R-Tree Based Pruning Process

R.Pushpalatha
Ph.D. Research Scholar in Computer Science,
Erode Arts and Science College, Erode
and Assistant Professor, Department of Computer
Science, Kongu Arts and Science College
(Autonomous), Erode, Tamilnadu.

Dr.K.Meenakshi Sundaram
Associate Professor, Department of Computer
Science, Erode Arts and Science College,
Erode,Tamilnadu.
(Email Id.:lecturerkms@yahoo.com)

Abstract - Clustering is an essential and significant topic in both machine learning and data mining. Recently, most of researches have been developed for clustering high dimensional data. However, there is a need for effective clustering model to cluster the densely populated high dimensional data objects with higher clustering accuracy. In order to overcome such limitation, Hilbert Space Clustered Indexing (HSCI) technique is proposed. HSCI technique work on the densely populated data objects in the Hilbert space dimensionality to easily identify symmetric objects and generates cluster with high accuracy level. Initially, HSCI technique is designed Hilbert Space Dimensional Clustering algorithm to efficiently cluster the densely populated high dimensional data objects. Next, The HSCI technique constructs an R-tree for indexing the clustered high dimensional data. Finally, the HSCI technique performs R-tree based pruning process to efficiently prune the user requested data with higher accuracy level. The performance of HSCI technique is measured in terms of clustering accuracy and clustering time. The experimental result shows that the HSCI technique is able to improve the clustering accuracy by 18% and also reduces the clustering time by 19% when compared to state-of-the-art-works.

Keywords-Clustering; Densely Populated High Dimensional Data Objects; Hilbert space; R-tree, indexing; R-tree based pruning process.

I. INTRODUCTION

Data mining refers to extorting or mining knowledge from large amounts of database. With the development of technology, it is easier to gather data in different fields, thus resulting in increasingly large and complex datasets. Clustering has been extensively utilized in numerous areas, ranging from science to engineering. The key objective of clustering is to assign data of similar patterns into the same cluster and expose the meaningful structure of the data.

In recent times, many research works has been designed for clustering high dimensional data. For example, a Predictive Subspace Clustering (PSC) was developed in [1] for clustering high-dimensional data. However, PSC is not considered for densely populated high dimensional data objects.The Discriminative Embedded Clustering (DEC) was applied in [2] for clustering high dimensional data which integrates subspace learning and clustering in a common procedure. But, DEC takes more time for clustering process.

A spectral clustering algorithm was designed in [3] based on collaborative representation coefficient vectors (CRSC) for clustering high-dimensional data and decreasing the time cost of clustering. But, the clustering accuracy is not at required level. A clustering-based feature subset selection algorithm was explained in [4] for clustering high dimensional data where each cluster is considered as a single feature and therefore the dimensionality was reduced. However, the computational complexity of this algorithm was higher.

A novel techniques was planned in [5] with the help of high-dimensional similarity based PCM with ant colony optimization intelligence to cluster the high dimensional data in projected space. But, this technique takes more memory for storing the high dimensional data. Besides, SNN similarity based smooth splicing clustering algorithm was introduced in [6] that used a complementary intensity-smoothness mechanism for clustering high-dimensional data.

A robust multi objective subspace clustering (MOSCL) algorithm was implemented in [7] for addressing the issues of high-dimensional clustering which results in improved accuracy of subspace clustering. An incremental semi supervised clustering ensemble approach (ISSCE) was explained in [8] that makes use of the gain of the random subspace technique, the constraint propagation approach to carry out high dimensional data clustering.

A novel H-K clustering algorithm was presented in [9] to lessen the computational complexity and to improve the accuracy of high dimensional data clustering. Though, space complexity remained unsolved. The Constraint-Partitioning K-Means (COP-KMEANS) clustering algorithm was designed in [10] for clustering high dimension dataset and to decrease computational cost via eliminating the noisy dimensions.

Based on the aforementioned techniques and methods presented, in this work we propose a novel framework called as Hilbert Space Clustered Indexing (HSCI) technique for effectively clustering the densely populated high dimensional data objects.

The rest of the paper organized as follows. In Section 2, a summary of different techniques designed for high dimensional data clustering is presented. In Section 3, the proposed HSCI technique is explained with the help of neat architecture diagram. In Section 4, simulation environment is discussed with exhaustive analysis of results described in Section 5. In Section 6, the concluding remarks are discussed.

II. RELATED WORKS

In spatial data mining, clustering is one of helpful techniques for identifying interesting data in the underlying data objects. Density-based Clustering algorithm was designed in [11] for data clustering with numerous properties and applications where clusters are made according to the density of the data. However, the clustering accuracy was poor. A feature selection based clustering method called IQRAM (Inter Quartile Range and Median) was designed in [12] for clustering high dimensional dataset and minimizing the effects of high dimensionality and choosing the initial clusters centres efficiently.

In addition, Hierarchical Accumulative Clustering Algorithm was presented in [13] for clustering high dimensional data which results in enhanced the clustering accuracy. But, it requires more memory space and also takes more running time for clustering process. A weighted clustering ensemble algorithm was developed in [14] to provide an enabling technique to support and to integrate any input partitions. But, this algorithm does not perform well due to information loss in the representation extraction.

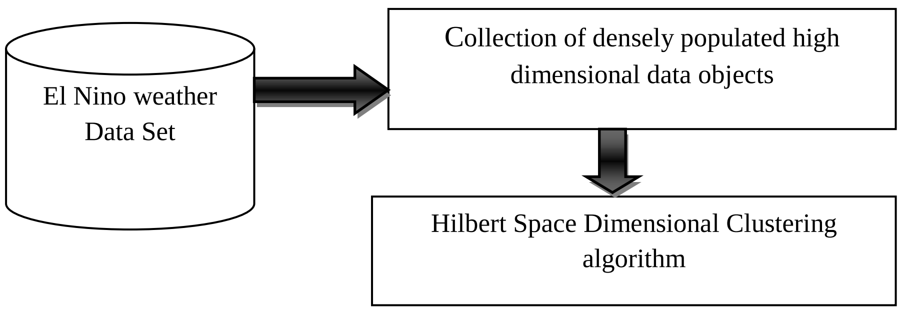
A novel algorithm was presented in [15] based on the combination of kernel mapping and hubness phenomenon to enhance the performance of clustering high dimensional data and to improve the clustering quality. However, the clustering time remained unaddressed. Auto-Associative Neural Networks (AANN) technique was planned in [16] to cluster the high-dimensional data and to reduce dimension of high-dimensional data. However, the clustering performance is not effective.

A modified PROCLUS algorithm called MPROCLUS was developed in [17] with objective of clustering high dimensional data and improving the running time and consistency. An Auto-Associative Neural Networks was designed in [18] to accomplish compression, clustering and visualization of high-dimensional data with the aim of improving data clustering accuracy. But, the clustering of complex multidimensional data remained unsolved.

Clustering ensemble method based on two-staged clustering algorithm was presented in [19] to enhance the efficiency and accuracy of clustering of high dimensional data. In [20], a classification algorithm for high dimensional data was designed to effectively handle large scale high dimensional data by using Kohonen neurons and to reduce dimensionality. Based on the above mentioned methods and techniques, the following proposed work is designed to provide an appropriate solution to solve the existing issues.

III.HILBERT SPACE CLUSTERED INDEXING TECHNIQUE

The Hilbert Space Clustered Indexing (HSCI) technique is developed to cluster densely populated high dimensional data objects. The HSCI technique is quite effective in clustering dense data points and improves the user's pruning efficiency for efficient data mining. The HSCI technique designs an effective Hilbert Space Dimensional Clustering Algorithm to group the densely populated high dimensional data objects with higher clustering accuracy. After performing the clustering process, HSCI technique build an R-tree for indexing the clustered high dimensional data. Finally, HSCI technique adopts R-tree based pruning process to extract the user requested data from the R-tree with higher levels of user pruning efficiency. The overall architecture diagram of HSCI Technique for clustering the densely populated high dimensional data objects is shown in below Figure 1



