Volume 3, Issue 10, October 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

# A Survey on Effect of Dimensionality Reduction Techniques on Data Clustering

**Nitin Soni, Mrs Abha Choubey**

*Computer Science & Engineering*

*Faculty Of Engg. & Tech., SSGI*, Bhilai, India

***Abstract— Clustering of data sets containing large number of variables presents challenges like cluster visualization and interpretation, computational complexity etc. Also when the scattering of data sets increase, distance measures used for assigning data sets to different clusters becomes less meaningful, leading to inaccurate clustering. So theoretically, it makes sense to reduce dimensionality of any large dimensionality data set before applying clustering techniques to it. Another area of challenge in data clustering is validation of clustering results for which a number of validity indices have been proposed. This paper does a survey on application of different dimensionality reduction techniques such as PCA, CMDS, ISOMAP, HLLE etc as a data pre-processing step on k-means data clustering algorithm and validates the results by observing the change in different Clustering Validity Indices.***

***Keywords — Data Clustering, Dimensionality Reduction Techniques, PCA, CMDS, ISOMAP, HLLE, Clustering Validity Indices.***

## I. INTRODUCTION

**Data Clustering** is a method of segmenting data sets into partitions called clusters in such a manner that data sets within a partition are more similar to each other than they are to data sets belonging to a different cluster. In data clustering process, no prior information is available about the number of partitions or groups (also called as classes), group membership criteria or group labels, or for that matter whether groups are present in the data sets or not. Hence Data Clustering is called **Unsupervised Learning**. Data clustering is one of the important techniques of pattern recognition. Recent development of explosion of data and data mining has increased manifold the applications and importance of data clustering.

Different type of data clustering algorithms include **Hierarchical based, Graph based, Density based, Centre based** etc. One of the most popular techniques of clustering **is K-means Algorithm** (Centre Based). K-means algorithm partitions a particular data set into k subsets. Each partition or cluster has a centre also called as centroid, which is actually the mean of all the data points in that particular partition. At the starting of computation, the algorithm randomly selects some data points as centroids (also called as seed centroids) and assigns to each centroid the data points closest to it. The centroid and the data points assigned to it represent a partition. Once all the data points of the data set to be clustered are assigned, the median or centroid for each cluster is re-calculated using the data points of a particular cluster. The process is repeated until a stopping criteria is met [1].

**Dimensionality Reduction Techniques** refer to methods for reducing the number of attributes or variables of the data items. One of the several ways to achieve dimensionality reduction is by creating new variables in a such a manner that these new variables can be expressed as **functions** of old variables[2]. Depending upon the linearity or non-linearity of these functions, the dimensionality reduction techniques can be classified into **linear** and **non-linear** respectively. Some of the linear techniques include **Principal Component Analysis (PCA)** etc. Some important non-linear techniques include **Classical Multi Dimensional Scaling (CMDS)**, **Isometric Feature Mapping (ISOMAP)**, **Hessian Eigen maps (HLLE)** etc [2]. This paper applies with all these techniques as a data pre-processing step to K-means Algorithm.

Clustering organizes a particular data set into groups, when no predefined groups or partitions are present in the data set to show that clustering is accurate or valid[3]. Many clustering validity indices have been proposed to overcome this problem. These clustering indices have broadly classified into internal, external and stability indices. Internal criteria are based on the constituent data points of the different clusters and can be applied to hierarchical and partition based clustering algorithms. This paper uses **Davies-Bouldin Index**[4], **Calinski-Harbasaz Index**[5], **Dunn's Index**[6] and **Silhouette Index**[7].

Section I of this paper deals with the introduction of concepts used in this paper. Section II deals with Literature Review, Section III deals with Problem Identification, Section IV deals with Methodology, Section V with Datasets used for experiments, Section VI with Experiments and Results, Section VII Conclusion followed by Acknowledgement.

## II. LITERATURE REVIEW

Many attempts have been made in past to combine Data Clustering and Dimensionality reduction techniques to improve the efficiency of Clustering algorithms. Chris Ding et al., in 2004 experimentally showed by conducting experiments that when dimensions of data set are reduced, data clustering results improve significantly [8]. Seong S. Chaea et al in 2006 showed that retrieval ability of clustering algorithms improved significantly by use Principal Coordinate Analysis (also known as classical multidimensional scaling) than Principal Component Analysis for data pre-processing [9]. Hai-Dong Meng et al., in 2010 showed that dimensionality reduction algorithms improve accuracy of K-means and Hierarchical clustering algorithms only when the number of attributes of the data set are less than 30 [10]. Rajashree Dash et al., in 2010 derived initial centroids for application of K-means algorithm from the reduced data set obtained by Principal Component Analysis [11]. More recently S. M. Shaharudin et al., in August 2013 showed that effective of Principal Component Analysis as a data pre-processing improves significantly if Tukey's biweight correlation matrix is used instead of Pearson correlation matrix in calculating principal components [12] . One of the most recent works in the field of cluster validation has been by Olatz Arbelaitz et al., in 2012, where comparison of 30 different clustering validity indices has been done [13].

