Volume 4, Issue 5, May 2014 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

An Improved K-Medoid Clustering Algorithm Using Feature Reduction Techniques and Clustering Validation Indices

Nitin Soni
Computer Science & Engineering
Faculty Of Engg. & Tech., SSGI, Bhilai

Prof Abha Choubey
Computer Science & Engineering
Faculty Of Engg. & Tech., SSGI, Bhilai

Abstract— Data organization techniques such as clustering are influenced by number of dimensions or features. Cluster visualization and interpretation, computational complexity, accurate cluster indexing of data sets directly depend on number of dimensions. Therefore feature extraction techniques can improve data clustering. Another problem in data clustering is validation of clustering results as clustering is unsupervised learning technique. This paper formulates a new model for data clustering using combination of feature extraction, data clustering algorithm and clustering validity index/indices. The different features reduction techniques used are PCA, CMDS, ISOMAP and HLLE. The clustering validity indices used are Silhouette Index, Dunn Index, Davies Bouldin Index and Calinski Harbasaz Index.

Keywords — Data Clustering, Dimensionality Reduction Techniques, PCA, CMDS, ISOMAP, HLLE, Silhouette Index, Dunn Index, Calinski Harbasaz Index.

I. INTRODUCTION

Clustering is a technique of organizing data instances into similar groups called clusters in such a manner that the instances within a cluster are more similar to each other than they are to instances belonging to a different cluster. Clustering technique is used to group data into clusters when no attributes of these clusters are known. So clustering is an example of **Unsupervised Learning**. Data clustering is one of most widely used techniques with applications in wide variety of fields such as archaeology, medicine, web mining etc[1].

Some of the clustering algorithms include **Hierarchical based algorithms, Partition based algorithms, Fuzzy based algorithms** etc. An important partition based techniques is **K-medoid algorithm**. K-medoid algorithm partitions a particular data set consisting of many instances into k subsets or clusters. The term medoid in the K-medoid algorithm is actually a data instance within in a cluster such that its average dissimilarity from all other data instances in the cluster is minimal. The algorithm starts working by random selection of data instances as medoids and each data point is associated with medoid closest to it. The medoid and the data points assigned to it represent a partition. After all the data points are associated to a cluster, for every cluster there is a swapping of any data point of that cluster with the initial randomly assigned medoid of that cluster. The total distances from this new medoid to all the points in the cluster are again calculated. The process is repeated for each data instance in cluster and all the clusters. The partitions containing medoids with lowest sum of all the distances (also sometimes termed as cost of clustering) is chosen as the final clustering in the end [1].

Feature Reduction Techniques refer to methods for reducing the number of attributes or variables of the data items. These techniques such as **PCA** might employ orthogonal linear transformations or by projections on a straight line or on a plane.Such techniques are called **Linear Dimensionality Reduction techniques. Non-Linear Dimensionality Reduction** techniques such as **ISOMAP** embed data in a non-linear manifold [2]. The techniques used by this paper for feature reduction are PCA, CMDS, ISOMAP, HLLE.

Since clustering is a unsupervised learning process, clustering validity indices have been proposed to validate the clustering algorithms[3]. These indices are of three types**:- internal, external and stability indices.** The indices based on constituent data instances within a cluster are called Internal Indices 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 on improved clustering algorithm Section VIII Conclusion followed by Acknowledgement.

II. LITERATURE REVIEW

Optimization of data clustering algorithms have been attempted in past by use of dimensionality reduction techniques. In 2004 Chris Ding et al., proved by conducting experiments that dimensionality reduction techniques improved results of data clustering algorithms [8]. In 2006 it was shown by Seong S. Chaea et al that Principal Coordinate Analysis (also known as classical multidimensional scaling) rather than Principal Component Analysis significantly improved data

clustering results.[9]. Hai-Dong Meng et al., in 2010 empirically proved accuracy of K-means and Hierarchical clustering algorithms improved by use of dimensionality reduction techniques only when the number of attributes of the data set are less than 30 [10]. Other improvements include derivation of initial centroids from reduced data set obtained by PCA for K-means algorithm by Rajashree Dash et al., in 2010 [11]. In August 2013,S. M. Shaharudin et al., showed that effectiveness of PCA 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] .