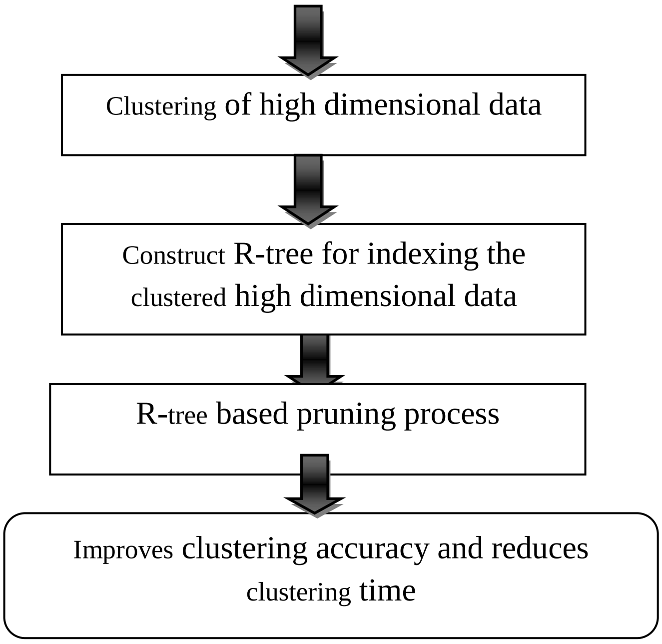


Figure 1 Overall Architecture Diagram of HSCI Technique for Clustering the Densely Populated High Dimensional Data Objects

As shown in Figure 1, HSCI Technique initially takes the El Nino weather dataset as input which comprises collections of densely populated high dimensional data objects. After that, HSCI Technique is used Hilbert Space Dimensional clustering algorithm with aiming at improving the clustering accuracy of high dimensional data and reducing clustering time. Then, HSCI Technique creates an R-tree for indexing the clustered high dimensional data. Finally, HSCI technique accomplishes R-tree based pruning process to efficiently retrieve the user requested data from the R-tree. This in turn improves the user pruning efficiency in an effective manner.

A.Hilbert Space Dimensional Clustering Algorithm

In HSCI Technique, Hilbert space dimensional clustering algorithm is used to efficiently cluster the densely populated high dimensional data in El Nino weather dataset. In Hilbert space dimensional clustering, Hilbert curve is generated to identify symmetric data objects and to generate cluster with high accuracy level. Hilbert curve is a continuous path that traverses each data point in a space once to link a one to one connection among the coordinates of the data points and the one-dimensional sequence numbers of the points on the curve. In HSCI Technique, Hilbert curve protect the distance of that data points which are close in space and indicate similar data should be stored close together in the linear order. This kind of mapping provides high speed for clustering densely populated high dimensional data which in turn reduces the clustering time of densely populated high dimensional data in an effective manner.

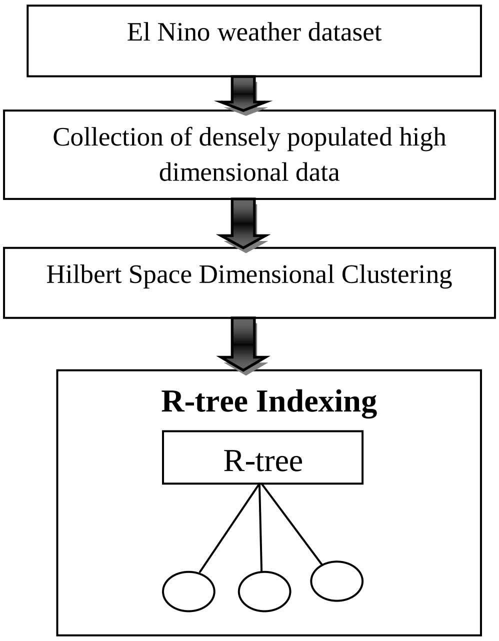
The HSCI technique is used El Nino weather dataset for clustering process where it comprises of variety of data for weather forecasting analysis such as air temperature, relative humidity, surface winds, sea surface temperatures and subsurface temperatures, rainfall and solar radiation etc.

Hilbert Space Dimensional Clustering Algorithm initially separates the densely populated high dimensional data objects into the rectangle blocks. If block i and the previous block l ($l < i$) which has objects are continuous, then place them with the same cluster number else the cluster number is incremented by one (i.e. it moves to next cluster). This is process repeated until all the rectangle blocks of Hilbert curve is reached. Therefore, the accuracy of clustering densely populated high dimensional data is significantly improved with minimum clustering time.

B. R-tree for Indexing Clustered High Dimensional Data

In HSCI Technique, R-tree indexing is an expansion of B+-tree which is used for indexing the clustered high dimensional data. R-tree indexing employs space-filling curves and exclusively the Hilbert curve to impose a linear ordering on the data rectangles. Let us consider high dimensional data are densely populated in Hilbert space. At first Hilbert value is calculated for each high space. At first Hilbert value is calculated for each high dimensional data. After that, the high dimensional data in Hilbert space are sorted based on Hilbert value for indexing clustered data. The sorted list of data rectangles are examined, then successive rectangles are allocated to the same R-tree leaf node until that node is filled. A new leaf node is then generated and the scanning of the sorted list continues. Hence, the nodes of the resulting R-tree will be fully packed with the potential exception of the last node at each level.