## III. PROBLEM IDENTIFICATION

Among different issues in data clustering, high dimensionality remains a big problem. An increase in number of attributes of the data set increases the scattering of data, reducing the accuracy of proximity or dissimilarity measures used for finding clusters. So as the number of the data set increases, its clustering becomes increasingly inaccurate. Even the application of Dimensionality reduction techniques does not alleviate the problem [10]. Another issue is choice of appropriate clustering validity index to validate the results. This is major issue because different clustering validity indices give different and many a times contrasting results.

## IV. METHODOLOGY

The below flow diagram summarizes the procedure followed.

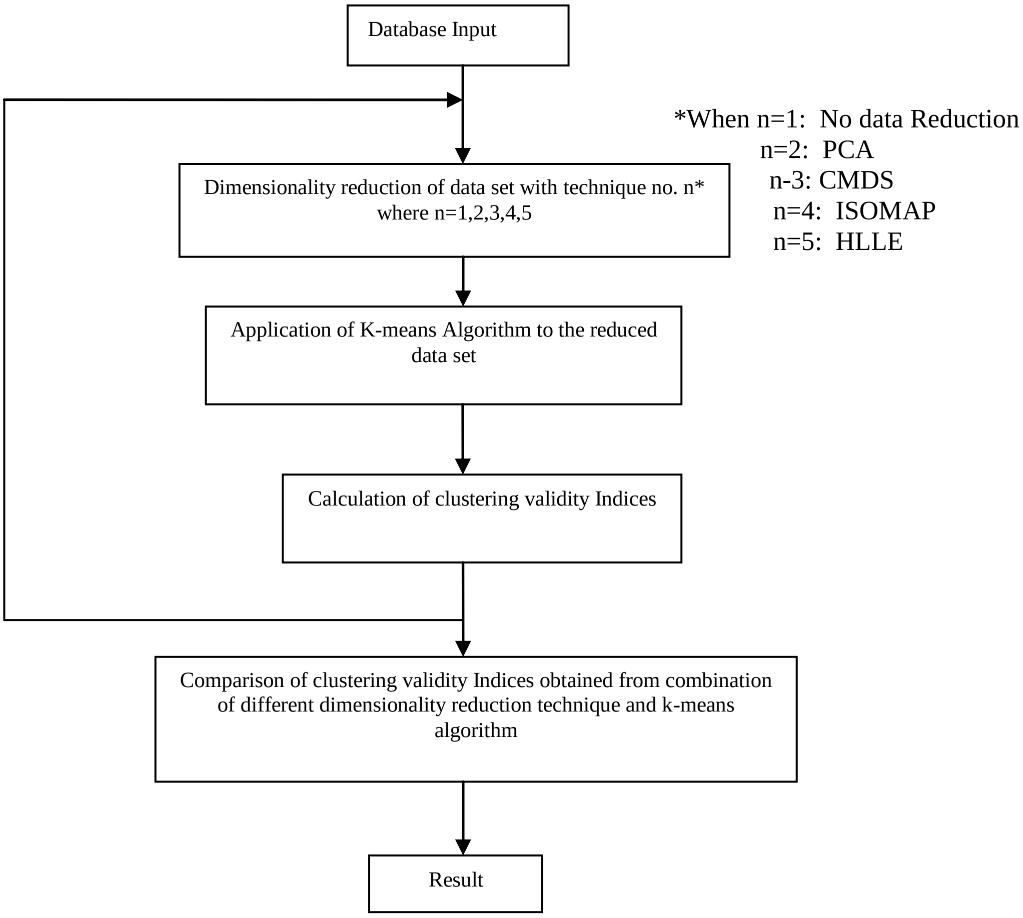

Database Input
Dimensionality reduction of data set with technique no. n* where n=1,2,3,4,5
*When n=1: No data Reduction
n=2: PCA
n-3: CMDS
n=4: ISOMAP
n=5: HLLE
Application of K-means Algorithm to the reduced data set
Calculation of clustering validity Indices
Comparison of clustering validity Indices obtained from combination of different dimensionality reduction technique and k-means algorithm
Result

FIGURE NO.1 METHODOLOGY

## IV. DATASET FOR EXPERIMENTS

The dataset used for experiment is **Libras Movement Data Set from UCI Machine Learning Repository** [14]. This data set has has 360 tupples, 90 attributes and 15 clusters.

## V. EXPERIMENTS AND RESULTS

The impact of dimensionality reduction techniques on k-means clustering algorithm is shown by studying the changes on different clustering validity indices. In each case the database taken is Libras-Movement database. To conduct experiments, packages from **MATLAB** and **R** software are used.

## VI A. **CHANGES IN SILHOUETTE INDEX**

The value of silhouette ranges from **-1 to +1**. A value of **+1** implies that any data tupple assigned to any particular cluster is **similar** to other tupples in that particular cluster and a value of -1 indicates **dissimilarity** [7]. In figures 2 to 6 shown below depict values of silhouette indices when different dimensionality reduction techniques are used and also when no dimensionality reduction is done. In each figure horizontal axis depicts number of partitions and vertical axis depicts Silhouette index.

