International Journal of Applied Engineering Research ISSN 0973-4562 Volume 12, Number 23 (2017) pp. 13186-13193

Dice Similarity Based Ensemble Clustering for Sparsely Distributed High Dimensional Data

R.Pushpalatha
Research Scholar in Computer Science, Erode Arts and Science College (Autonomous), Erode and Assistant Professor, Department of Computer Science, Kongu Arts and Science College (Autonomous), Nanjanapuram, Erode, Tamil Nadu, India.

Orcid: 0000-0001-8834-7284

Dr.K.MeenakshiSundaram
Associate Professor, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode,Tamil Nadu, India.

Abstract

Data mining is the process of extracting the valuable patterns from large volume of data. Clustering in data mining is the process of dividing the data points depending on their similarity level. Clustering techniques for managing the high dimensional data is more complicated because of intrinsic sparsity nature of high dimensional data. However, the clustering accuracy and similarity measurement time was not improved using existing clustering techniques such as fuzzy c-means and spectral clustering. In order to overcome these limitations, Dice Similarity Threshold based Ensemble Clustering (DST-EC) Technique is introduced. DST-EC technique clusters the sparsely distributed high dimensional data points based on the similarity value. Initially in DST-EC technique, Dice Similarity Coefficient Measurement Algorithm is introduced to measure the similarity between two high dimensional data points with minimum similarity measurement time consumption and higher true positive rate. After finding the similarity, the different similarity threshold range is set for clustering the data points. Finally based on the similarity threshold value, Similarity Threshold Ensemble Clustering Algorithm clusters the similar data points to form number ofclusters with higher clustering accuracy. The performance of DST-EC technique is measured in terms oftrue positive rate, similarity measurement time and clustering accuracy with El Nino weather data sets from UCI Machine Learning Repository. The experimental result explains that the DST-EC technique improves the clustering accuracy by 15% and reduces the similarity measurement time by 21% when compared to state-of-the-art-works.

Keywords: data mining, clustering, similarity, threshold range, dice similarity coefficient measurement, similarity threshold ensemble clustering.

Introduction

Clustering is the process of grouping the similar data items together to form the cluster. Clustering is a challenging issue forthe discovery of identical groups of data based on similarity measure. Fuzzy c-means (FCM) model was introduced in [1] with sparse regularizationby varying the FCM objective function into weighted between-cluster sum of square form. But, the clustering accuracy was not improved using FCM model.

Two weight matrix constructions were introduced for spectral clustering algorithm in [2] using the similarity of sparse representation vectors. Though the clustering accuracy was improved, the algorithm consumes large amount of time for finding similar data points.

The local input space histogram was introduced in [3] to enhance the prototype-based vector quantization techniques to collect large amount of information about structure of relevant input space. Though, designed technique increased the performance of visualization and hierarchical clustering, the true positive rate was not improved. Modified fuzzy c-medoid clustering algorithm was introduced in [4] with geodesic distance measure and selected thepotential cluster between the central objects.

The designed techniques managed the data that lie on low dimensional manifold of high dimensional feature space. But, the clustering accuracy was not improved as it employed geodesic distance measure.Constraint Partitioning K-Means algorithm was introduced in [5] to form accurate clusters by means of Principal Component Analysis. But, the clustering was not carried out in efficient manner.

A stratified sampling method was introduced in [6] for creating the subspace component datasetsduring the ensemble clustering of data. A new development of Extreme learning machine (ELM) was surveyed and its applications in high dimensional as well as in large data were studied in [7].Discriminative embedded clustering (DEC) approach was introduced in [8] for addressing the formulated nonconvex optimization issues. But, the clustering accuracy was not

improved using DEC approach.Sparse Subspace Clustering was carried out in [9] that cluster the data points through sparse optimization program to assume clustering of data into subspaces for minimizing the outliers. Group Sparse graph (GSgraph) method was introduced in [10] with Kernel tricks for constructing an informative graph by auto-grouped sparse regularization depending on ℓ1-graph. But, sparse graph construction failed to satisfy the locality limitation as well as failed to combine nonzero coefficients locality and sparsity.

In order to overcome the above mentioned issues, Dice Similarity Threshold based Ensemble Clustering (DST-EC) Technique is introducedfor efficient clustering of similar data points from sparsely distributed high dimensional dataset.

