International Journal of Advanced Research Methodology in Engineering & Technology,
Special Issue, March 2017, ISBN 978-1-63535-324-2

Hilbert Space Clustered Indexing Technique for Densely Populated High Dimensional Data Objects

R.Pushpalatha[1]

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, India.
(Email Id.: rpljour@gmail.com)

Dr.K.Meenakshi Sundaram[2]

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

***ABSTRACT* - Clustering is an essential and significant topic in both machine learning and data mining. 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. 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.**

Keywords – Clustering; Densely Populated High Dimensional Data Objects; Hilbert Space Clustered Indexing (HSCI) technique.

I.INTRODUCTION

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.

II. RELATED WORKS

In spatial data mining, clustering is one of the 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. 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.

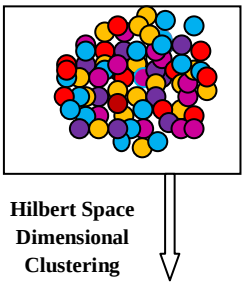
HSCI Technique initially takes the El Nino weather dataset as input which comprises collections of densely populated high dimensional data objects. After that, HSCI Technique is used Hilbert Space Dimensional clustering algorithm with aiming at improving the clustering accuracy of high dimensional data and reducing clustering time.

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. Let us assume El Nino weather dataset be composed of densely populated high dimensional data objects. The clustering of densely populated weather data using Hilbert Space Dimensional Clustering algorithm is shown in below Fig 1.

Densely populated weather data



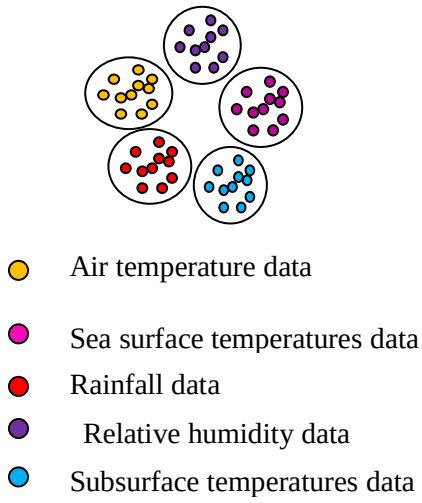
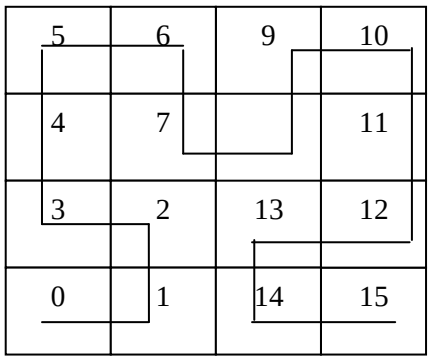


Fig. 1. Hilbert Space Dimensional Clustering Technique for Densely Populated Weather Data

Fig 1 shows the clustering of densely populated weather data. The Hilbert space dimensional clustering algorithm initially splits the densely populated high dimensional data objects into rectangle blocks. Here, the number of the data blocks is represented as m ($m = 2^{2h}$), where h indicates the order of Hilbert curve. The size of the blocks is dependent on the density of high dimensional data objects. The path of a space filling curve is a linear ordering which starts at one end of the curve and follows the path to the other end. The term h-ordering represents the ordering of Hilbert curve. The numbers in the rectangle block denotes the h-ordering and the gray block indicates the block which has data objects. The Hilbert curve order 2 of densely populated weather data is shown in below Fig. 2.



5	6(4)	9(3)	10
4	7(4)	8	11(7)
3(5)	2	13(7)	12(7)
0	1(5)	14	15

Fig. 2. Hilbert Curve of Order 2 for Densely Populated High Dimensional Data

As said by the property of Hilbert curve, two numbers are contiguous in the two dimensional space, they are always close together in one dimensional space. Depends on this property, if two block numbers are continuous, they must be clustered in the same cluster. Besides, Blocks 1 and 3 have data objects, but their Hilbert curve numbers are not continuous, hence they are clustered in different cluster number. In addition, Blocks 6 and 7 have data objects and also their Hilbert curve numbers are continues; so they are clustered in same cluster number. This is process continued until all the rectangle blocks of Hilbert curve is attained. $B\,[i] = 1$ represents block i has objects and $B[i] = 0$ denotes block i does not have data objects therefore it does not belong to any cluster. As shown in figure, The 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.

IV 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 tables and 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,

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

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.

TABLE 1 TABULATION FOR CLUSTERING ACCURACY

Number of Data Objects	Cluster accuracy (%)		
	PSC	***DEC***	***HSCI technique***
50	64.58	72.45	83.21
100	66.87	74.58	85.69
150	67.89	75.89	86.54
200	68.47	77.47	89.47
250	70.56	78.98	90.12
300	71.45	79.14	91.23
350	73.47	81.25	93.54
400	74.86	83.54	95.87
450	75.21	85.47	96.85
500	77.26	87.26	99.45

Table 1 illustrate the result analysis of clustering accuracy of densely populated high dimensional data using three methods based on the different number of data taken in the range of 50-100. From the table, while the 150 data objects is taken for clustering process, HSCI technique has acquires 86 % clustering accuracy whereas PSC [1], DEC [2] acquires 68 %, 76 % respectively. Therefore, clustering accuracy of densely populated high dimensional data using proposed HSCI technique is higher when compared to other existing methods [1], [2].

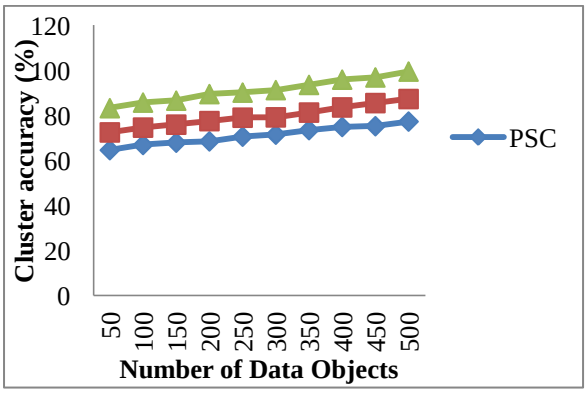


Fig. 3. Measurement of Clustering Accuracy

Fig.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.

V 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.

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.