Olatz Arbelaitz et al., in 2012, have done a quite comprehensive work in the field of clustering validation. In their work comparison of 30 different clustering validity indices has been done [13].

III. PROBLEM IDENTIFICATION

Large number of attributes causes machine learning algorithms such as data clustering to underperform. The main reason is that as the number of variables increase, dissimilarity or distance measures used for creating clusters become meaningless.The situation increasingly worsens as the number of variables increase and even the application of Dimensionality reduction techniques such as PCA does not improve clustering [10].

A large number of clustering validity indices (over 100 internal indices alone) compound the problem of validating clustering. As a matter of fact, different clustering validity indices give contrasting result.

IV. METHODOLOGY

The below flow diagram summarizes the procedure followed.

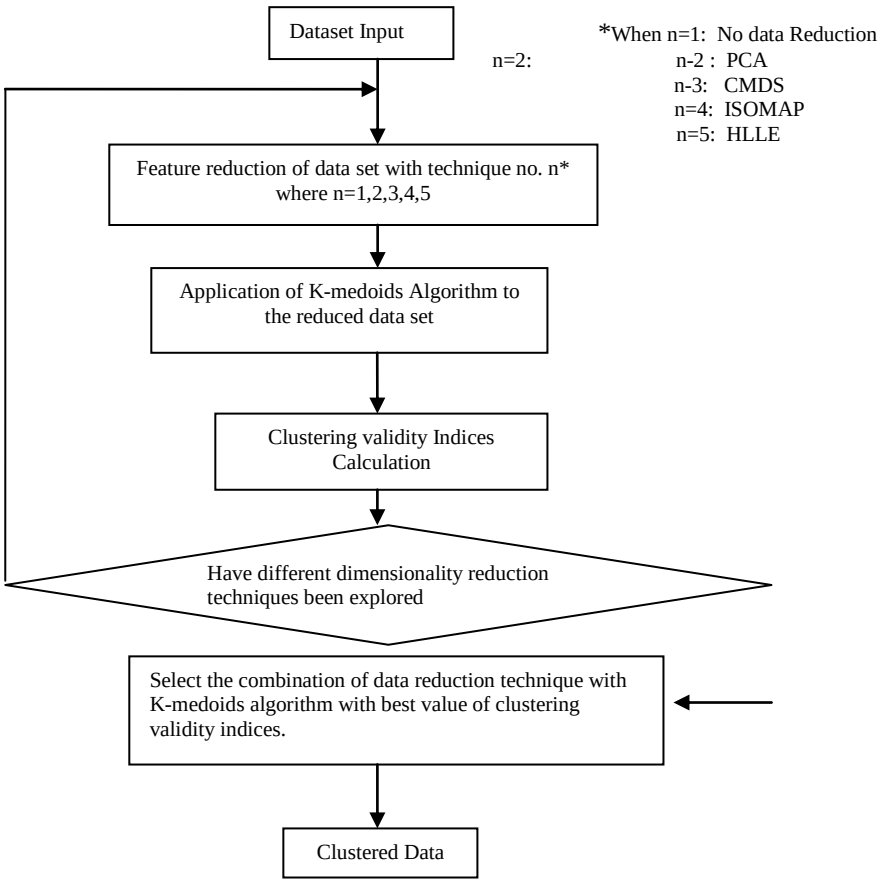


FIGURE NO.1 METHODOLOGY

V. 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.

VI. EXPERIMENTS AND RESULTS

The impact of dimensionality reduction techniques on k-medoids clustering algorithm is shown by studying variation on clustering validity index with respect to number of partitions k. If a clustering validity index for a particular dimensionality reduction techniques accurately predicts the number of partition, that particular feature reduction method and clustering validity index is considered to be effective. 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 index ranges from **-1 to +1**. A value of **+1** implies that any data instance assigned to any particular cluster is **similar** to other instances 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.