In R-tree indexing, the Hilbert curve uses a linear ordering on the data rectangles and then passes through the sorted list, allocates each set of data rectangles to a node in the R-tree. The final result is that the collection of data rectangles on the similar node will be close to each other in the linear ordering and most likely in the native space; hence the resulting R-tree nodes have smaller spaces. This in turn reduces the space complexity in an effective manner. The indexing process of clustered high dimensional data objects is shown in below Fig.2 process of R tree indexing High Dimensional Data.



In HSCI Technique, R-tree indexing significantly reduces the overlap space and also optimizes the retrieval path of index. Therefore, the space complexity and user's pruning efficiency of HSCI Technique is improved in an effective manner.

C.R-Tree Based Pruning Process for Data Retrieval

After indexing the clustered high dimensional data, the HSCI technique performs R-tree based pruning process to cover the lower left, upper left, upper right, lower right places of R-tree for efficient data retrieval process. Thus, HSCI technique provide better results in the effectively pruning and extracting varied data ranges for efficient clustering of various data mining applications such as medical data analysis, aerial data extraction, weather forecasting, etc. The R-tree builds a set of dense data objects in the plane and then performs kernel mapping strategy to prune the user requested data with higher levels of accuracy in minimal time. The R-tree based pruning process removes unwanted data and mines only the user requested query data result from the R-tree.

In HSCI technique, the kernel mapping strategy is employed to map the user query x into a feature space of higher dimensional data objects indexed in R-tree which is mathematically formulated as

$$x \rightarrow \emptyset(x)$$

(1)

From the equation (1), $\emptyset(x)$ represents the mapping function which maps the user query into the specified nodes that contains the user requested data.

R-tree based pruning algorithm initially takes the user query as input and then search specified nodes which contain the user requested data in indexed R-tree. Finally, R-tree based pruning algorithm extracts the data from that searched nodes and provides mined data to the corresponding user which results in improved the user pruning accuracy in an efficient manner.

IV. EXPERIMENTAL SETTING

The Hilbert Space Clustered Indexing (HSCI) technique is implemented in Java Language using El Nino dataset from UCI machine learning repository. The El Nino dataset includes oceanographic and surface meteorological readings taken from a series of buoys located throughout the equatorial Pacific. The El Nino dataset comprise of following variables such as 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. Data is obtained from the buoys from as early as 1980 for some locations. Other data that was obtained from different locations are rainfall, solar radiation, current levels and subsurface temperatures.

The performance of HSCI technique is compared against with the existing two methods namely Predictive Subspace Clustering (PSC) [1] and Discriminative Embedded Clustering (DEC) [2] respectively. The experimental evaluation of HSCI technique is conducted on various factors such as clustering accuracy, clustering time, space complexity and user pruning efficiency.

V. RESULTS AND DISCUSSIONS

The efficiency of HSCI technique is compared against with exiting two methods namely Predictive Subspace Clustering (PSC) [1] and Discriminative Embedded Clustering (DEC) [2]. The performance of HSCI technique is evaluated along with the following metrics with the help of graphs.

A. Measurement of Clustering Accuracy

In HSCI technique, clustering accuracy is defined as the ratio of number of correctly clustered data objects to the total number of data objects taken. The clustering accuracy is measured in terms of percentage (%) and mathematically formulated as follows,

$$Clustering\ accuracy = \frac{\text{number of correctly clustered data objects}}{\text{total number of data objects taken}} * 100 \quad \text{---} \ (1)$$

From the equation (1), clustering accuracy of densely populated high dimensional data is obtained. While the clustering accuracy of densely populated high dimensional data is higher, the method is said to be more efficient.