The contribution of our research is given as:Dice Similarity Threshold based Ensemble Clustering (DST-EC) Technique clusters the sparselydistributed high dimensional data points based on similarity measure between the data points. In DST-EC technique, Dice Similarity Coefficient Measurement Algorithm is used to find the similarity between two high dimensional data points with minimum similarity measurement time consumption and higher true positive rate. After finding the similarity between data points, the different similarity thresholdrange is set. Finally based on the similarity threshold value, the clustering process is carried outusing Similarity Threshold Ensemble Clustering Algorithm with higher clustering accuracy.

The rest of the paper ordered as follows. In Section 2, the proposed DST-EC technique is described with the help of architectural diagram. In Section 3, experimental evaluation is discussed and result analysis is carried out with help of tables and graph in Section 4. A review of different high dimensional data clusteringis studied in section 5. The Section 6 concludes the designed work.

Dice Similarity Based Ensemble Clustering Technique

The Dice Similarity Threshold based Ensemble Clustering (DST-EC) Technique is introduced to cluster the sparselydistributed high dimensional data points. DST-EC technique is employed for clustering the sparse data points with higher clustering accuracy. DST-ECtechnique introduces Dice Similarity Coefficient Measurement Algorithm to find the similarity betweentwo sparsely distributed high dimensional data points with lesser similarity measurement time consumption. Then, DST-EC technique assigns different similarity threshold range. Finally based on the similarity threshold range assigned, Similarity Threshold Ensemble Clustering Algorithm clusters the similar data points with higher clustering accuracy. The overall architectural diagram of DST-EC Technique for clustering the sparselydistributed high dimensional data points is explained in Figure 1. From Figure 1, DST-EC Technique initially collects the sparsely distributed high dimensional data points from El Nino weather dataset as input.

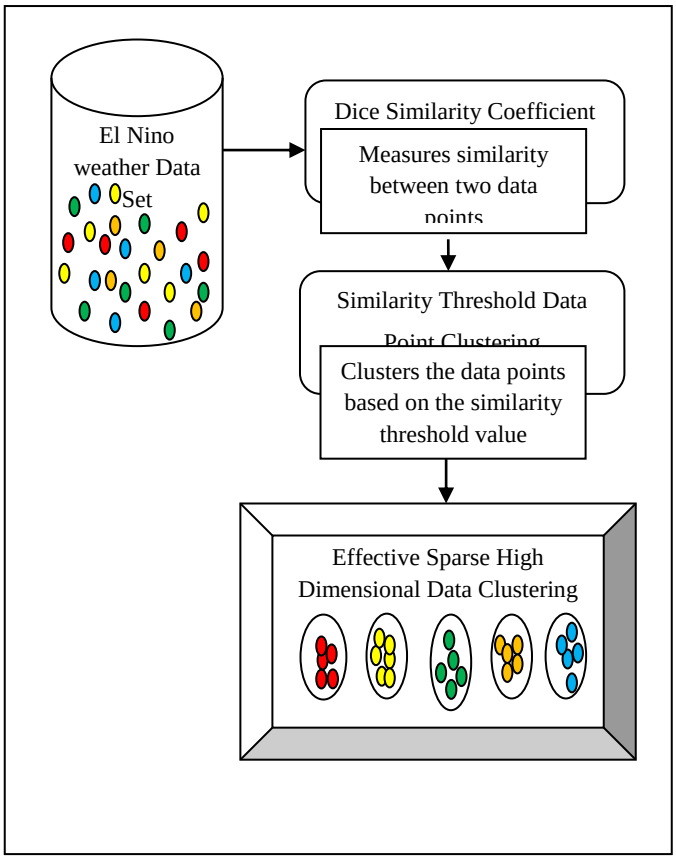

Figure 1: Overall Architectural Design of Dice Similarity Threshold based Ensemble Clustering (DST-EC) Technique

After that, DST-EC Technique usedDice Similarity Coefficient Measurement Algorithmfor measuring thesimilarity between two data points from sparse high dimensional database. Then, DST-EC technique set different similarity threshold range. Finally based on the similarity threshold range assigned, Similarity Threshold Ensemble Clustering Algorithm clusters the similar data points with higher clustering accuracy.The brief description of dice similarity coefficient measurement and similarity threshold ensemble clustering are explained in upcoming section.

Dice similarity coefficient measurement

The DST-EC Technique used Dice Similarity Coefficient to increase the performance of similar data pointsfrom the sparsely distributed high dimensional dataset (i.e., El Nino weather dataset). The Dice Similarity Coefficient measures the similarity between the two data points in sparsely distributed high dimensional dataset for clustering of similar data points with higher clustering accuracy. The input sparselyhigh dimensional datasetis initially divided into number of data pointsand then Dice Similarity Coefficient is measured based on data type such as sea surface temperatures, relative humidity,rainfall, subsurface temperatures air temperature data etc. The process involved in Dice Similarity

Coefficient measurement for finding the similar data points in a given dataset is described in Figure 2.

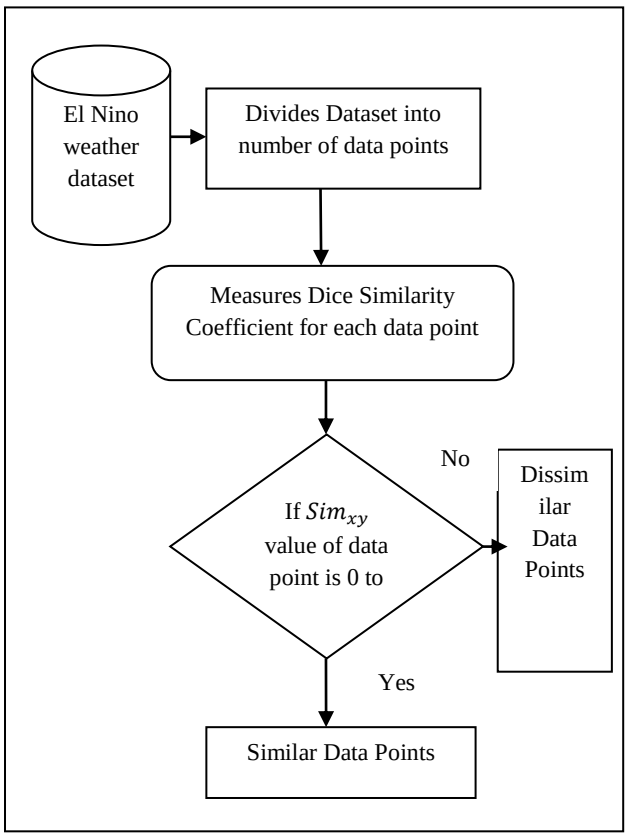

Figure 2: Dice Similarity Coefficient Measurement

From Figure 2, DST-EC Technique initially measures thedice similarity for each data pointin dataset.As a result, the Dice Similarity Coefficient of each data pointis measuredby using mathematical formula,

$$Sim_{i,j} = \frac{2{x_i}^T * x_j}{\|x_i\|^2 + \|x_j\|^2} \quad (1)$$

From equation (1) 'x_i' and 'x_j'representsdata points from given sparse high dimensional dataset.

The Dice Similarity Values between two data points rangesfrom0 to 1.0. Therefore, data point with dicesimilarity value between 0 and 1.0 is taken as similar data points. Then, the remainingdata points are considered as dissimilar data pointsfor clustering process. The algorithmic process of Dice Similarity Coefficient for finding the similardata pointis described below,

// Dice Similarity Coefficient Measurement Algorithm

Input: El Nino weather dataset

Output: Similar data points

Step 1:Begin

Step 2: For each data point in given dataset

Step 3: MeasureDiceSimilarity Coefficient using (1)

Step 4: If (Dice similarity value Sim_{ij}lies between 0 to 0.1)**then**

Step 5: Data point is said to be similar

Step 6:Else

Step 7: Data point is said to be dissimilar

Step 8:End if

Step 9: End for

Step 10:End

Algorithm 1 Dice Similarity Coefficient Measurement Algorithm

From algorithm 1,DST-EC Technique efficiently identifiessimilar data points for clustering process. After measuring the DiceSimilarity Coefficient between data points, similarity threshold ensemble clusteringprocess is carriedin order to improve the performance of clustering accuracy.

Similarity threshold ensemble clustering

The DST-EC Technique used Similarity Threshold Ensemble Clustering for improving theclustering accuracy of sparsely distributed high dimensional data. In Similarity Threshold Ensemble Clustering process, data pointclusteringis carried out based on different similarity threshold values

$$(\text{i.e},(T_0, \ldots, T_{i-1}), (T_i, \ldots . . T_{j-1}), (T_j \ldots ., T_n)).$$

In first cluster,the data points with the dice similarity coefficient values ranging fromT_0 to T_{i-1}get clustered. Then in the second cluster, data points with dice similarity coefficient values ranging from T_i to T_{j-1}get grouped.

Likewise, the data points get clustered based on the similarity threshold range. The process involved in Similarity Threshold Ensemble Clusteringis explained in below Figure 3.

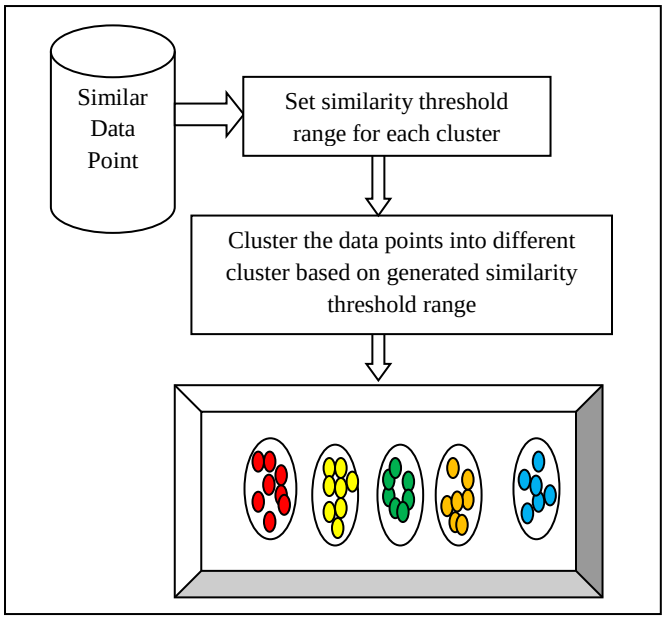

Figure 3: Similarity Threshold Ensemble Cluster Process

As shown in Figure 3, Similarity Threshold Ensemble Cluster initially collects the similar data pointsas input from Dice Similarity Coefficient Measurement Algorithm. Then, clustering of sparse high dimensionaldata points is carried outusingdifferentsimilarity threshold values for improving the clustering accuracy. The clustering result of Similarity Threshold Ensemble Clustering process comprises k-number of clusters with different number of data points based on similaritythreshold values. The algorithmic process of similarity threshold ensemble clustering is describedbelow,

// **Similarity Threshold Ensemble Cluster Algorithm**

Input: Similar data points, DiceSimilarity Threshold range

Output: Groups Similar data point based on dice similarity threshold

Step 1:Begin

Step 2:For similar data points in dataset

Step 3:Repeat

Step 4: Generate new Dice Similarity Threshold Range

Step 5: Cluster data points based on the generated new similarity threshold

Step 6:Until a stop criterion is met

Step 7:Return Many cluster of similar data points based onsimilarity threshold value

Step 9: End for

Step 10:End

Algorithm 2 Similarity Threshold Ensemble Clustering Algorithm

From Algorithm 2,Similarity Threshold Ensemble Clustering Algorithm in DST-EC Technique initially collects the similar data pointsas input. In Similarity Threshold Ensemble Clustering Algorithm, Similarity Threshold value range is initialized. During the sparsely distributed high dimensional data point clustering process, the dice similarity coefficient measure of data point within the similarity threshold range is clustered. This Similarity Threshold Ensemble Clustering processgets repeated until the similarity threshold value$T$ reaches the maximum value of 1.0. This in turn increases the clustering accuracy using DST-EC Technique.

Experimental evaluation

In this section, Dice Similarity Threshold based Ensemble Clustering (DST-EC) Techniqueis implemented in Java Language with Inter 4-core 2.6GHz CPU and 12 GB RAM. The proposed DST-EC Technique is compared with two existing methods such as Fuzzy C-Means (FCM) model [1] and Spectral Clustering Algorithm [2]. The proposed technique uses E1 Nino dataset from UCI Machine Learning Repository to conduct experiments. This E1 Nino dataset from Tropical Atmosphere Ocean (TAO) array was developed by international Tropical Ocean Global Atmosphere (TOGA) program. The dataset comprises 12 attributes and 178080 instances. The data comprises of the following variables: date, latitude, longitude, zonal winds (west<0, east>0), meridional winds (south<0, north>0), relative humidity, air temperature, sea surface temperature and subsurface temperatures down to a depth of 500 meters. The dataset characteristics is spatio-temporal and attribute characteristics is both real and integer.

RESULT AND DISCUSSION

The result analysis of DST-EC technique is compared against with existing two approaches namely Fuzzy C-Means (FCM) model [1] and Spectral Clustering Algorithm [2] respectively. The performance of DST-EC technique is evaluated on various factors such as clustering accuracy, true positive rate and similarity measurement time with help of tables and graphs.

Impact of similarity measurement time

Similarity Measurement Time(SMT)is defined as the amount of time taken for finding the similarity between two data points. It is measured in terms of milliseconds (ms).The similarity measurement time is mathematically formulated as,

$$SMT = n * time\ for\ finding\ similarity \quad (2)$$

From (2), 'n' represents number of data points. When the similarity measurement time is lesser, the method is said to be more efficient.

Table 1: Tabulation for Similarity Measurement Time

Number of Data Points	Similarity Measurement Time (ms)		
	FCM Model	Spectral Clustering Algorithm	DST-EC technique
50	52	42	36
100	56	46	38
150	60	48	41
200	62	51	44
250	65	55	47
300	67	58	49
350	70	61	52
400	73	63	54
450	76	67	57
500	79	71	59

Table 1illustrates the similarity measurement time with respect to number of data points ranging from 50 to 500. When the number of data point gets increased, the similarity measurement time also gets increased correspondingly. But, the similarity measurement time of proposed DST-EC technique is comparatively lesser than that of the Fuzzy C-Means (FCM) model [1] and Spectral Clustering Algorithm [2].

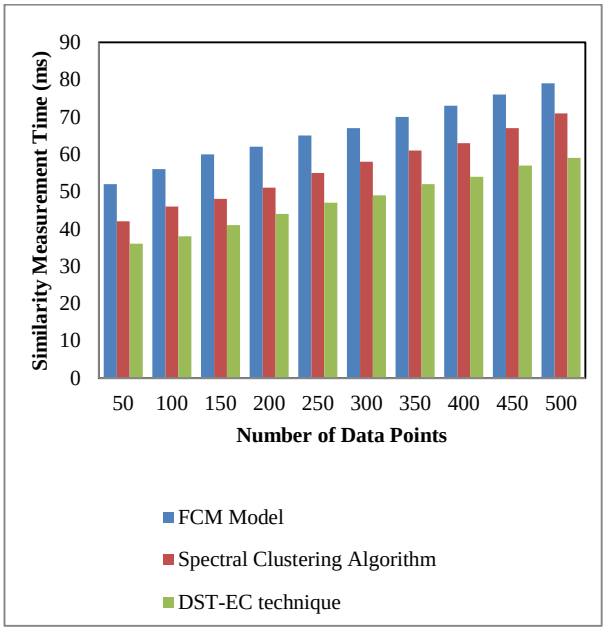

Figure 4: Measure of Similarity Measurement Time

Figure 4explains the similarity measurement time measure of sparse high dimensional data versus number of data points in range of 50-500. From the figure, proposed DST-EC technique has lesser similarity measurement time for clustering the similar data points when compared to Fuzzy C-Means (FCM) model [1] and Spectral Clustering Algorithm [2]. In addition, when the number of data points increases, thesimilarity measurement time also gets increased in all three methods. However, the similarity measurement time using proposed DST-EC technique is lesser. This is because of application of Dice Similarity Coefficient Measurement Algorithm in DST-EC technique where it efficiently calculates the similaritybetween two data points. This in turn helps to reduce thesimilarity measurement time of sparse high dimensional data in an efficient way. As a result, proposed DST-EC technique reducesthe similarity measurement time of sparse high dimensional data by 28% as compared to Fuzzy C-Means (FCM) model [1] and 15% as compared to Spectral Clustering Algorithm [2] respectively.

Impact of true positive rate

True Positive Rate (TPR) of similarity measurement is described as the ratio of number of data points that correctly identified as similar to the total number of data points. The true positive rate of is measured in terms of percentages (%) and formulated as,

$$TPR = \frac{number\ of\ data\ points\ that\ are\ correctly\ identified\ as\ similar}{total\ number\ of\ data\ points} * 100 \quad (3)$$

When the true positive rate is higher, the technique is said to be more efficient.

Table 2: Tabulation for True Positive Rate

Number of Data Points	True Positive Rate (%)		
	FCM Model	Spectral Clustering Algorithm	DST-EC technique
50	69.12	72.65	85.54
100	70.98	73.17	86.41
150	72.56	74.56	87.86
200	73.64	75.96	89.14
250	74.79	77.21	90.72
300	75.14	78.63	91.89
350	76.38	79.14	92.47
400	77.46	80.39	93.61
450	78.23	81.45	94.52
500	79.64	82.87	95.12

Table 2 shows the true positive rate with respect to number of data points ranging from 50 to 500. When the number of data point gets increased, the true positive rate also gets increased correspondingly. But, the true positive rate of proposed DST-EC technique is comparatively higher than that of the Fuzzy

C-Means (FCM) model [1] and Spectral Clustering Algorithm [2].

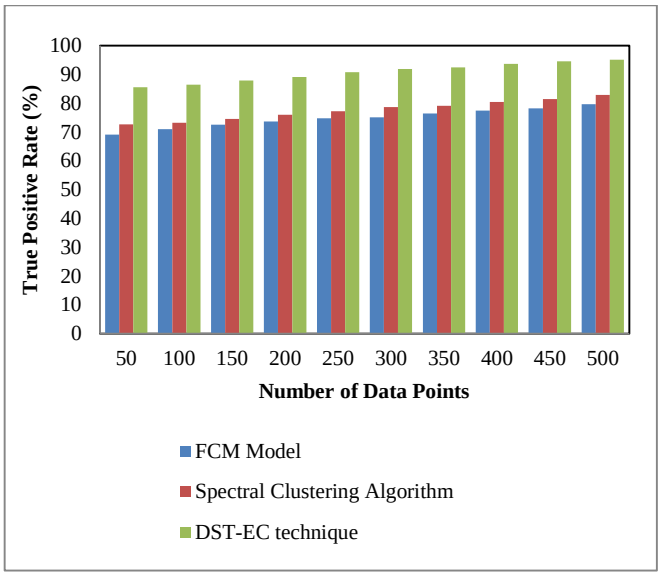

Figure 5: Measure of True Positive Rate

Figure 5portrays the true positive rate measure of sparse high dimensional data versus number of data points in range of 50-500. From figure, proposed DST-EC technique has higher true positive ratewhile finding the similar data points when compared to Fuzzy C-Means (FCM) model [1] and Spectral Clustering Algorithm [2]. In addition, when the number of data points increases, the true positive rate also gets increased in all three methods. However, the true positive rate using proposed DST-EC technique is higher. This is because of application of Dice Similarity Coefficient Measurement Algorithm in DST-EC technique where it efficiently and correctlyfinds the similar data points. This in turn helps to increase the true positive rate of sparse high dimensional data in an efficient way. As a result, proposed DST-EC technique increases the true positive rate of sparse high dimensional data by 21% as compared to Fuzzy C-Means (FCM) model [1] and 17% as compared to Spectral Clustering Algorithm [2] respectively.

Impact of clustering accuracy

Clustering accuracy is defined as ratio of number of similar data points that are clustered correctly to the total number of data points. It is measured in terms of percentage (%).The clustering accuracy is mathematically formulated as,

$$ClusteringAccuracy = \frac{Numberofsimilardatapointscorrectlyclustered}{Totalnumberofdatapoints} \quad (4)$$

When the clustering accuracy is higher, the method is said to be more efficient.

Table 3: Tabulation for Clustering Accuracy

Number of Data Points	Clustering Accuracy (%)		
	FCM Model	Spectral Clustering Algorithm	DST-EC technique
50	71.24	78.25	86.24
100	72.35	79.63	87.81
150	73.69	80.17	88.32
200	74.87	81.54	89.64
250	75.96	82.36	90.11
300	76.32	83.67	91.28
350	77.98	84.23	92.83
400	78.21	85.74	93.34
450	79.44	86.98	94.24
500	80.79	87.45	95.38

Table 3 shows the clustering accuracy with respect to number of data points ranging from 50 to 500. When the number of data point gets increased, the clustering accuracy also gets increased correspondingly. But, the clustering accuracy of proposed DST-EC technique is comparatively higher than that of the Fuzzy C-Means (FCM) model [1] and Spectral Clustering Algorithm [2].

