	MUFOLD-CL
1000	6	10	20
2000	5	11	16
3000	3	14	22
4000	8	13	21
5000	5	12	20
6000	4	14	21
7000	4	17	25
8000	5	14	22
9000	3	13	19
10000	4	16	20

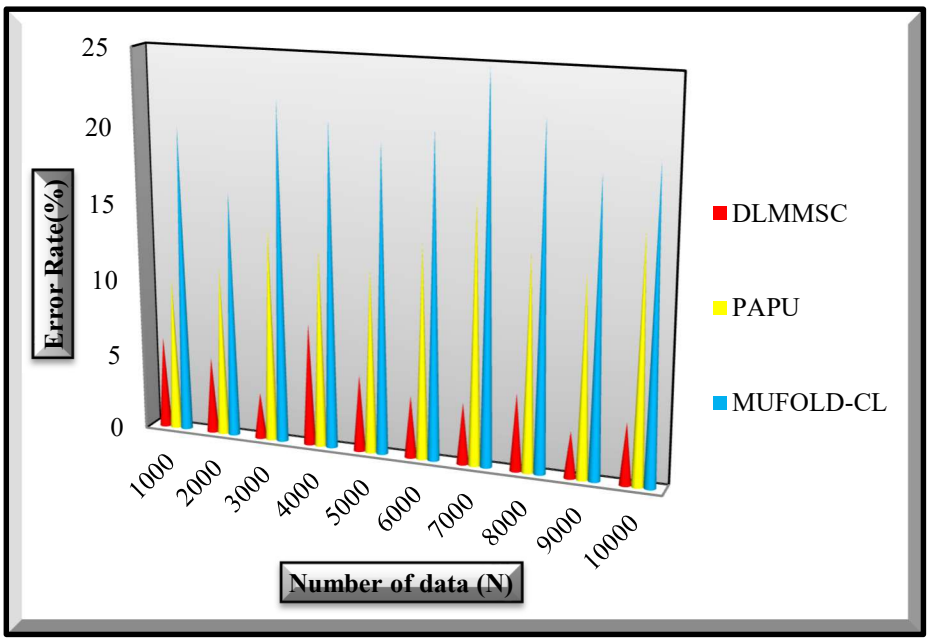

Figure 5 Experimental Result of Error Rate versus Number of Data

Table 3 and Figure 5 presents the tabulation and graphical depiction of error rate result obtained for three techniques namely proposed DLMMSC technique and existing PAPU [1] and MUFOLD-CL [2]. As presented in the above experimental result, proposed DLMMSC technique gives minimum error rate for efficient clustering of big e-commerce data in input dataset when compared to conventional PAPU [1] and MUFOLD-CL [2]. This is because of application of MapReduced Mean Shift Clustering in DLMMSC technique.

On the contrary to conventional algorithms, proposed DLMMSC technique map each input e-commerce data in two dimensional space to nearby cluster mean through a Gaussian kernel function. This supports for proposed DLMMSC technique to correctly map the data to corresponding cluster mean. From that, proposed DLMMSC technique takes away the unrelated data and thereby gives optimal e-commerce data clustering result. Hence, DLMMSC technique reduces the ratio of number of e-commerce data mistakenly grouped when compared to other works [1] and [2]. As a result, proposed DLMMSC technique decreases the error rate of big data clustering by 63 % when compared to PAPU [1] and 77 % when compared to MUFOLD-CL [2].

4.4 Measurement of Space Complexity

Space Complexity 'SC' determines memory space taken for storing the clustered e-commerce data. The space complexity is mathematically computed using below,

$$SC = N * Mem(SSD) \qquad (11)$$

From the above equation (11), '$Mem(SSD)$' designates a memory needed to store single clustered data and 'M' refers a total number of data. The space complexity is evaluated in terms of mega bytes (MB).

Table 4 Comparative Result of Space Complexity for Three Techniques

Number of data (N)	Space Complexity (MB)		
	DLMMSC	PAPU	MUFOLD-CL

1000	18	25	29
2000	20	29	31
3000	23	31	33
4000	25	33	36
5000	27	36	38
6000	28	39	43
7000	31	41	46
8000	35	45	48
9000	37	48	51
10000	44	52	54