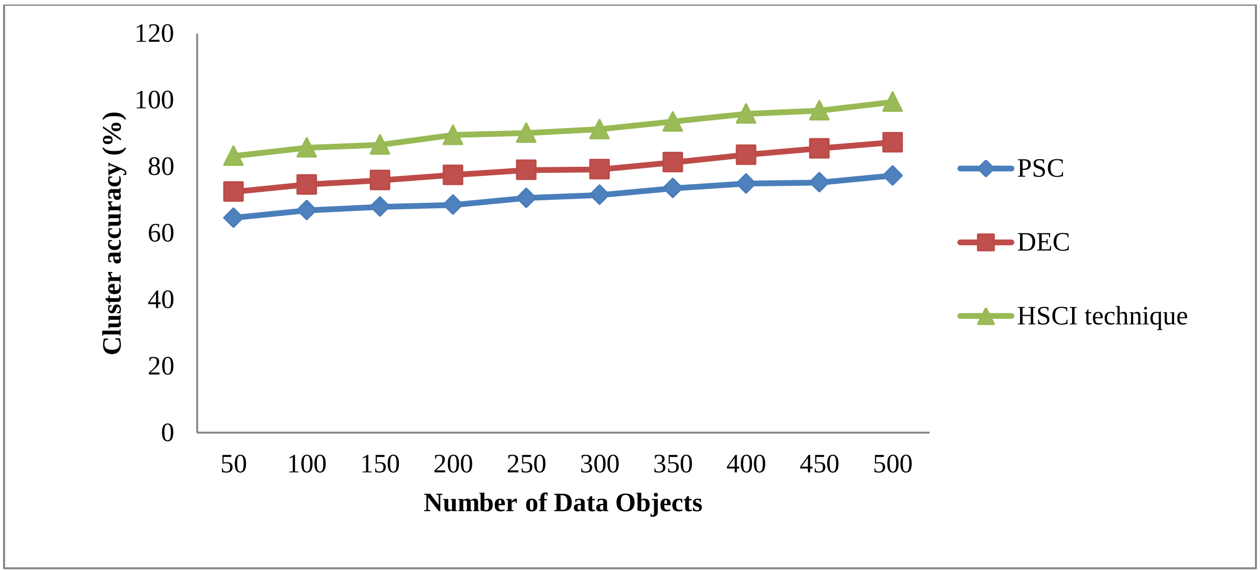


Figure 3 Measurement of Clustering Accuracy

Figure 3 portrays the impact of clustering accuracy of densely populated high dimensional data versus diverse number of data objects in the range of 50-100. As shown in figure, proposed HSCI technique provides higher clustering accuracy for clustering the densely populated high dimensional data when compared to PSC [1], DEC [2] respectively. In addition, while increasing the number of data objects for clustering process, the clustering accuracy is also gets increased using all the three methods. But comparatively, the clustering accuracy using proposed HSCI technique is higher. This is because of application of Hilbert space clustering algorithm in HSCI technique where it employs property of Hilbert curve to efficiently group the high dimensional data. Based on this property, if two block numbers are continuous, they must be group in the same cluster. This in turn helps to improve the clustering accuracy of densely populated high dimensional data in an effective manner. Therefore, proposed HSCI technique improves the clustering accuracy of densely populated high dimensional data by 22% as compared to PSC [1] and 13% as compared to DEC [2] respectively.

B. Measurement of Clustering Time

In HSCI technique, clustering time measures the amount of time taken for clustering the densely populated high dimensional data. The clustering accuracy is measured in terms of milliseconds (ms) and mathematically formulated as follows,

$$\text{clustering time} = n * \text{time for clustering one data object} \qquad (2)$$

From the equation (2), clustering time of densely populated high dimensional data is obtained where n represents the number of data objects taken for clustering process. While the clustering time of densely populated high dimensional data is lower, the method is said to be more efficient.