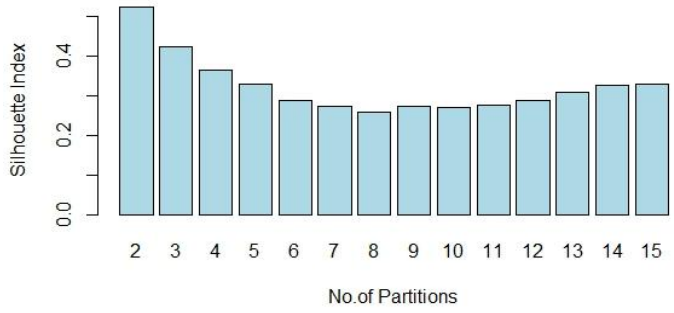


FIGURE NO.2 CHANGES IN SILHOUETTE INDEX WHEN NO DIMENSIONALITY REDUCTION TECHNIQUE IS EMPLOYED

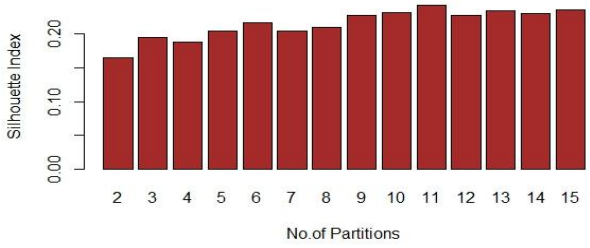


FIGURE NO.3 CHANGES IN SILHOUETTE INDEX WHEN PRINCIPAL COMPONENT ANALYSIS IS EMPLOYED

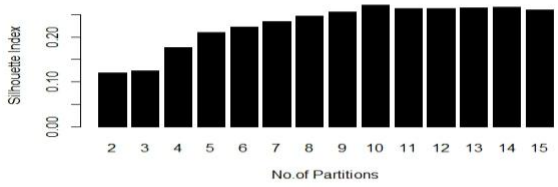


FIGURE NO.4 CHANGES IN SILHOUETTE INDEX WHEN CLASSICAL MULTI DIMENSIONAL SCALING IS EMPLOYED

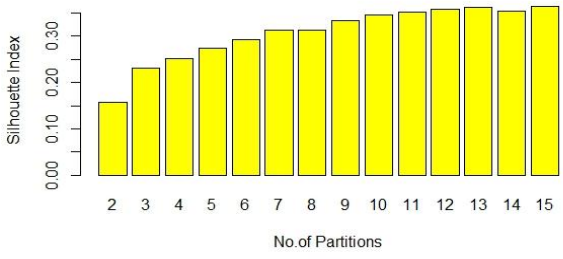


FIGURE NO.5 CHANGES IN SILHOUETTE INDEX WHEN ISOMAP IS EMPLOYED

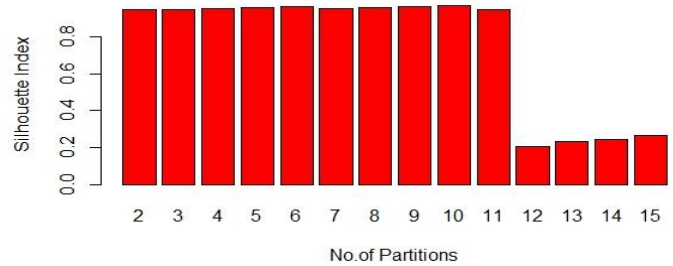


FIGURE NO.6 CHANGES IN SILHOUETTE INDEX WHEN HLLE IS EMPLOYED

It is evident from above bar graphs, the silhouette index range when HLLE is applied as a dimensionality reduction technique is **0.2 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 (**0.3**) 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