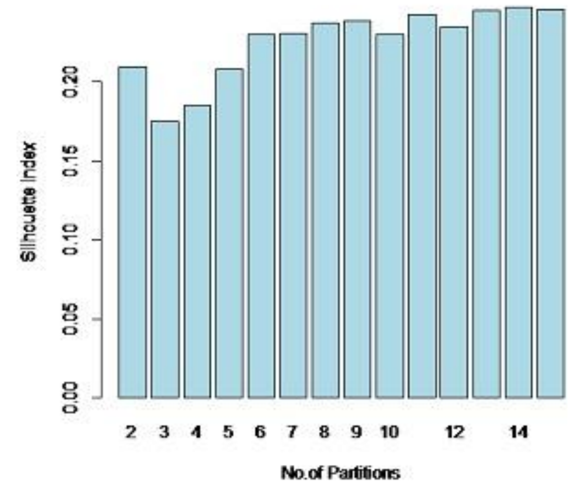


Figure No.2 Silhouette Plot With No Dimensionality Reduction

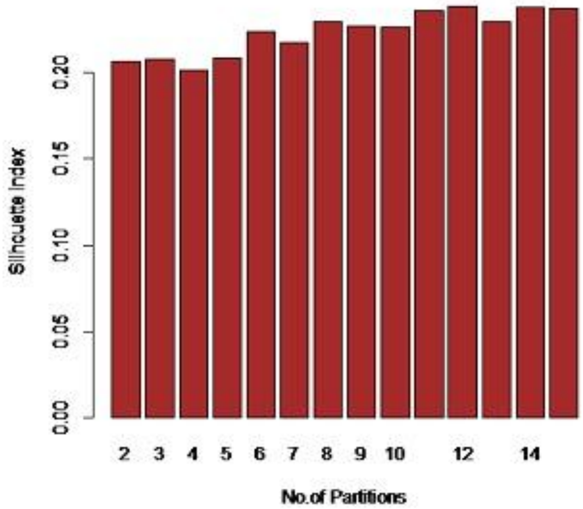


Figure No.3 Silhouette Plot With Pca As Dimensionality Reduction Technique

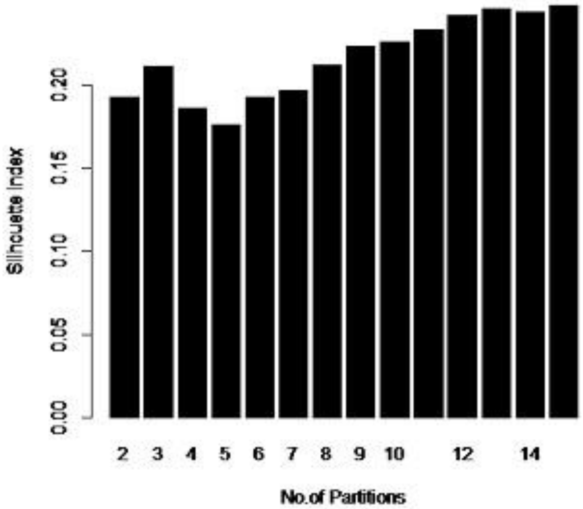


Figure No.4 Silhouette Plot With Classical Multi Dimensional Scaling As Dimensionality Reduction