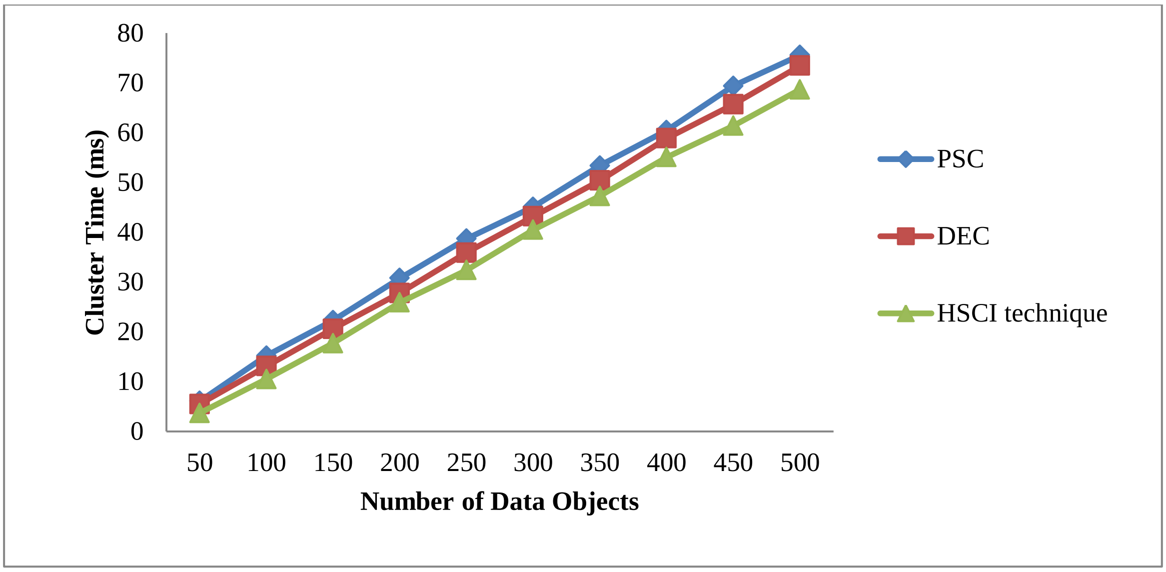


Figure 4 Measurement of Clustering Time

Figure 4 represents the impact of clustering time of densely populated high dimensional data versus varied number of data objects in the range of 50-100. As shown in figure, proposed HSCI technique provides minimum clustering time for clustering the densely populated high dimensional data when compared to PSC [1], DEC [2] respectively. As well, while increasing the number of data objects for clustering process, the clustering time is also gets increased using all the three methods. But comparatively, the clustering time using proposed HSCI technique is lower. This is due to application of Hilbert space clustering algorithm in HSCI technique. In HSCI Technique, Hilbert curve safeguard the distance of that data point which are close in space and designate similar data should be stored close together in the linear order. This type of mapping provides high speed for clustering densely populated high dimensional data. This in turn supports to reduce the clustering time in an efficient manner. As a result, proposed HSCI technique reduces the clustering time of densely populated high dimensional data by 23% as compared to PSC [1] and 14% as compared to DEC [2] respectively.

C. Measurement of Space Complexity

In HSCI technique, space complexity measures the amount of memory required for clustering the densely populated high dimensional data. The space complexity is measured in terms of Mega Bytes (MB) and mathematically formulated as follows,

$$\mathit{Space\ complexity} = n * \text{memory for stroing one clusered object} \qquad (3)$$

From the equation (3), space complexity for the clustering process is obtained where n represents the number of data objects taken for clustering process. While the space complexity, the method is said to be more efficient.

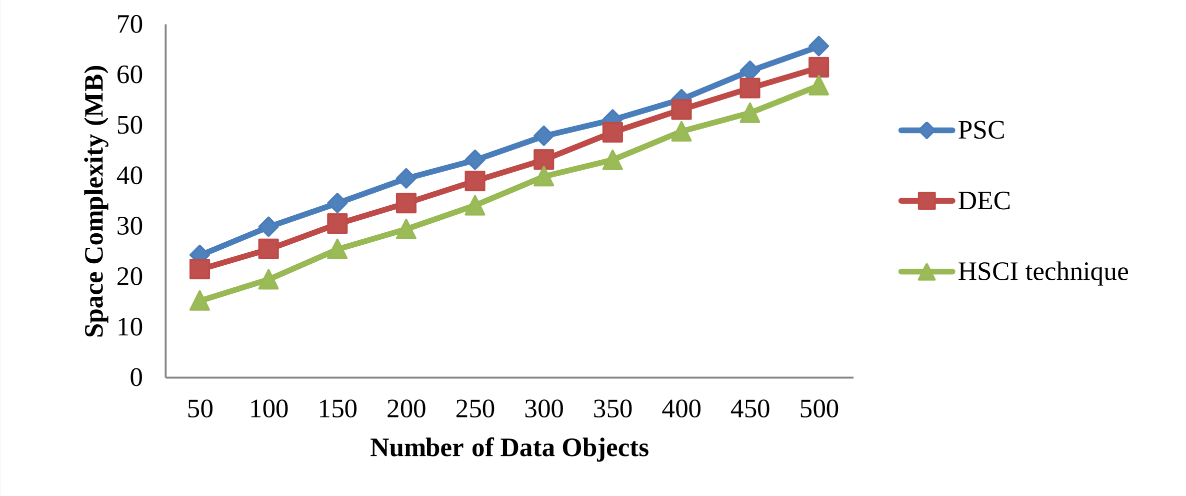


Figure 5 Measurement of Space Complexity

Figure 5 explains the impact of space complexity of densely populated high dimensional data versus diverse number of data objects in the range of 50-100. As shown in figure, proposed HSCI technique provides better space complexity for clustering the densely populated high dimensional data when compared to PSC [1], DEC [2] respectively. Besides, while increasing the number of data objects for clustering process, the space complexity is also gets increased using all the three methods. But comparatively, the space complexity using proposed HSCI technique is lower. This is due to application of R-tree indexing in HSCI technique where the Hilbert curve imposes a linear ordering on the data rectangles and then traverses the sorted list, allocating each set of C rectangles to a node in the R-tree. The final result is that the set of data rectangles on the similar node will be close to each other in the linear ordering and probably in the native space so the resulting R-tree nodes will have lesser areas. This in turn helps to minimize the space complexity of HSCI Technique in a significant manner. Therefore, proposed HSCI technique reduces the space complexity of densely populated high dimensional data by 29% as compared to PSC [1] and 17% as compared to DEC [2] respectively.

D. Measurement of User Pruning Efficiency

In HSCI technique, user pruning efficiency is defined as the ratio of number of data objects that are correctly extracted based on user query to the total number of data objects taken. The user pruning efficiency is measured in terms of percentages (%) and mathematically formulated as follows,

$$\text{user pruning efficiency} = \frac{\text{number of data objects that are correctly extracted based on user query}}{\text{total number of data objects taken}} * 100 \quad (4)$$

From the equation (4), the user pruning efficiency is obtained. While the user pruning efficiency, the method is said to be more efficient.

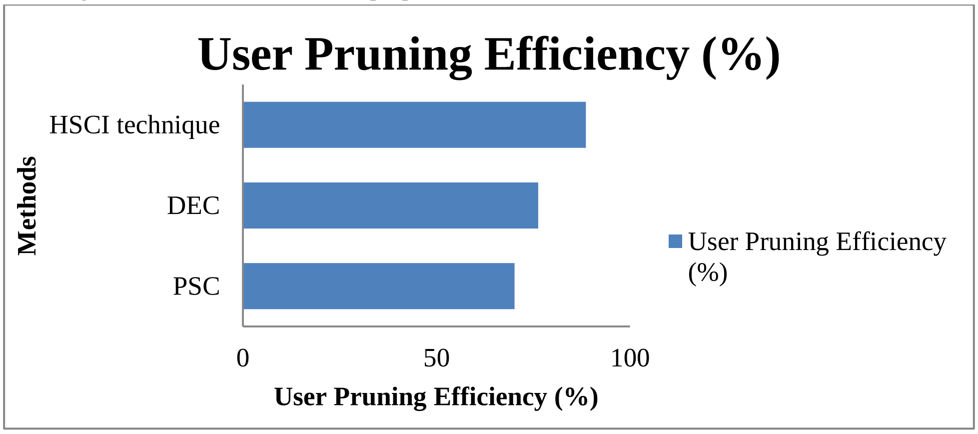


Figure 6 Measurement of User Pruning Efficiency

Figure 6 explains the impact of user pruning efficiency versus different number of data objects in the range of 50-100. As shown in figure, proposed HSCI technique provides better user pruning efficiency when compared to PSC [1], DEC [2] respectively. This is because of R-tree based pruning process is performed in HSCI technique where it cover the lower left, upper left, upper right, lower right places in the densely populated high dimensional data plane for efficient data retrieval. Besides, R-tree based pruning process efficiently removes unwanted data and extracts only the user requested data result from the R-tree. This in turn assists for improving the user pruning efficiency in an effective manner. Therefore, proposed HSCI technique improves the user pruning efficiency by 21% as compared to PSC [1] and 14% as compared to DEC [2] respectively.