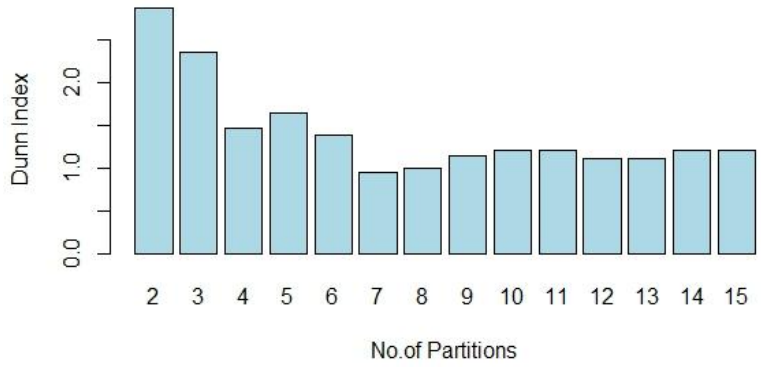


FIGURE NO.7 CHANGES IN DUNN INDEX WITH NO DIMENSIONALITY REDUCTION

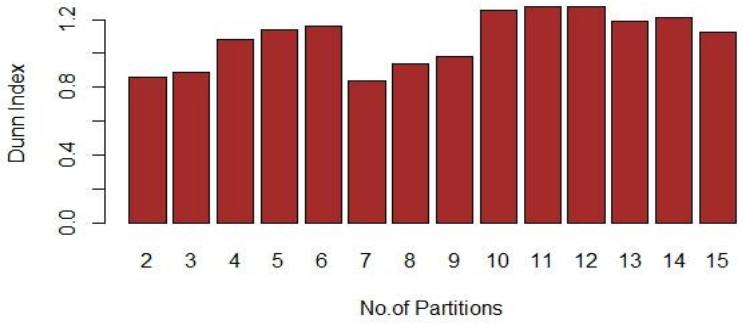


FIGURE NO. 8 CHANGES IN DUNN INDEX WHEN PRINCIPAL COMPONENT ANALYSIS IS USED

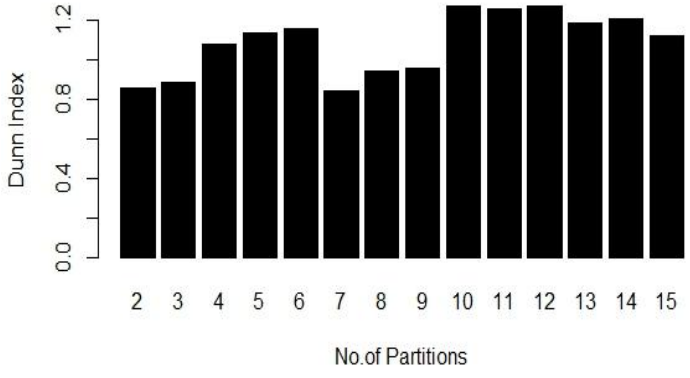


FIGURE NO.9 CHANGES INDUNN INDEX WHEN CMDS IS USED

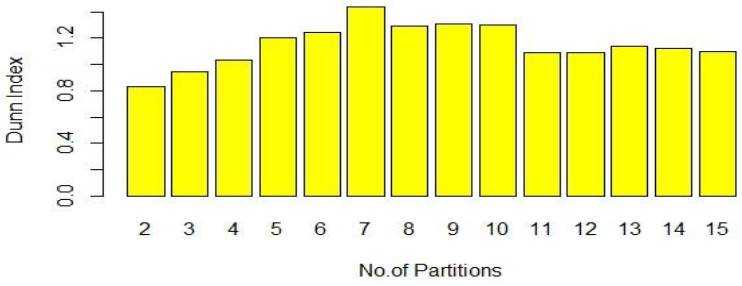


FIGURE NO.10 CHANGES IN DUNN INDEX WHEN ISOMAP IS APPLIED

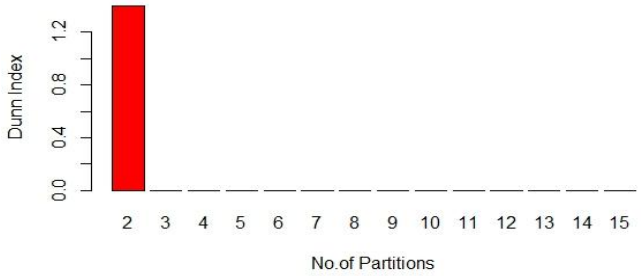


FIGURE NO.11 DUNN INDEX WITH HLLE AS DIMENSIONALITY REDUCTION TECHNIQUE

It is evident from figure 6 to figure 11that Dunn Index values from range from **0 to 2.8** for different dimensionality reduction technique (the values are highest when no dimensionality reduction technique is applies). **So according to dunn index values, dimensionality reduction techniques don't improve performance of k-medoids clustering algorithm.**

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.

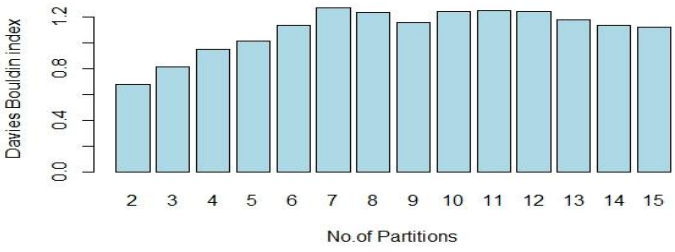


FIGURE NO. 12. CHANGES IN DAVIES-BOULDIN INDEX WHEN NO DIMENSIONALITY REDUCTION TECHNIQUE IS EMPLOYED

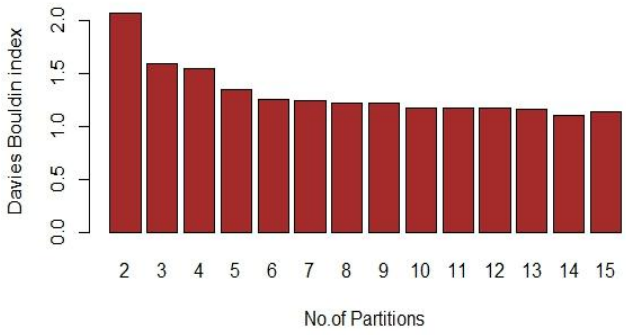


FIGURE NO.13. CHANGES IN DAVIES-BOULDIN INDEX WHEN PRINCIPAL COMPONENT ANALYSIS IS EMPLOYED

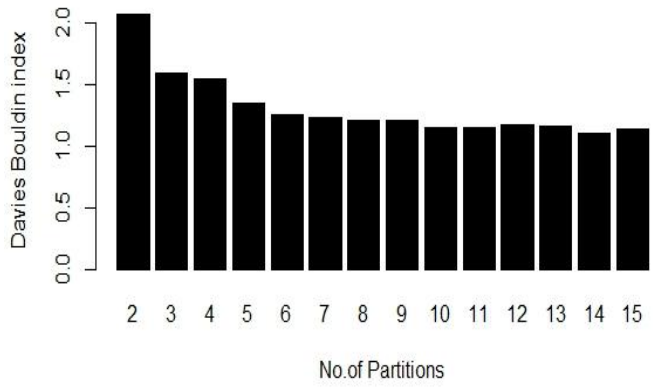


FIGURE NO. 14. CHANGES IN DAVIES-BOULDIN INDEX WHEN CMDS IS EMPLOYED

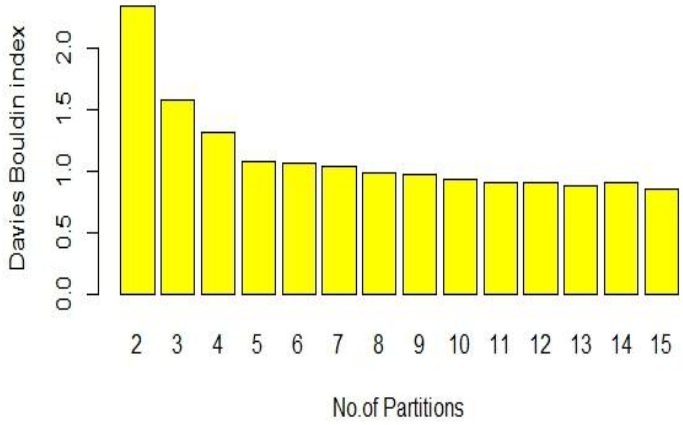


FIGURE NO. 15. CHANGES IN DAVIES-BOULDIN INDEX WHEN ISOMAP IS EMPLOYED

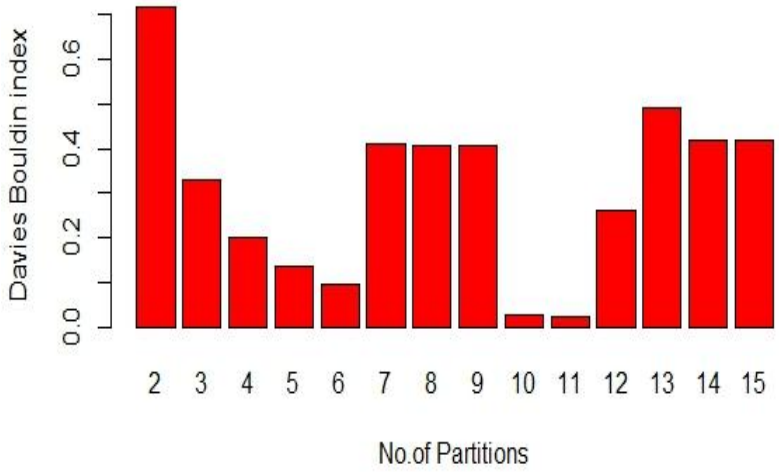


FIGURE NO. 16. CHANGES IN DAVIES-BOULDIN INDEX WHEN HLLE IS EMPLOYED

The figures 12 to 16 show that values of Davies-Bouldin Index approaches a **lower value** for correct number of clusters (The exception being FIGURE16 where Davies-Bouldin index approaches lowest value for inaccurate number of clusters). But the variation in Davies-Bouldin indices is not much and so nothing can be concluded about efficacy of dimensionality reduction techniques.

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.