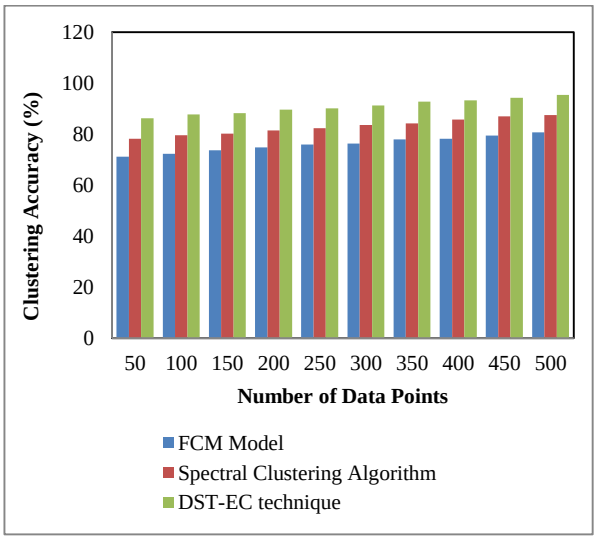

Figure 6: Measure of Clustering Accuracy

Figure 6 describes the clustering accuracy measure of sparse high dimensional data versus number of data points in range of 50-500. From figure, proposed DST-EC technique has higher clustering accuracy during clustering process when compared to Fuzzy C-Means (FCM) model [1] and Spectral Clustering Algorithm [2]. In addition, when the number of data points during clustering increases, the clustering accuracy also gets increased in all three methods. However, the clustering accuracy using proposed DST-EC technique is

higher. This is because of application of Similarity Threshold Ensemble Clustering Algorithm in DST-EC technique where it efficiently clusters the similar data points based on the similarity threshold values. This in turn helps to increase the clustering accuracy of sparse high dimensional data in an efficient way. As a result, proposed DST-EC technique increases the clustering accuracy of sparse high dimensional data by 19% as compared to Fuzzy C-Means (FCM) model [1] and 9% as compared to Spectral Clustering Algorithm [2] respectively.

RELATED WORKS

PEEDR and CPGS Clustering Algorithms for Probabilistic Graphs (CA-PG) [11] were designed to address the clustering correlated probabilistic graph issues and to increase the clustering efficiency. But, the true positive rate was not improved using clustering process. A multiview point-based similarity measure with two related clustering methods were introduced in [12]. A modified PROCLUS algorithm termed as MPROCLUS was introduced in [13] for clustering the high dimensional data with higher performance on running time and consistency. But, the clustering accuracy was not improved in above mentioned techniques.

With kernel mapping and hubness phenomenon, a new algorithm was introduced in [14] for increasing the clustering quality. However, the similarity measurement process was not carried out to cluster the similar data. H-K clustering algorithm was introduced in [15] to reduce the computational complexity and to increase the high dimensional data clustering accuracy. But, complexity remained unaddressed. An incremental semi supervised clustering ensemble approach (ISSCE) was introduced in [16] that utilized the gain of random subspace method for high dimensional data clustering. A CDIM algorithmic framework was presented in [17] for partitioned document clustering to increase the sum of discrimination information given by document. But, the clustering was not carried out in efficient manner.

Similarity based Possibility C-means (SPCM) algorithm was introduced in [18] to cluster high dimensional data with ant colony optimization techniques. But, the similarity measurement time was not reduced using SPCM algorithm. Distance-based clustering algorithm was introduced in [19] to examine and evaluate different similarity measures. Though the similarity measurement time was reduced, the clustering accuracy was not improved. A robust multi objective subspace clustering (MOSCL) algorithm was introduced in [20] for addressing the high-dimensional clustering issues. But, the clustering accuracy was not improved at the required level.

CONCLUSION

An efficient Dice Similarity Threshold based Ensemble Clustering (DST-EC) Technique is developed to cluster the sparsely distributed high dimensional data. DST-EC technique uses Similarity Coefficient Measurement Algorithm to measure the similarity between two data points with lesser similarity measurement time consumption. After finding the similarity between the data points, the different similarity threshold range is set for clustering the data points. Finally based on the similarity threshold value, Similarity Threshold Ensemble Clustering Algorithm clusters thedata points with higherclustering accuracy. The efficiency of DST-EC technique is evaluated with two exiting methods in terms of similarity measurement time, true positive rate and clustering accuracy.The experimental results show that DST-EC technique provides better performance with an enhancement of clustering accuracy as well as true positive rate and reduced the similarity measurement time when compared to state-of-the-art works.

