International Journal of Advanced Research in Engineering and Technology (IJARET)
Volume 11, Issue 11, November 2020, pp. 1337-1352. