Journal of Theoretical and Applied Information Technology
28th February 2018. Vol.96. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

HILBERT SPACE RELATIONAL SCATTERED DISTANCE CLUSTERING FOR DENSELY POPULATED AND SPARSELY DISTRIBUTED HIGH DIMENSIONAL DATA OBJECTS

[1]R.PUSHPALATHA, [2]Dr.K.MEENAKSHI SUNDARAM
[1]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.

[2]Associate Professor, Department of Computer Science, Erode Arts and Science College (Autonomous), Erode, Tamil Nadu, India.
E-mail: [1]rpljour@gmail.com, [2]lecturerkms@yahoo.com

ABSTRACT

Clustering high-dimensional data is the process of grouping the similar data from large amounts of database. Hence it is essential and significant issue in both machine learning and data mining. But, Clustering on high dimensional data has low accuracy and quality of the clustering algorithm is reduced due to the data objects from a variety of clusters in different subspaces consisting of dissimilar groupings of dimensions. The Different clustering algorithm is designed to solve the difficulties, however the cluster objects hides in the subspaces due to it sparseness and low dimensionality. In order to evaluate both sparsely distributed and densely populated data objects in any given plane, Hilbert Space Relational Scattered Distance Clustering (HS-RSDC) technique is introduced. The HS-RSDC using Controlled Effort Boundary Operation namely UNION_BOUND, INTERSECT_BOUND and PARTITION_BOUND to improve the clustering accuracy. This helps to improve multi space data object mining quality in various real world clustering applications. The Hilbert space relational cluster objects are processed for producing the accurate cluster by reducing subspaces in the data plane. After that, Scattered Distance measures are employed to calculate the distance of geometric median. Finally, the HS-RSDC boundary computation operations associate unlabeled data objects to more appropriate cluster using relational object number. The relational object number is assigned internally and globally in HS-RSDC method to remove unlabelled data objects and to discover different types of correlation events over the cluster objects. HS-RSDC handles the boundary overhead and improves the efficiency of user pruning queries. By minimizing the number of intermediate pruning, the total traversal length of data object search is reduced. Experimental results show that the proposed HS-RSDC method increases the performance in terms of clustering accuracy, clustering time, space complexity.

Keywords: *Data Mining, High Dimensional Data, Cluster Analysis, Hilbert Space, Pruning Process, Scattered Distance, Controlled Effort Boundary Operation.*

1. INTRODUCTION

Clustering is a significant task of data mining. The function of clustering is determining and grouping similar objects with the principle and the objects in the same group (cluster) are similar. Clustering high dimensional data is a promising research field.

Kernelized Group Sparse graph (KGS-graph) was developed in [1] to exhibits a contextual information of a data manifold. However, sparse graph construction failed to satisfy the locality restriction and it does not combine non-zero coefficients locality and sparsity. Predictive Subspace Clustering (PSC) was developed in [2] to cluster the high-dimensional data. However, it failed to consider both dense and sparse high dimensional data objects simultaneously.

To perform both sparse and dense high dimensional data objects, a robust multi objective subspace clustering (MOSCL) algorithm was designed in [3]. However, the analysis of attribute relevancy remained unsolved.

An Incremental Semi-Supervised Clustering Ensemble (ISSCE) approach was introduced in [4] for high dimensional data clustering. However, the effectiveness of the framework was remained unaddressed. Fuzzy Clustering Algorithms was developed in [5] functioned on relational input data. But, the fuzzy clustering using the prototypes or

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Gaussians mixture was not appropriate to sentence clustering.

A fast clustering-based feature selection algorithm (FAST) was introduced in [6] for high dimensional data. But, it failed to perform the correlation measures between the high dimensional data events in feature space.

The Principal Component Analysis (PCA) was described in [7] for detecting the data's in sentence clustering. But, it failed to consider a class labels, and hence fewer related components may still contains high discriminatory value.

A Sparse Subspace Clustering (SSC) was introduced in [8] using the sparse representation to cluster data points. However, the sparse optimization program and spectral clustering was not applicable to large datasets.

A graph-based clustering method was designed in [9] for multidimensional datasets but it failed to improve clustering efficiency at a required level. A generalization of the Gaussian mixture method was introduced in [10] to automatically recognize natural aspects of data and group the data along with features concurrently. But it was not possible that the single clustering produced by a variable selection method.

The several issues are identified from above reviews such as difficult to handles both dense and sparse high dimensional data, lack of clustering accuracy, and difficult to find attribute relevancy and correlation.

High dimensional data includes many of the dimensions that introduce less correlation to clustering process. This lack of correlation complicates clustering process by hiding clusters in noisy data. Besides, curse of dimensionality is handled by high dimensional data clustering algorithms. This means that distance measures become insignificant as the number of dimensions improves in the dataset.

In order to overcome such kind of issues, Hilbert Space Clustered relational scattered distance clustering (HS-RSDC) technique is introduced.

The contribution of the paper is described as follows,To improve the clustering accuracy of sparsely distributed and densely populated high dimensional data objects, the Hilbert Space Clustered relational scattered distance clustering (HS-RSDC) technique is introduced.

Hilbert space relational cluster objects are processed to produce accurate and time efficient cluster by measuring Geometric Median of similar sparse and dense high dimensional data and scattered distance between the two data objects. This helps to reduce the cluster subspace.

The relational object number is allocated for each cluster objects internally and globally for removing the unlabelled data objects and determining the correlation events over the cluster objects.

The rest of the paper is ordered as follows: In Section 2, the proposed Hilbert Space Clustered relational scattered distance clustering (HS-RSDC) technique is described with neat diagram. In Section 3, Experimental settings are presented and the result and discussion is explained in Section 4. Section 5 introduces the background and reviews the related works. Section 6 provides the conclusion.

2. HILBERT SPACE RELATIONAL SCATTERED DISTANCE CLUSTERING TECHNIQUE

The Hilbert space relational scattered distance clustering (HS-RSDC) technique is developed to cluster sparsely and densely populated high dimensional data objects.

Initially, Hilbert Space and Controlled Effort Boundary Operation are used to perform the clustering process. This increases the clustering accuracy in an effective manner.

After that, the geometric median is calculated to specify the clustering data points for high dimensional data. This aids to reduce the space complexity.

Next, scattered distance measure is performed to compute the distance between the inner and outer cluster objects with clustering minimum time.

Finally, the relational object number is assigned for each cluster objects to eliminate the unlabelled data to find the correlation between the events over the cluster object.

This further improves the performance of high dimensional data clustering in HS-RSDC technique.

The overall architecture diagram of HS-RSDC technique is shown in below Figure 1. The high dimensional data are taken from the El Nino dataset to perform clustering. The dataset consists of the collection of densely and sparsely populated high dimensional data objects.

Journal of Theoretical and Applied Information Technology
28th February 2018. Vol.96. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

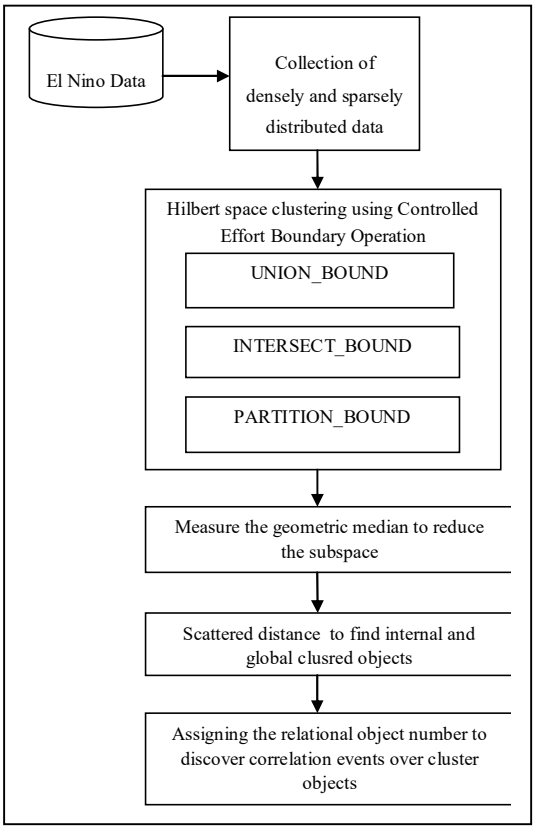

Figure 1: Overall Architecture Diagram Of Hilbert Space Relational Scattered Distance Clustering Technique

The clustering of data objects (i.e. data points) is performed using Hilbert space clustering technique through the three types of boundary operations. After that, the HS-RSDC technique identifies the geometric median of similar sparse data for reducing subspace of data objects in clustered plane. Then the Scattered Distance is used to measure the distance of geometric median for improving the clustering accuracy with minimum time. Finally, correlation events over the cluster objects are determined through the relational object number. This helps to improve the efficacy of user pruning queries. The brief description about HS-RSDC technique is explained in forth coming sections.

2.1 Hilbert Space Clustering Technique

The first step in design of the Hilbert space clustering is performed using Controlled Effort Boundary Operation namely UNION_BOUND, INTERSECT_BOUND and PARTITION_BOUND. Hilbert space clustering is used to cluster the Sparsely Distributed and Densely Populated High Dimensional Data Objects. The El Nino Data set is a spatio temporal dataset which is used to predict the real time data such as Weather conditions around the world. These data are not aligned to predict the weather condition effectively. Therefore, the Hilbert space clustering technique provides the perfect alignment of the different weather data such as air temperature, relative humidity, latitude and so on. These data are clustered separately to evaluate the user query.

During the clustering, the data points are mapped from data space into a high-dimensional Hilbert space. Let us consider the smallest sphere in the Hilbert space that encloses the representation of the data. This sphere is mapped reverse to data space, where it forms a set of contours which enclose the data objects.

These contours are defined as cluster boundaries. Let us consider the set of data points $DP_1, DP_2, \ldots, DP_n \in \{x_i\}$ in high dimensional data space. Mapping function (φ) is used for mapping the data points from data space into Hilbert space (H). Therefore, the mapping process is described as follows,

$$\varphi: x_i \rightarrow H \qquad (1)$$

In high dimensional data clustering, Hilbert curve is used to identify similar data objects and to form a cluster with high accuracy. Hilbert curve is a continuous path that goes across each data objects in a Hilbert space provides direct link among the coordinates of the data objects. It also achieves better ordering of high dimensional data objects in the node. A Hilbert curve is termed as Hilbert space-filling curve.

The Hilbert curve on a 2x2 grid which is represented by H_1 and the next 4x4 grid is represented by H_2. The Hilbert curve with 2x2 grids is shown in figure 2.

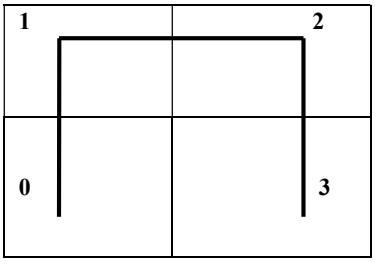

Figure 2: Hilbert Curve On First Order (H_1)

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

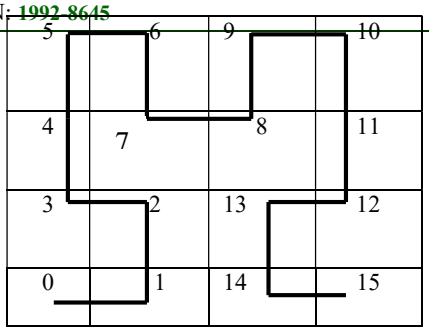

Figure 3: Hilbert Curve On Second Order (H_2) For Sparse And High Dimensional Data Objects

Figure 2 and 3 shows the Hilbert curve on first order and second order for sparse and high dimensional data objects. In Hilbert space, the high dimensional data objects are arranged in a rectangular block. Therefore, the actual end of the each data point is calculated. As shown in figure, the two numbers (i.e. data) are continuous in the two dimensional Hilbert space, then the data objects are grouped in one cluster. If it is not continuous, they are clustered in different cluster. This process is continuous for all the rectangular blocks. During the clustering, the Controlled Effort Boundary Operation is performed to improve the clustering accuracy using UNION_BOUND, INTERSECT_BOUND and PARTITION_BOUND.

When sparsely distributed and densely populated high dimensional data points are combined together in Hilbert space, the UNION_BOUND process is used for improving the clustering process. INTERSECT_BOUND operation in HS-RSDC technique is used to control the data points while two or more data points are intersected. Finally, the PARTITION_BOUND is used to control the data points which are separated into a disjoint cluster.

As shown in figure 4, the sparse and dense high dimensional data objects are clustered using Hilbert space clustering. From the figure, it is clearly illustrates that the clustering on data space into Hilbert space with mapping function (φ). The clustered data with rose colour indicates the sea surface temperature data whereas yellow coloured data indicates the air temperature data. In addition, the red colour clustered data is a rainfall data, violet and blue colour clustered data shows that the relative humidity data and surface temperature data respectively. Similarly, all the weather data in El Nino dataset are clustered separately to prune the user requested data with higher levels of accuracy with minimal time.

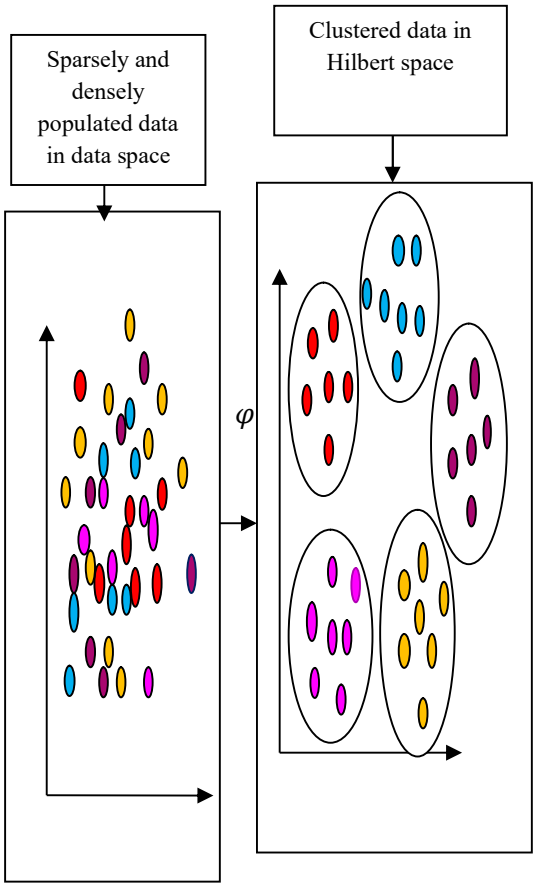

Figure 4 Hilbert spaces clustering on sparse and dense high dimensional data

The Hilbert spaces clustering algorithm is described as follows,

Input: Collection of sparse and dense populated high dimensional data in El nino dataset, data block B[i] is varied from 0 to n-1.
Output: Improve clustering accuracy
Step 1:Begin
Step 2: For each sparse and dense populated high dimensional data
Step 3: Mapping the data space into Hilbert space using (1)
Step 4: Perform Controlled Effort Boundary Operation
Step 5: For each rectangle box in Hilbert space
Step 6: if block status B[i]≠ 0 then
Step 7: if block 'i' objects and their previous block 'i-1' objects are continuous
Step 8: Group the two data objects in same cluster
Step 9: else
Step 10: Group the two objects in different cluster
Step 11: End if
Step 12: End if

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Step 13: End for
Step 14: End for
Step 15: End

Figure 5 :Hilbert Spaces Clustering Algorithm

Figure 5 shows the step by step process of the Hilbert space clustering algorithm. For each sparse and dense high dimensional data, the mapping of data space into Hilbert space is performed using mapping function. The Controlled Effort Boundary Operation is used in Hilbert spaces clustering to identify combined data, intersect data and separated data. Then the each rectangular box in Hilbert space is checked if it is empty or not. If the block status is not equal to zero (i.e. contain the data objects), then it verifies each block and their previous block are continuous or not. If it is continuous, then it is grouped in similar cluster. Otherwise, it grouped into different cluster. As a result, the clustering accuracy is increased.

2.2 Geometric Median And Scattered Distance Measures

After the clustering on high dimensional data, The Hilbert space relational cluster objects are processed to produce the accurate and time efficient cluster by reducing subspaces in the data plane. In HS-RSDC technique, Geometric Median of similar sparse and dense high dimensional data is measured to reduce the cluster subspace. Let us consider, 'd' denotes the dimensional data points '$DP = DP_1, DP_2, \dots, DP_n$' in data space '$DS$' $DP \in DS$. With the given set of data points, the geometric median of the sparsely distributed and densely populated high dimensional data points are measured to determine the objects to be placed on each cluster is as given below.

$$Geometric\ median = arg\min \sum_{i=1}^{n} ||p_i - q||_2, q \in D \quad (2)$$

From (2), '$arg\ min$' specifies the value of the argument 'q' which minimizes the sum. In this case, the point 'q' from where the sum of all Euclidean distances to the 'p_i' is minimum. Therefore, the geometric median of similar sparse and dense data is obtained. Once, the similar sparse and dense data objects are obtained, then the non similar sparse are measured as follows.

$$q = \sum_{i=1}^{n} \frac{p_i - q}{||p_i - q||} \quad (3)$$

$$where\ q = \frac{\left(\sum_{i=1}^{n} \frac{p_i}{p_i - q}\right)}{\left(\sum_{i=1}^{n} \frac{1}{||p_i - q||}\right)}, p_i \neq q \quad (4)$$

From (4), 'q' denotes a non similar sparse and dense data objects. Finally, the geometric medians of similar sparse and dense data are selected to obtain less subspace of the cluster objects in a plane.

Input: High dimensional data points '$DP = DP_1, DP_2, \dots, DP_n$', data space 'DS'

Output : Reduce the space complexity
Step 1: Begin
Step 2: For each data points 'DP' in cluster
Step 3: Calculate geometric median for selecting similar sparse data objects using (2)
Step 4: Calculate non similar sparse and dense data objects using (3)
Step 5: End for
Step 6: End

Figure 6: Geometric Median Algorithm

Figure 6 describes the algorithmic representation of the geometric median to find the data object is placed on each cluster for reducing the subspace. For each data point in cluster, the geometric median is performed to select the similar sparse and dense data objects inside the cluster. Moreover, the HS-RSDC technique also calculates the non similar sparse and dense data objects. As a result, the similar sparse and dense data objects provide the relational data. This helps to reduce the space complexity on high dimensional data clustering.

2.2.1 Scattered distance measure

After the geometric median measure, scattered distance metric is applied in HS-RSDC technique that reflects within cluster scatter (i.e.,) inner object and between the cluster scatter (i.e.,) outer object. The scattered distance measure is performed between the two data points 'DP_i' and 'DP_j' in two different clusters '$Cl(i = K)$' and '$Cl(j = K)$' is as given below.

$$ST = \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{m} Dis\left(DP_i, DP_j\right) \quad (5)$$

$$ST = \frac{1}{2}\sum_{k=1}^{K}\sum_{Cl(i=K)}\left(\sum_{Cl(j=K)} Dis\left(DP_i, DP_j\right) + \sum_{Cl(j\neq K)} Dis\left(DP_i, DP_j\right)\right) \quad (6)$$

From (5) and (6), the scatter distance is measured (ST) based on within the cluster scatter

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

CS_W and between the clusters CS_B are measured as follows.

$$CS_W = \sum_{Cl(i=K)}(p_i - r_k)^2 \quad (7)$$

$$CS_B = \sum_{Cl(i=K)}(r_k - m)^2 \quad (8)$$

From (7) and (8), within cluster scatter and between cluster scatter with similar object position is mathematically obtained. 'p', 'r' and m denotes an uncertain data objects. Different data objects are mapped into various clusters. This in turn minimizes the clustering time.

2.3 Relational Object Number Assignment For Correlation Events Determination

In HS-RSDC, a boundary computation operation includes unlabeled data objects. These objects are eliminated through assigning the relational object number. The relational object number is assigned internally and globally to eliminate unlabelled data objects and to determine different types of correlation events over the cluster objects.

The clustered data objects are labelled as the relational number as $C_1, C_2, C_3, C_4 \ldots . C_n$. These labelled cluster objects are used to identify the types of correlation events over the cluster objects. In general, correlation provides the linear relationship between two cluster objects. Therefore, the relationships between two correlation events are measured as follows,

$$Correlation = \rho = \frac{COV\,(e_1,e_2)}{\sigma_{e_1}\sigma_{e_2}} \quad (9)$$

From (9),$COV\,(e_1, e_2)$ is the covariance function which is the number that measures the common variation of two events e_1 and e_2. $\sigma_{e_1} and\ \sigma_{e_2}$ is the standard deviation of the e_1 and e_2 respectively. The correlation measurement provides the value between -1 and 1. If the correlation is 1, then the types of relation events are discovered. Otherwise, there is no relations events are determined.

As a result, HS-RSDC handles the boundary overhead based on the number of boundary objects and the user pruning queries. By minimizing the number of intermediate pruning, the total traversal length of data object search is reduced and the query evaluation rate improves the performance significantly.

3. EXPERIMENTAL EVALUATION

The Hilbert Space Relational Scattered Distance Clustering (HS-RSDC) technique is implemented in Java Language using El Nino dataset from UCI machine learning repository. El Nino Data Set consist of oceanographic and surface meteorological readings considered from a series of uncertain buoys positioned all over the equatorial pacific with 178080 instances. El Nino Data Set consists of 12 attribute set includes and subsurface temperatures down to a depth of 500 meters. The characteristics of attributes are represented as integer, real. Similarly, the dataset characteristics are specified as Spatio-temporal.

4. RESULT AND DISCUSSION

Result analysis of Hilbert Space Relational Scattered Distance Clustering (HS-RSDC) technique is described in this section. The HS-RSDC is compared against with the existing Kernelized Group Sparse graph (KGS-graph) [1], Predictive Subspace Clustering (PSC) [2] and Multi Objective Subspace CLustering (MOSCL) algorithm [3]. An experiment is conducted on the factors such as clustering accuracy, clustering time, and space complexity with the number of data objects (i.e. attributes). Experimental results are compared and analyzed with the help of table and graph.

4.1 Impact Of Clustering Accuracy

Clustering accuracy is measured as the ratio of number of data objects correctly clustered to the total number of data objects. It is measured in terms of percentage (%) and it is expressed as follows,

$$Clustering\ accuracy = \frac{No.of\ \ data\ objects\ correctly\ clustered}{Total\ no.of\ data\ objects} * 100 \quad (10)$$

By using (10), clustering accuracy is measured. While the clustering accuracy of sparse and dense populated high dimensional data is higher, the method is said to be more efficient.

Table 1 describes the clustering accuracy with number of iterations. Among the data objects in El-Nino dataset, the number of data objects is correctly clustered. Therefore, the clustering accuracy is improved using HS-RSDC technique than the existing KGS-graph [1], PSC [2] and MOSCL algorithm [3].

Journal of Theoretical and Applied Information Technology
28th February 2018. Vol.96. No 4

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Table 1:Tabulation For Clustering Accuracy

Number of iterations	Clustering accuracy (%)			
	HS-RSDC	KGS-graph	PSC	MOSCL
5	91.36	79.28	82.45	83.12
10	92.68	83.16	85.12	88.64
15	94.35	86.29	87.07	89.10
20	90.36	82.12	83.46	85.65
25	93.69	86.26	88.58	90.46
30	95.57	85.19	87.34	91.12
35	97.12	89.84	92.47	93.14

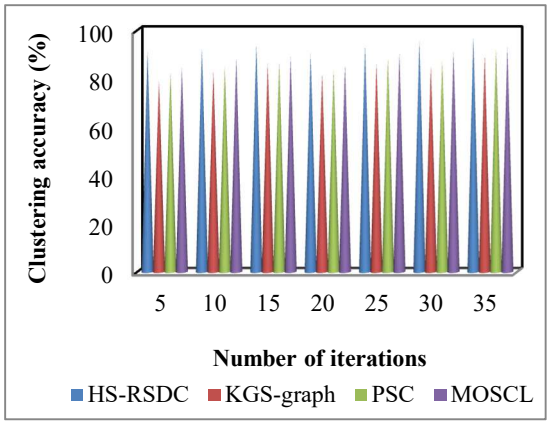

Figure 7: Measure Of Clustering Accuracy

Figure 7 shows the Clustering accuracy analysis based on the number of data objects in El Nino Dataset. From the figure, the clustering accuracy is increased using proposed HS-RSDC technique than the existing methods. This is because, Hilbert space clustering algorithm is applied in HS-RSDC technique. During the clustering, the Hilbert curve is plotted to efficiently group the high dimensional sparse and dense data objects. Followed by this, the rectangular block with data objects in Hilbert space is verified whether the two data objects are continuous or not. This helps to improve the clustering the data objects. This can also increase the clustering accuracy. The clustering accuracy is increased by 11 %, 8% and 5% compared to existing KGS-graph [1], PSC [2] and MOSCL algorithm [3] respectively.

4.2 Impact Of Clustering Time

Clustering time is the amount of time taken to cluster the data objects with respect to number of data objects. The formula for clustering time is measured as follows,

$$CT = No.of\ data\ objects * Time\ (clustering) \quad (11)$$

From (11), clustering time 'CT' is measured in terms of milliseconds (ms). Less clustering time, more efficient the method is said to be.

Table 2: Tabulation For Clustering Time

Number of data objects	Clustering time (ms)			
	HS-RSDC	KGS-graph	PSC	MOSCL
2	3.5	5.2	6.7	4.8
4	5.3	8.9	10.6	6.7
6	9.8	13.5	15.7	11.3
8	12.9	17.9	20.6	15.7
10	10.6	15.2	18.9	12.5
12	16.5	21.4	25.8	19.6

Table 2 shows the measurement of clustering time with respect to number of data objects. The data objects are varied from 2 to 12. From the table value, it is clearly illustrates that the clustering time of proposed HS-RSDC is reduced when compared to the existing methods KGS-graph [1], PSC [2] and MOSCL algorithm [3].

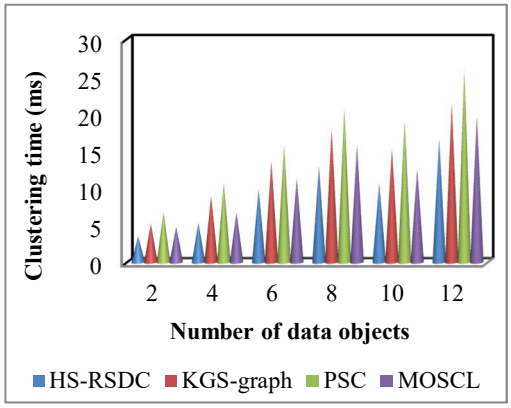

Figure 8: Measure Of Clustering Time

Figure 8 illustrates the performance analysis of clustering time. As shown in figure, proposed HS-RSDC technique achieves higher clustering accuracy with minimum time when compared to existing methods [1] [2] [3]. By applying clustering, The Hilbert space relational cluster objects are processed through the Scattered Distance measures the distance of inner object and similar object position to reduce the clustering time. The data objects are closed in Hilbert space and select similar data objects close together. As a result, the clustering accuracy is significantly improved with minimum time. The proposed HS-

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

RSDC technique considerably reduces the clustering time of sparely distributed and densely populated high dimensional data by 30%, 42% and 18% as compared to existing KGS-graph [1], PSC [2] and MOSCL algorithm [3] respectively.

4.3 Impact Of Space Complexity

Space complexity is defined as the amount of memory used for storing the clustered object in the cluster with respect to number of data objects. The formula for space complexity is measured as follows,

$$SC = No.of\ data\ objects * Memory\ (clustered\ objects) \quad (12)$$

From (12), SC represents the space complexity which is measured in terms of Mega Bytes (MB). Less space complexity, the method is said to be more efficient.

Table 3 Tabulation for Space complexity

Number of data objects	Space complexity (MB)			
	HS-RSDC	KGS-graph	PSC	MOSCL
2	12	19	23	15
4	15	30	35	25
6	22	45	47	36
8	26	51	56	42
10	32	68	70	55
12	37	79	83	65

Table 3 describes the comparative result analysis of space complexity for clustering the sparely distributed and densely populated high dimensional data using four methods namely HS-RSDC technique, KGS-graph [1], PSC [2] and MOSCL algorithm [3]. Space complexity is reduced using proposed HS-RSDC technique than the existing methods.

Figure 9 depicts the performance comparison results of space complexity with respect to the number of data objects. The significant improvement is achieved in proposed HS-RSDC technique than the existing methods. This is because, the geometric median between the data objects are measured to reduce the subspace in cluster plane.

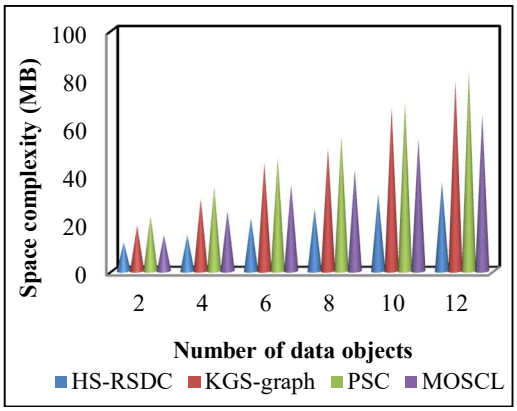

Figure 9: Measure Of Space Complexity

Therefore, the similar sparse and dense selected data objects provide the relational data for improving the user pruning queries. This helps to reduce the space complexity on high dimensional data clustering. Therefore, the space complexity is significantly reduced by 49%, 54% and 37% compared to existing KGS-graph [1], PSC [2] and MOSCL algorithm [3] respectively.

5. RELATED WORKS

A new clustering algorithm (SUBSCALE) was introduced in [11] for discovering a non-trivial subspace clusters with minimum cost for a k-dimensional data set. It successfully determines quality subspace clusters without requiring exclusive database scans. However, it failed to perform clustering on densely populated high dimensional data objects. The proposed HS-RSDC technique uses Hilbert space clustering to cluster both spare and dense populated high dimensional data objects.

Density based Subspace clustering algorithms was introduced in [12] to perform clustering on high dimensional data. It performs better by producing the clusters of adaptive size, shape, densities and dimensionalities. But it takes more space for storing the clustered objects. The proposed HS-RSDC reduces the space complexity through the measure of geometric median between the data objects. The mutual subspace clustering was introduced in [13] using bottom up approach and top down approach. It provides the inherent connection between the two different set features. But, the time for clustering the data objects was not solved. The HS-RSDC takes minimum amount of time for clustering on high dimensional data objects.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

A new clustering algorithm was presented in [14] using FG-k-means to cluster high-dimensional data from subspaces of feature groups and individual features. But the algorithm achieved minimal clustering accuracy. The HS-RSDC technique significantly achieves higher accuracy through Hilbert space clustering.

Hub-based algorithm was designed in [15] for improving the clustering on high dimensional data. However, it failed to accurately handle the clusters of random shapes. The proposed HS-RSDC technique effectively handles the structure of cluster using Hilbert space clustering.

Hierarchical clustering was performed in [16] which contain highly correlated variables. But it takes more time to perform clustering with higher complexity. The proposed HS-RSDC techniques significantly reduce the clustering time and complexity.

An improved Density Based Spatial Clustering of Application of Noise (IDBSCAN) algorithm is designed in [17]. But it failed to consider the neighbourhood objects and graphs. The HS-RSDC technique considered the neighbourhood data objects to improve the clustering accuracy.

A novel constraint based multi-dimensional data-clustering algorithm is developed in [18] to discover the number of clusters in a multi-dimensional data set. However, It takes more time complexity. The HS-RSDC technique achieves higher clustering accuracy with minimum time.

An efficient measure known as Joint optimum similarity eccentric is designed in [19] to accurately calculate the distance between points in multiple forms of data. However, the space complexity of the method is not considerably reduced. The proposed HS-RSDC technique performs efficient analysis on space complexity parameter.

A semi-supervised clustering algorithm based on active learning is designed in [20] to increase clustering efficiency. But, this algorithm is not sufficient to process the massive data sets. The HS-RSDC technique is developed for handling high dimensional data objects to improve the clustering accuracy.

As a result, a Hilbert Space Relational Scattered Distance Clustering (HS-RSDC) technique is developed to perform clustering on sparse and dense high dimensional data object with minimum time and space complexity.

6. CONCLUSION

An efficient HS-RSDC technique is developed to cluster the sparsely distributed and densely populated high dimensional data. The HS-RSDC uses Controlled Effort Boundary Operations to achieve higher clustering accuracy. In clustering, the Hilbert space-filling curve uses a continuous path that traverses each data objects in a Hilbert space to provide the direct link to coordinate of the data objects. After that, a new Geometric Median is measured between data objects by selecting similar and non similar data objects to reduce the cluster subspace. This in turns space complexity is minimized in HS-RSDC technique. Next, efficient Scattered Distance measure the distance between the two data points and within cluster to reduce the clustering time. Finally, the relational cluster object number is assigned for determining the types of correlation events over the cluster objects. Experimental evaluation is performed using El Nino weather data sets from UCI research repository. The performance results show that the proposed HS-RSDC significantly improves the clustering accuracy with minimum time and also reduces the space complexity than the state-of-the-art methods.

7. CONFLICTS

The problem with this HS-RSDC technique is that they convert many dimensions to a single set of dimensions which later makes it difficult to interpret the results. Also, this technique is insufficient if the clusters are in diverse subspaces of the dimension space.

REFERENCES:

[1] 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.

[2] Brian McWilliams, Giovanni Montana, “Subspace clustering of high-dimensional data: a predictive approach”, Data Mining and Knowledge Discovery, Springer, Volume 28, Issue 3, 2013, Pages 736–772

[3] Singh Vijendra and Sahoo Laxman, “Subspace Clustering of High-Dimensional Data: An Evolutionary Approach”, Hindawi Publishing Corporation, Applied Computational Intelligence and Soft Computing , Volume 2013, November 2013, Pages 1-12

[4] Zhiwen Yu, Peinan Luo, Jane You, Hau-San Wong, Hareton Leung, Si Wu, Jun Zhang, Guoqiang Han, “Incremental Semi-supervised Clustering Ensemble for High Dimensional

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

Data Clustering", IEEE Transactions on Knowledge and Data Engineering, Volume: 28, Issue: 3, March 2016, Pages 701 – 714

[5] Andrew Skabar, Member, and Khaled Abdalgader, "Clustering Sentence-Level Text Using a Novel Fuzzy Relational Clustering Algorithm", IEEE Transactions On Knowledge and Data Engineering, Volume 25, Issue 1, 2013, Pages 62 - 75

[6] Qinbao Song, Jingjie Ni, and Guangtao Wang, "A Fast Clustering-Based Feature Subset Selection Algorithm for High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering , Volume 25, Issue 1, 2013, Pages 1 - 14

[7] Ludmila I. Kuncheva, Member, and William J. Faithfull, "PCA Feature Extraction for Change Detection in Multidimensional Unlabeled Data", IEEE Transactions On Neural Networks and Learning Systems, Volume 25, Issue 1, 2014, Pages 69 - 80

[8] Ehsan Elhamifar and René Vidal, "Sparse Subspace Clustering: Algorithm, Theory, and Applications", IEEE Transactions on Pattern Analysis and Machine Intelligence , Volume 35, Issue 11, Nov. 2013, Pages 2765 – 2781

[9] Peixin Zhao, Cun-Quan Zhang, Di Wan, and Xin Zhang, "A Multidimensional and Multimembership Clustering Method for Social Networks and Its Application in Customer Relationship Management", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2013, August 2013, Pages 1-8

[10] Leonard K.M. Poon, Nevin L. Zhang, Tengfei Liu, April H. Liu, "Model-based clustering of high-dimensional data: Variable selection versus facet determination", International Journal of Approximate Reasoning, Elsevier, Volume 54, 2013, Pages 196–215

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

[12] Sunita Jahirabadkar and Parag Kulkarni, "Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms", International Journal of Computer Applications, Volume 63, Issue 20, 2013, Pages 29-35

[13] Ming Hua a, JianPei, "Clustering in applications with multiple data sources—A mutual Subspace clustering approach" Neuro computing, Elsevier, Volume 92 , 2012, Pages 133–144

[14] Xiaojun Chen, YunmingYe , XiaofeiXu , Joshua Zhexue Huang, "A feature group weighting method for subspace clustering of high-dimensional data" Pattern Recognition, Elsevier, Volume 45, 2012, Pages 434–446

[15] Nenad Tomasev , Milos Radovanovic, Dunja Mladenic, Mirjana Ivanovic, "The Role of Hubness in Clustering High-Dimensional Data", IEEE Transactions on Knowledge and Data Engineering ,Volume 26, Issue 3, March 2014, Pages 739 - 751

[16] Jie Zhang and Meng Pan, "A high-dimension two-sample test for the mean using cluster subspaces", Computational Statistics and Data Analysis, Elsevier, Volume 97, 2016, Pages 87–97

[17] Arvind Sharma, R. K. Gupta, and Akhilesh Tiwari, "Improved Density Based Spatial Clustering of Applications of Noise Clustering Algorithm for Knowledge Discovery in Spatial Data", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2016, June 2016, Pages 1-9

[18] B.Kranthi Kiran, A. Vinaya Babu, "An Algorithm to Constraints Based Multidimensional Data Clustering Aided With Associative Clustering", Journal of Theoretical and Applied Information Technology (JATIT), Volume 70, Issue 2, 2014, Pages 241-250

[19] M.Suguna, Dr.S.Palaniammal, "Jose Measure Based High Dimensional Data Clustering for Real World Conditions", Journal of Theoretical and Applied Information Technology (JATIT), Volume 67, Issue 2, 2014, Pages 361-368

[20] Zhang Chun Na, Zhu Yong Yong, Li Yi Ran, "An Improved Semi-Supervised Clustering Algorithm Based on Active Learning", Journal of Theoretical and Applied Information Technology (JATIT), Volume 48, Issue 2, 2013, Pages 741-748