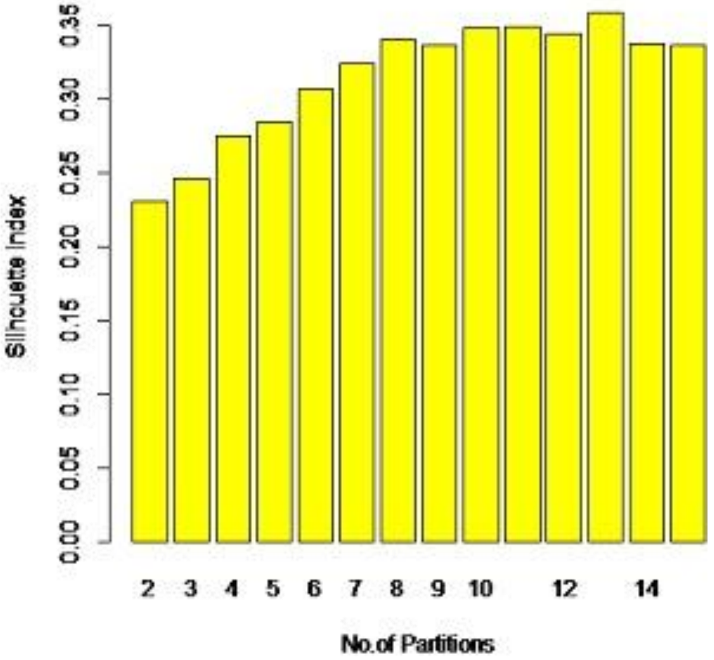


Figure No.5 Silhouette Plot With Isomap As Dimensionality Reduction Technque

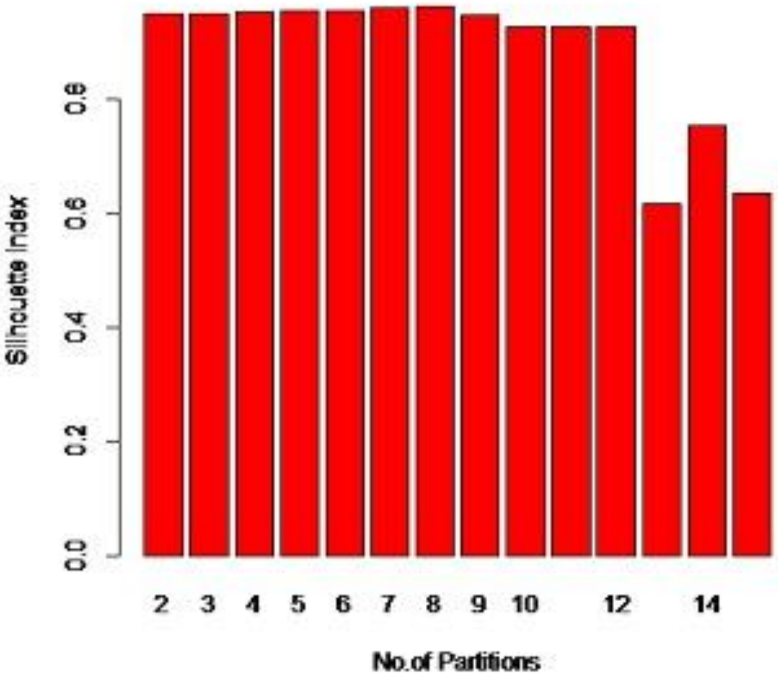


Figure No.6 Silhouette Plot With Hlle As Dimensionality Reduction Technique

It is evident from above bar graphs, the silhouette index range when HLLE is applied as a dimensionality reduction technique is **0.6 to 0.9**. For all other dimensionality reduction techniques and when no dimensionality reduction technique is applied the Silhouette index ranges from **0.2 to 0.3**. So the bar graphs show dramatic improvement in silhouette index values when HLLE is applied as the dimensionality reduction technique as compared to other techniques or when no technique is applied, thereby showing improvement in quality of clustering. But the drawback of HLLE is that is unable to predict the accurate number of clusters. When the number of partitions is 15(the accurate number of clusters in Libras-Movement Database), the value of silhouette index is quite low (**0.6393**) as compared to its value for inaccurate number of partitions (**0.9 and above**).

### VI B. **CHANGES IN DUNN INDEX**

Next we consider changes in Dunn index. A higher value of Dunn index indicates better clustering results [5]. Figures 7 to 11 depict Dunn indices for different approaches. Horizontal axis shows different number of partitions and vertical axis depicts values of Dunn Index

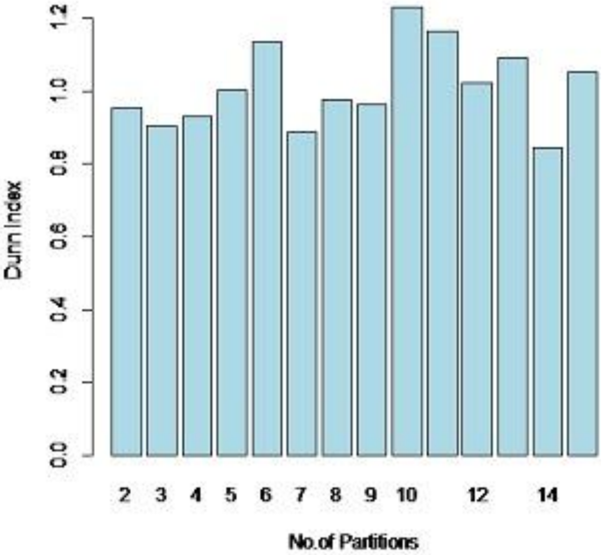


Figure No.7 Dunn Index With No Dimensionality Reduction

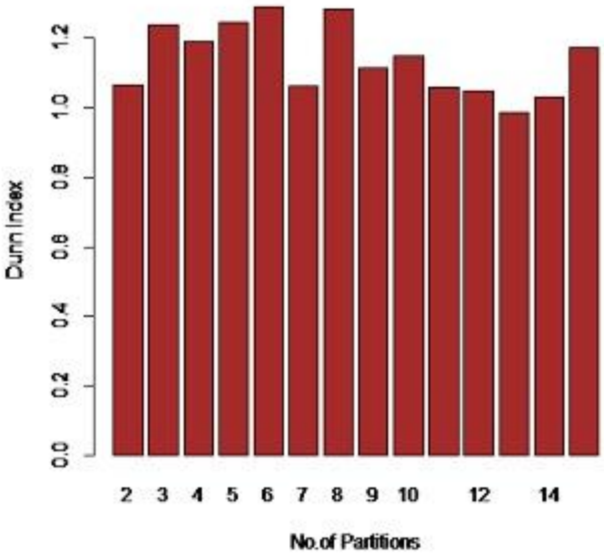


Figure No. 8 Dunn Index With Pca As Dimensionality Reduction Technique

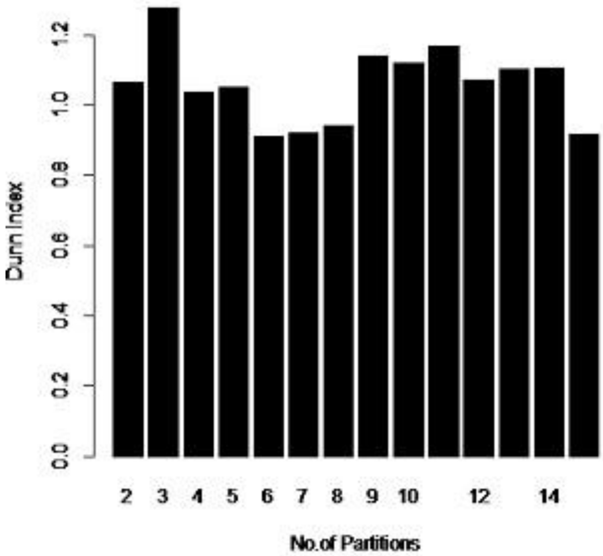


Figure No.9 Dunn Index With Cmds As Dimensionality Reduction Technique

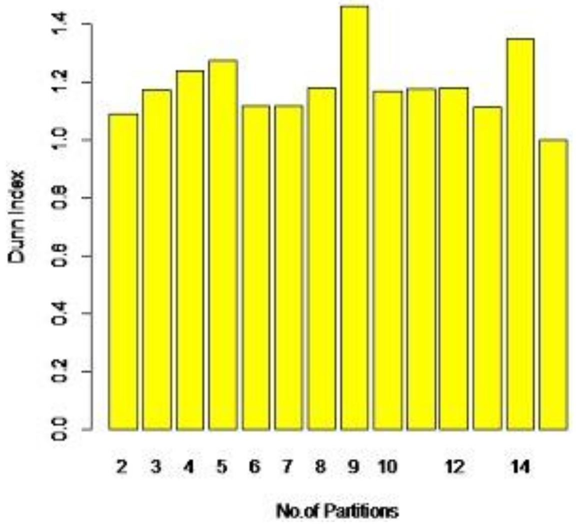


Figure No.10 Dunn Index With Isomap As Dimensionality Reduction Technique

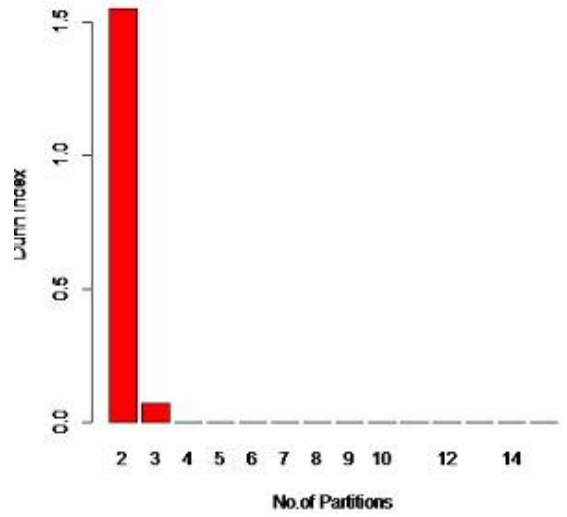
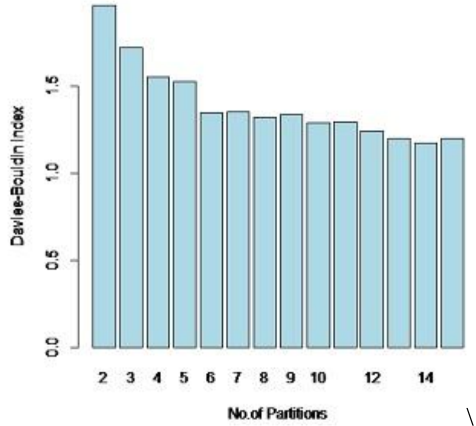


Figure No.11 Dunn Index With Hlle As Dimensionality Reduction Technique

It is evident from figure 6 to figure 11that Dunn Index values vary in the same narrow range from **0.9 to 1.3** for different dimensionality reduction technique (except for HLLE where it varies from 0 to 1.5). So nothing can be said conclusively from these values.

### VI C. **CHANGES IN DAVIES–BOULDIN INDEX**

Next index considered is Davies-Bouldin index. The smaller the value of this index, the better the clustering results [4]. Figures 12- 16 depict show variation in values of Davies-Bouldin index for different values of partitions.



\

Figure No. 12. Davies-Bouldin Index With No Dimensionality Reduction

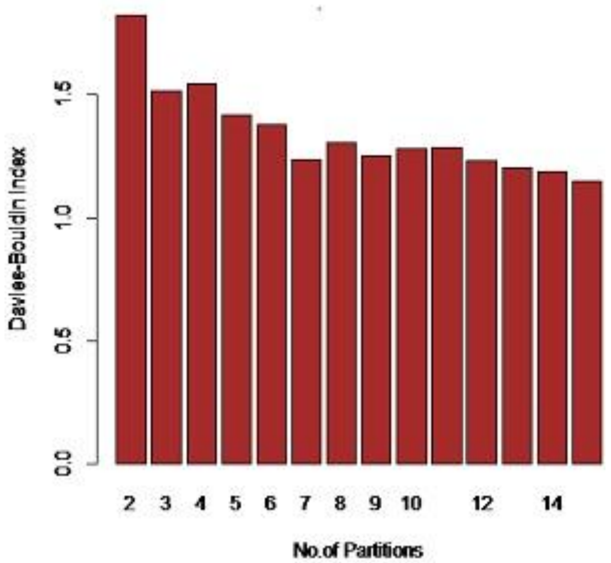


Figure No.13. Davies-Bouldin Index With Pca As Dimensionality Reduction Technique

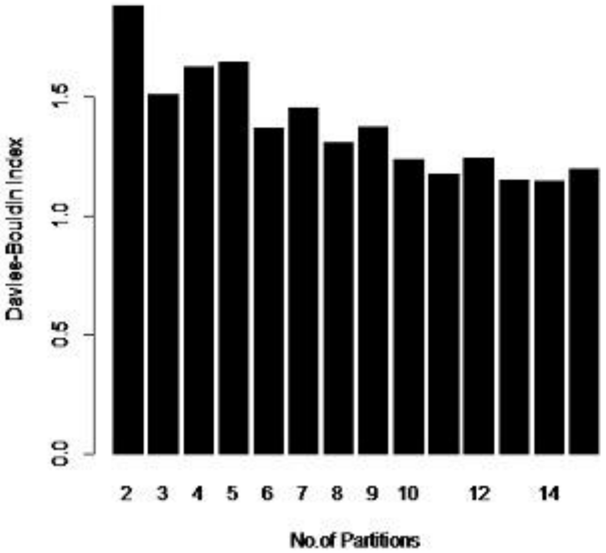


Figure No. 14. Davies-Bouldin Index With Cmds As Dimensionality Reduction Technique

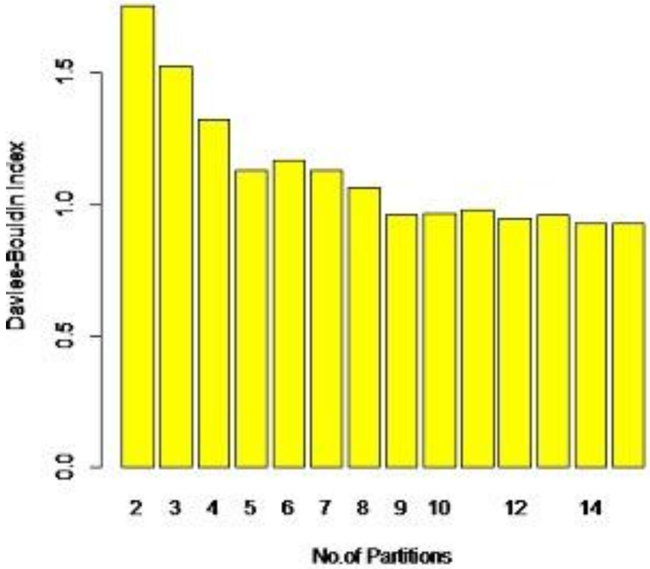


Figure No. 15. Davies-Bouldin Index With Isomap As Dimensionality Reduction Technique

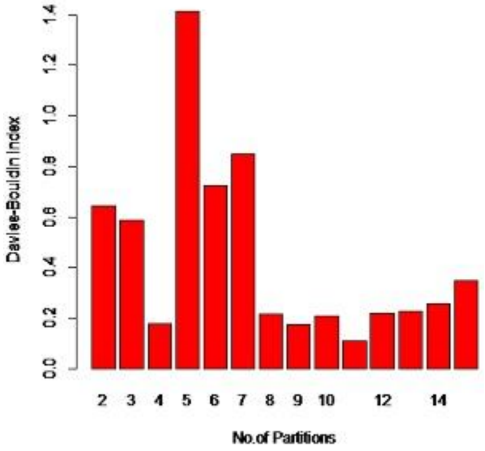


Figure No. 16. Davies-Bouldin Index With Hlle As Dimensionality Reduction Technique

The figures 12 to 16 show that values of Davies-Bouldin Index approaches a **lower value** as the number of clusters approach their correct value. Also the value of this index is less when HLLE as a dimensionality reduction technique as compared to other techniques (especially for accurate value of partitions). This indicates that HLLE is a better dimensionality reduction technique than others. Still, there is not much difference in the range of values for HLLE (**0.1 to 0.9**).

VI D. **CHANGES IN CALINSKI-HARABASZ INDEX**

Next we consider changes in Calinski-Harabasz Index. A higher value of Calinski-Harbasaz Index indicates better clustering results [6]. Figures 17 to 21 depict the results.

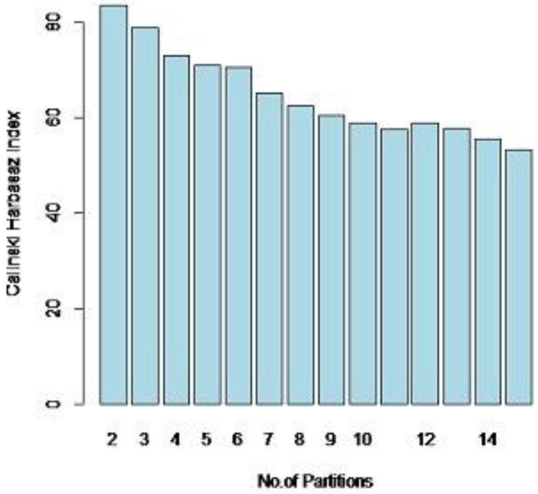


Figure No.17. Calinski Harbasaz Index With No Dimensionality Reduction

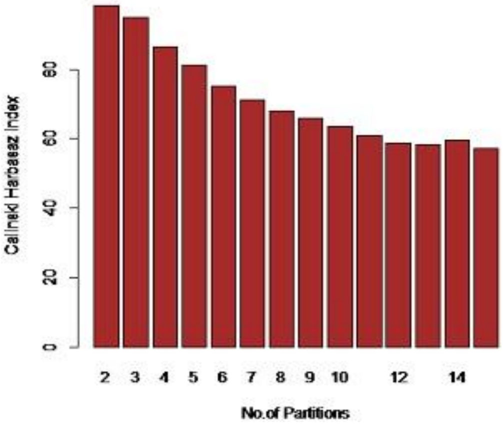


Figure No.18. Calinski Harbasaz Index With Pca As Dimensionality Reduction Technique

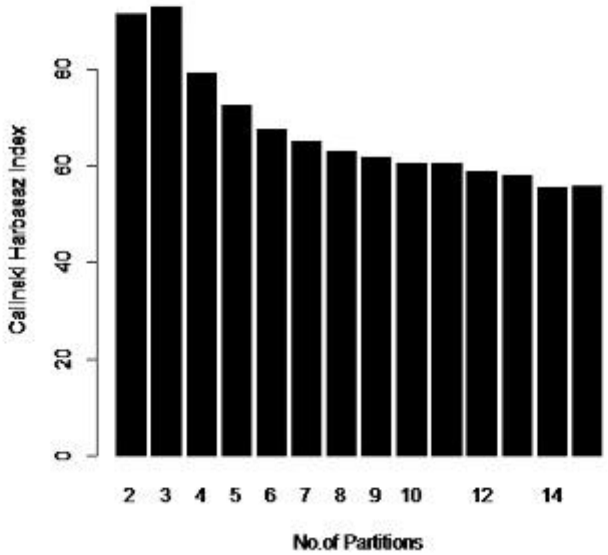


Figure No.19. Calinski Harbasaz Index With Cmds As Dimensionality Reduction Technique

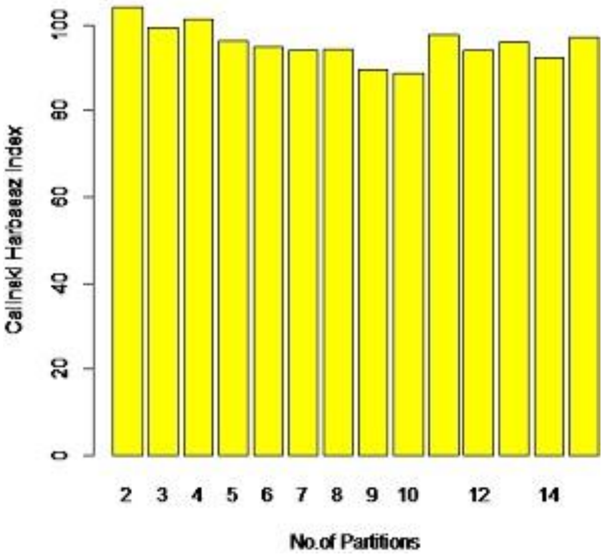


Figure No.20. Calinski Harbasaz Index With Isomap As Dimensionality Reduction Technique

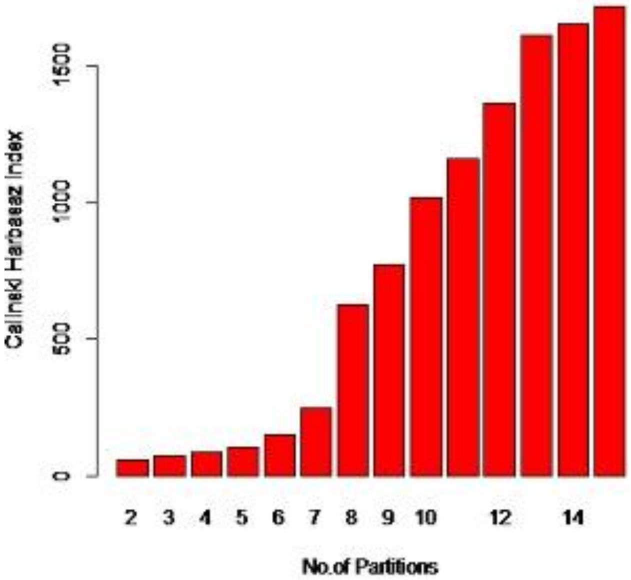


Figure No.21. Calinski Harbasaz Index With Hlle As Dimensionality Reduction Technique

Figures 17 to 20 are inconclusive. The range of values for Calinski-Harbasaz (CH) Index remains the same from 60 to 100. In some cases the Index values decrease for accurate number of clusters instead of increasing. But figure number 21 stands out from the rest. First of all the range of values are from 60 to 1715, 1715 being the value for number of partitions equal to 15 and 60 for number of partitions equal to 2. **So the CH Index is accurately predicting the number of partitions in the Libras Movement database**. Also the jump in the range of values ( 60-100 to 60-1715) indicates a corresponding jump in the quality of clustering. This indicates that for Libras data base with 90 attributes and 15 classes, **application of HLLE improves clustering**.

.

## VII. CONCLUSIONS

In this paper different techniques of dimensionality reduction are applied on Libras Movement database. To detect their effectiveness, four clustering validity indices are used. From the results obtained it can be concluded that **HLLE** is a better technique than Dunn Index, Davies-Bouldin Index and Silhouette plot for improvement of clustering results. Also **Calinski-Harbasaz** Index outperforms the other indices mentioned earlier in depicting the improvement in data clustering.

## ACKNOWLEDGMENT

I would like to give special thanks to Mrs Samta Gajbhiye, Head of Department, Computer Science and Engineering, Faculty of Engineering and Technology, Shri Shankaracharya Group of Institutions(SSGI), Bhilai and Dr P.B.Deshmukh, Director, SSGI for their continuous motivation and support.

## REFERENCES

[1] Bing Liu, *Web Data Mining : Exploring Hyperlinks, Contents and Usage Data Second Edition,* Springer-Verlag Berlin Heildberg 2007,2011.

[2] Wendy L. Martinez, Angel R. Martinez, Jeffery L. Solka, *Exploratory Data Analysis with MATLAB Second Edition,* CRC Press, Taylor & Francis Group, 6000 Broken Sound Parkway NW, Suite 300, Boca Raton, FL 33487-2742, 2011.

[3] Goujon Gan, Chaoqun Ma, and Jianhong Wu, *Data Clustering Theory, Algorithms, and Applications,* ASA-SIAM Series on Statistics and Applied Probability, SIAM, Philadelphia, ASA, Alexandria, VA, 2007.

[4] D. L. Davies and D. W. Bouldin, *A cluster separation measure*, IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1, no. 2:224-227, 1979.

[5] J. Dunn, *Well separated clusters and optimal fuzzy partitions*, Journal of Cybernetics, 4:95-104, 1974.

[6] T. Calinski and J. Harabasz, *A dendrite method for cluster analysis,* Communications in Statistics, 3, no. 1:1-27, 1974

[7] Rousseeuw P.J, *Silhouettes: a graphical aid to the interpretation and validation of cluster analysis*, Journal of Computational and Applied Mathematics , 20:53-65, 1987

[8] Chris Ding and Xiaofeng He, *Principal Component Analysis and k-means Clustering,* SIAM, Philadelphia, ASA, Alexandria, VA, 2004

[9] S.S. Chae, W.D. Warde, *Effect of using principal coordinates and principal components on retrieval of clusters,* Computational Statistics & Data Analysis 50 (2006) 1407 – 1417.

[10] Hai-Dong Meng, Jin-Hui Ma, Guan-Dong Xu, *Experimental Research on Impacts of Dimensionality on Clustering Algorithms,* 978-1-4244-5392-4/10 , IEEE 2010.

[11] Rajashree Dash, Debahuti Mishra, Amiya Kumar Rath, Milu Acharya, *A hybridized K-means clustering approach for high dimensional dataset,* International Journal of Engineering, Science and Technology,Vol. 2, No. 2, 2010, pp. 59-66

[12] S. M. Shaharudin, N. Ahmad, F. Yusof, *Improved Cluster Partition in Principal Component Analysis Guided Clustering,* International Journal of Computer Applications (0975 – 8887), Volume 75– No.11, August 2013.

[13] Olatz Arbelaitz ,IbaiGurrutxagan, Javier Muguerza, Jesu´s M.Pe´ rez, In˜igo Perona, *An extensive comparative study of cluster validity indices,* Pattern Recognition, Elsevier 2012.

[14] K. Bache and M.Lichman, (2013), UCI Machine Learning Repository, [http://archive.ics.uci.edu/ml]: Irvine, CA University of California, School of Information and Computer Science.