VI. CONCLUSION

An effective novel framework called as Hilbert Space Clustered Indexing (HSCI) technique is designed for clustering the densely populated high dimensional data with higher clustering accuracy and minimum clustering time. At first, HSCI technique is used Hilbert Space Dimensional Clustering algorithm to cluster the densely populated high dimensional data which results in enhanced clustering accuracy with reduced clustering time. After performing the clustering process, HSCI technique builds an R-tree for indexing the clustered high dimensional data which results in reduced space complexity. Finally, the HSCI technique performs R-tree based pruning process to efficiently mine the user requested data from the indexed R-tree which in turn improved the user pruning efficiency. The performance of HSCI technique is measured in terms of clustering accuracy, clustering time, space complexity and user pruning efficiency by using El Nino weather data sets from UCI research repository and compared with two exiting methods. With the experiments conducted for HSCI technique, it is observed that the clustering accuracy of densely populated high dimensional data objects provides more accurate results while compared to state-of-the-art works. The experimental results demonstrate that HSCI technique provides better performance with an improvement of clustering accuracy by 18% and reduces the clustering time by 19% when compared to state-of-the-art works.

REFERENCES

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

[2] Chenping Hou, Feiping Nie, 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, Pages 1287 – 1299, June 2015

[3] Shulin Wang, Jinchao Gu, and Fang Chen, "Clustering High-Dimensional Data via Spectral Clustering Using Collaborative Representation Coefficients", Intelligent Computing Theories and Methodologies, Springer, Volume 9226 of the series Lecture Notes in Computer Science, Pages 248-258, 2015

[4] 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, Pages 1 – 14, January 2013

[5] Thenmozhi Srinivasan and Balasubramanie Palanisamy, "Scalable Clustering of High-Dimensional Data Technique Using SPCM with Ant Colony Optimization Intelligence", Hindawi Publishing Corporation, the Scientific World Journal, Volume 2015, Article ID 107650, 5 pages

[6] JingDong Tan and RuJingWang, "Smooth Splicing: A Robust SNN-Based Method for Clustering High-Dimensional Data", Hindawi Publishing Corporation, Mathematical Problems in Engineering, Volume 2013, Article ID 295067, 9 pages

[7] 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, 12 pages

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

[9] Rashmi Paithankar 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

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

[11] M Pallavi, "Efficient Density-Based Subspace Algorithms for High-Dimensional Data", International Journal of Engineering Development and Research, Volume 3, Issue 1, Pages 225-260

[12] Dharmveer Singh Rajput, Pramod Kumar Singh and Mahua Bhattacharya, "IQRAM: a high dimensional data clustering technique", International Journal of Knowledge Engineering and Data Mining, Volume 2, Issue 2/3, Pages 117-136, 2012

[13] k.Kaarguzhali, k.Sargunavathy, c.Jayanavithraa, "Hierarchical Accumulative Clustering Algorithm For High Dimensional Data", International Journal of Emerging Technology in Computer Science & Electronics (IJETCSE), Volume 12, Issue 4, Pages 201-203, 2015

[14] Yun Yang and Ke Chen, "Temporal Data Clustering via Weighted Clustering Ensemble with Different Representations", IEEE Transactions on Knowledge and Data Engineering, Volume 23, Issue 2, Pages 307 – 320, February 2011

[15] 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

[16] Zalhan Mohd Zin, Rubiyah Yusof, Ehsan Mesbahi, "Dimension Reduction and Clustering of High Dimensional Data using Auto-Associative Neural Networks", International Journal of Computer Applications (0975 – 8887), Volume 72, Issue 11, Pages 31-37, May 2013

[17] R. G. Mehta, N. J. Mistry, M. Raghuwanshi, "Towards Unsupervised and Consistent High Dimensional Data Clustering", International Journal of Computer Applications (0975 – 8887), Volume 87 – No.2, Pages 40-44, February 2014

[18] Zalhan Mohd Zin, Rubiyah Yusof, Ehsan Mesbahi, "Dimension Reduction and Clustering of High Dimensional Data using Auto-Associative Neural Networks", International Journal of Computer Applications (0975 – 8887), Volume 72, No.11, Pages 31-37, May 2013

[19] B.A Tidke, R.G Mehta, D.P Rana, "A Novel Approach for High Dimensional Data Clustering", International Journal of Engineering Science & Advanced Technology, Volume 2, Issue 3, Pages 645 – 651, 2012

[20] Asim Roy, "A classification algorithm for high-dimensional data", Procedia Computer Science, Elsevier, Volume 53, Pages 345–355, 2015.