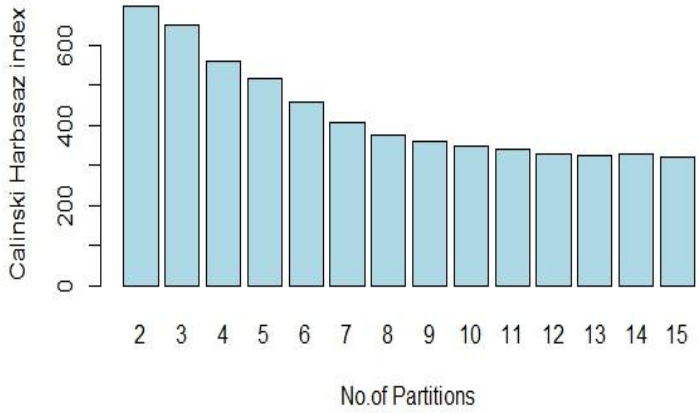


FIGURE NO.17. CHANGES IN CALINSKI HARBASAZ INDEX WHEN NO DIMENSIONALITY REDUCTION IS EMPLOYED

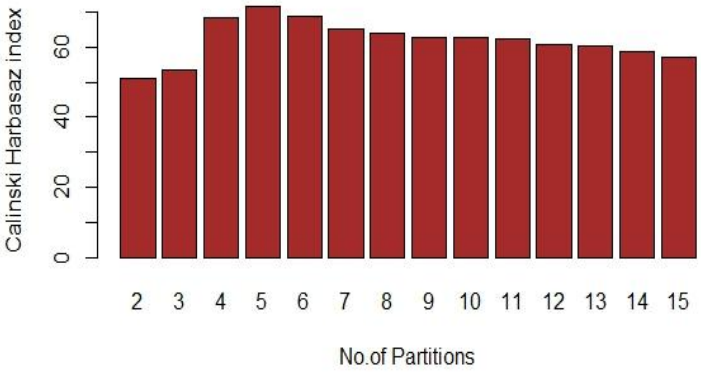


FIGURE NO.18. CHNAGES IN CALINSKI HARBASAZ INDEX WHEN PCA IS EMPLOYED

FIGURE NO.19. CHANGES IN CALINSKI HARBASAZ INDEX WHEN CMDS IS EMPLOYED

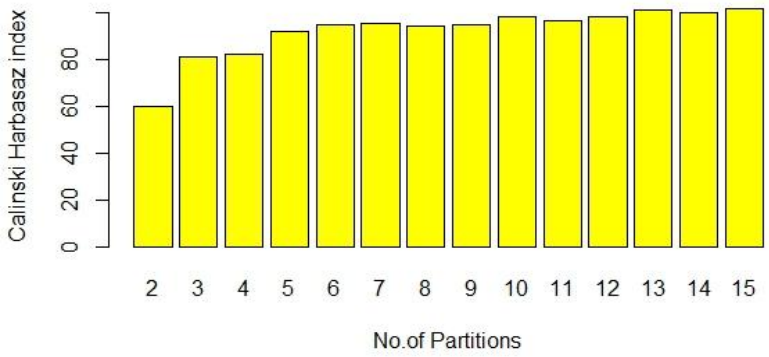
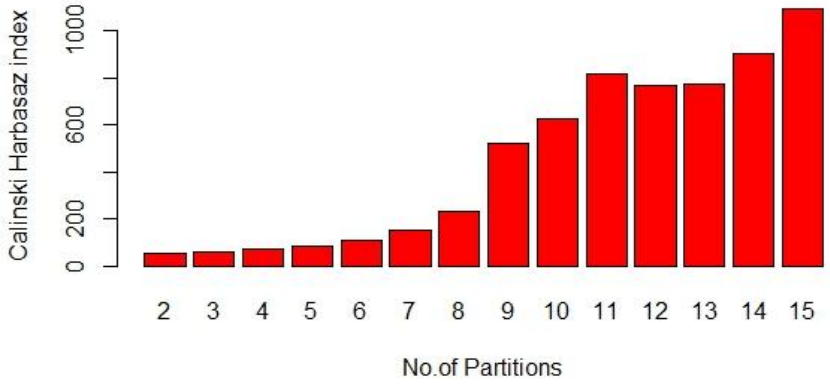


FIGURE NO.20. CHANGES IN CALINSKI HARBASAZ INDEX WHEN ISOMAP IS EMPLOYED
FIGURE NO.21. CHANGES IN CALINSKI HARBASAZ INDEX WHEN HLLE IS EMPLOYED



Figures 17 to 20 are inconclusive. The range of values for Calinski-Harbasaz (CH) Index remains the same from 50 to 100 (except in case of FIGURE 17 where the values range from 650 to 320). 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 50 to 1100, 1100 being the value for number of partitions equal to 15 and 50 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 (50-100 to 50-1100) indicate 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. K-MEDOID DATA CLUSTERING ALGORITHM