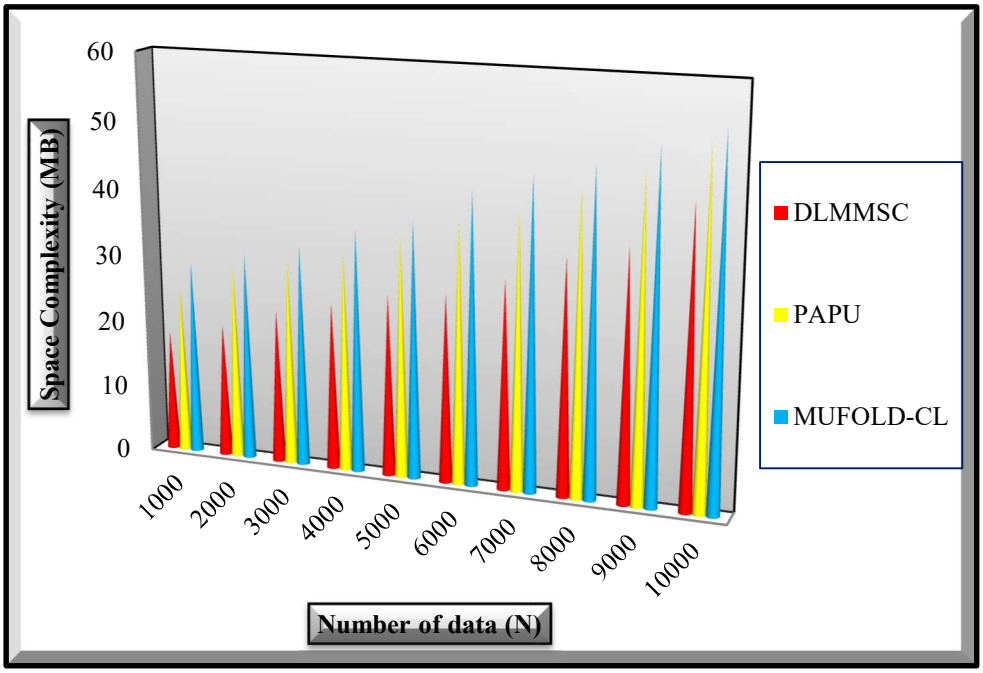

Figure 6 Experimental Result of Space Complexity versus Number of Data

Table 4 and Figure 6 depict the tabulation and graphical representation of space complexity result obtained when implementing the three techniques i.e. proposed DLMMSC technique and existing PAPU [1] and MUFOLD-CL [2] in Java language using Amazon-E-Commerce-Data-set. As presented in the above performance result, proposed DLMMSC technique attains lower space complexity during the clustering process of big e-commerce data when compared to conventional PAPU [1] and MUFOLD-CL [2]. This is due to application of deep neural learning and MapReduced Mean Shift Clustering in DLMMSC technique on the contrary to conventional algorithms.

By using the deep neural learning and MapReduced Mean Shift Clustering concepts, proposed DLMMSC technique accurately group only the related data to corresponding cluster mean. As a result, final optimal set of clusters does not contain an unrelated data. Therefore,

DLMMSC technique reduces the memory space taken for storing the clustered e-commerce data when compared to other works [1] and [2]. Accordingly, proposed DLMMSC technique minimizes the space complexity of big data clustering by 23 % when compared to PAPU [1] and 30 % when compared to MUFOLD-CL [2].

5. LITERATURE SURVEY

Two clustering validity indices were employed in [11] with the intention of clustering a vast amount of data in an input dataset with a lower error rate. However, clustering accuracy was poor. Parallel power iteration clustering was designed in [12] for minimizing the computational time of big data clustering. But, accuracy using this algorithm was poor.

K-Means Hadoop MapReduce (KM-HMR) was applied in [13] for increasing the quality of clusters to create clusters with maximum intra-cluster and minimum inter-cluster distances for large datasets. However, space complexity involved during big data clustering process was higher. Scalable Random Sampling with Iterative Optimization Fuzzy c-Means algorithm (SRSIO-FCM) was designed in [14] using an Apache Spark Cluster to handle the issues related with big data grouping. But, processing time required for performing big data clustering process was not reduced.

A Hybrid Approach was developed in [15] to perform big data clustering process with minimal computational complexity. But, space complexity using this approach was not considered. A Comprehensive analysis of clustering algorithms developed for big data was presented in [16].

A Data-Aware HDFS and Evolutionary Clustering Technique was introduced in [17] for enhancing accuracy of big data clustering. But, execution time of this technique was higher. Clustering big IoT data was carried out in [18] with application of metaheuristic optimized mini-batch and parallel partition-based DGC algorithm. However, clustering performance was not at required level.

A high-order CFS algorithm (HOCFS) was implemented in [19] to group similar heterogeneous data. However, accuracy was lower when taking big data as input. Multivariate-Rank-Based Techniques was presented in [20] for minimizing the false positive rate while performing big data clustering. But, error rate using this technique was not minimized.

6. CONCLUSION

The DLMMSC technique is proposed with aiming at enhancing the accuracy of clustering to achieve better big e-commerce data analytics performance as compared to conventional works. The aim of DLMMSC technique is attained with application of deep neural learning and MapReduce and Mean Shift Clustering. The designed DLMMSC technique enhances the clustering performance to efficiently group the related e-commerce data with higher accuracy as compared to state-of-the-art works. As well, the DLMMSC technique minimizes the time and space complexity of the big data clustering as compared to existing works. By an efficient clustering of big e-commerce data, DLMMSC technique determines which product, price and advertising is best to increase their profits in Amazon as compared to state-of-the-art works. The effectiveness of DLMMSC technique is measured in terms of clustering accuracy, time complexity, and error rate and space complexity and compared with two conventional works. The experimental

result demonstrates that DLMMSC technique presents better performance with an enhancement of clustering accuracy and reduction of time complexity to effectively analyze the big e-commerce data when compared to state-of-the-art works.