REFERENCES

[1] Xiangyu Chang, Qingnan Wang, Yuewen Liu and Yu Wang “Sparse Regularization in Fuzzy c-Means for High-Dimensional Data Clustering”, IEEE Transactions on Cybernetics, Volume 47, Issue 9, September 2017, Pages 2616 – 2627

[2] Sen Wu, Xiaodong Feng and Wenjun Zhou, “Spectral clustering of high-dimensional data exploiting sparse representation vectors”, Neurocomputing, Elsevier, Volume 135, 2014, Pages 229-239

[3] JochenKerdels, Gabriele Peters, “Analysis of high-dimensional data using local input space histograms”, Neurocomputing, Elsevier, Volume 169, December 2015, Pages 272–280

[4] AndrasKiraly, Agnes Vathy-Fogarassy and Janos Abonyia, “Geodesic distance based fuzzy c-medoid clustering – searching forcentral points in graphs and high dimensional data”, Fuzzy Sets and Systems, Elsevier, Volume 286, March 2016, Pages 157-172

[5] Aloysius George, “Efficient High Dimension Data Clustering using Constraint-Partitioning K-Means Algorithm”, The International Arab Journal of Information Technology, Volume 10, Issue 5, September 2013, Pages 467-476

[6] Liping Jing, KuangTian and Joshua Z. Huang, “Stratified Feature Sampling Method for Ensemble Clustering of High Dimensional Data”, Pattern Recognition, Elsevier, Volume 48, Issue 11, November 2015, Pages 3688-3702

[7] Zhiping Lin, Jiuwen Cao, Tao Chen, Yi Jin, Zhan-Li Sun, and AmauryLendasse, “Extreme Learning Machine on High Dimensional and Large Data Applications”, Hindawi Publishing Corporation,

Mathematical Problems in Engineering, Volume 2015, Pages 1-3.

[8] ChenpingHou, FeipingNie, Dongyun Yi and Dacheng Tao, "Discriminative Embedded Clustering: A Framework for Grouping High-Dimensional Data", IEEE Transactions on Neural Networks and Learning Systems, Volume 26, Issue 6, June 2015, Pages 1287-1299

[9] EhsanElhamifar, and Ren´e Vidal, "Sparse Subspace Clustering:Algorithm, Theory, and Applications", IEEE Transactions on Pattern Analysis & Machine Intelligence, Volume 35, Issue 11, November 2013, Pages 2765-2781

[10] Yuqiang Fang, Ruili Wang, Bin Dai, and Xindong Wu, "Graph-Based Learning via Auto-Grouped Sparse Regularization and Kernelized Extension", IEEE Transactions on Knowledge and Data Engineering, Volume 27, Issue 1, January 2015, Pages 142-154.

[11] Yu Gu, ChunpengGao, Gao Cong, and Ge Yu, "Effective and Efficient Clustering Methods for Correlated Probabilistic Graphs", IEEE Transactions on Knowledge and Data Engineering, Volume 26, Issue 5, May 2014, Pages 1117-1130.

[12] DucThang Nguyen, Lihui Chen and Chee Keong Chan, "Clustering with Multiviewpoint-Based Similarity Measure", IEEE Transactions on Knowledge and Data Engineering, Volume 24, Issue 6, June 2012, Pages 988-1001.

[13] R. G. Mehta, N. J. Mistry, M. Raghuwanshi, "Towards Unsupervised and Consistent High Dimensional Data Clustering", International Journal of Computer Applications, Volume 87, Issue 2, February 2014, Pages 40-44

[14] R.Shenbakapriya, M. Kalimuthu, P. Sengottuvelan, "Improving Clustering Performance on High Dimensional Data using Kernel Hubness", International Journal of Computer Applications (IJCA), Pages 27-30, 2014

[15] RashmiPaithankar and Bharat Tidke, "An H-K Clustering Algorithm for High Dimensional Data Using Ensemble Learning", International Journal of Information Technology Convergence and Services (IJITCS) Volume 4, Issue 5/6, Pages 1-9, December 2014

[16] Zhiwen Yu, PeinanLuo, 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, March 2016, Pages 701 – 714

[17] Malik Tahir Hassan, Asim Karim, Jeong-Bae Kim and MoonguJeon, "CDIM: Document Clustering by Discrimination Information Maximization", Information Sciences, Elsevier, Volume 316, September 2015, Pages 87–106.

[18] Thenmozhi Srinivasan and BalasubramaniePalanisamy, "Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence", Hindawi Publishing Corporation, The Scientific World Journal, Volume 2015, April 2015, Pages 1-6.

[19] Ali SeyedShirkhorshidi, Saeed Aghabozorgi, Teh Ying Wah, "A Comparison Study on Similarity and Dissimilarity Measures in Clustering Continuous Data", Plos One, Volume 10, Issue 12, December 2015, Pages 1-20.

[20] Singh Vijendra and SahooLaxman, "Subspace Clustering of High-Dimensional Data: An Evolutionary Approach", Hindawi Publishing corporation,Applied Computational Intelligence and Soft Computing, Volume 2013, 2013, Pages 1-12