Based on above result the following figure 22 **An improved** depicts an optimized data clustering K-medoid algorithm.

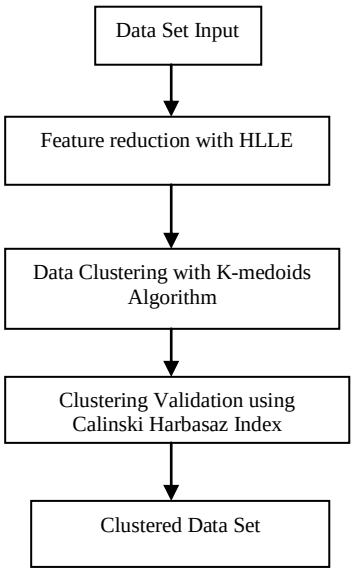


FIGURE NO.22. AN OPTIMIZED DATA CLUSTERING ALGORITHM

As it is clear from the above figure an optimized k-medoid data clustering algorithm would require feature reduction with HLLE and clustering validation using HLLE.

VIII. CONCLUSIONS

In this paper different techniques of dimensionality reduction in conjunction with K-medoid algorithm 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 PCA, CMDS and ISOMAP for improvement of clustering results. Also **Calinski-Harbasaz** Index outperforms Dunn Index, Davies Bouldin Index and Silhouette Index for validating data clustering.

ACKNOWLEDGMENT

We 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.