REFERENCES

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

[2] Yun Wu, Zhiquan He, Hao Lin, Yufei Zheng, Jingfen Zhang,Dong Xu, "A Fast Projection-Based Algorithm for Clustering Big Data", Interdisciplinary Sciences: Computational Life Sciences, Springer, Pages 1–7, 2018

[3] Muruganantham S, Elango N M, "Fully Recurrent Deep Neural Learning Based Uncertain Data Clustering", Journal of Theoretical and Applied Information Technology, Volume 97, Issue 4, February 2019

[4] Erxue Min, Xifeng Guo, Qiang Liu, Gen Zhang, Jianjing Cui, and Jun Long, "A Survey of Clustering With Deep Learning: From the Perspective of Network Architecture", Volume 6, Pages 39501-39514, 2018

[5] Chin-Yi Chenand Jih-Jeng Huang, "Double Deep Autoencoder for Heterogeneous Distributed Clustering", Information, Volume 10, Issue 144, Pages 1-14, 2019

[6] Singh Vijendra1 and Sahoo Laxman, "Subspace Clustering of High-Dimensional Data: An Evolutionary Approach", Hindawi Publishing Corporation, Applied Computational Intelligence and Soft Computing, Volume 2013, Article ID 863146, Pages 1-12, 2013

[7] MinyarSassi Hidri, Mohamed AliZoghlami, RahmaBen Ayed, "Speeding up the large-scale consensus fuzzy clustering for handling Big Data", Fuzzy Sets and Systems, Elsevier, Volume 348, Pages 50-74, October 2018

[8] FanyuBu, "An efficient fuzzy c-means approach based on canonical polyadic decomposition for clustering big data in IoT", Future Generation Computer Systems, Volume 88, Pages 675-682, November 2018

[9] 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, Pages 1–18, July 2017

[10] Qingchen Zhang, Laurence T.Yang, Arcangelo Castiglione, Zhikui Chend PengLi, "Secure weighted possibilistic c-means algorithm on cloud for clustering big data", Information Sciences, Elsevier, Pages 1-15, 2018

[11] José María Luna-Romera, Jorge García-Gutiérrez, María Martínez-Ballesteros, José C. Riquelme Santos, "An approach to validity indices for clustering techniques in Big Data", Progress in Artificial Intelligence, Springer, Volume 7, Issue 2, Pages 81–94, June 2018

[12] Weizhong Yan, Umang Brahmakshatriya, Ya Xue, Mark Gilder, Bowden Wise, "p-PIC: Parallel power iteration clustering for big data", Journal of Parallel and Distributed Computing, Elsevier, Volume 73, Issue 3, Pages 352-359, March 2013

[13] Chowdam Sreedhar, Nagulapally Kasiviswanath, Pakanti Chenna Reddy, "Clustering large datasets using K-means modified inter and intra clustering (KM-I2C) in Hadoop", Journal of Big Data, Springer, Volume 4, Issue 27, December 2017
[14] Neha Bharill, Aruna Tiwari, Aayushi Malviya, "Fuzzy Based Scalable Clustering Algorithms for Handling Big Data Using Apache Spark", IEEE Transactions on Big Data, Volume2 , Issue 4, Pages 339 – 352, 2016
[15] Dheeraj Kumar, James C. Bezdek, Marimuthu Palaniswami, Sutharshan Rajasegarar, "A Hybrid Approach to Clustering in Big Data", IEEE Transactions on Cybernetics, Volume 46, Issue 10, Pages 2372 – 2385, 2016
[16] Dongkuan Xu, Yingjie Tian, "A Comprehensive Survey of Clustering Algorithms", Annals of Data Science, Springer, Volume 2, Issue 2, Pages 165–193, June 2015
[17] Mustafa Hajeer, Dipankar Dasgupta, "Handling Big Data Using a Data-Aware HDFS and Evolutionary Clustering Technique", IEEE Transactions on Big Data, Volume 5, Issue 2, Pages 134 – 147, June 2019
[18] Rui Tang, Simon Fong, "Clustering big IoT data by metaheuristic optimized mini-batch and parallel partition-based DGC in Hadoop", Future Generation Computer Systems, Volume 86, Pages 1395-1412, September 2018
[19] Fanyu Bu ,Zhikui Chen, Peng Li, Tong Tang, and Ying Zhang, "A High-Order CFS Algorithm for Clustering Big Data", Mobile Information Systems, Hindawi Publishing Corporation, Volume 2016, Article ID 4356127, Pages 1-8, 2016
[20] Pritha Guha, "Application of Multivariate-Rank-Based Techniques in Clustering of Big Data", VIKALPA, The Journal for Decision Makers, Volume 43, Issue 4, Pages 179–190, 2018
[21]Amazon-E-commerce-Data-set:https://github.com/SamTube405/Amazon-E-commerce-Data-set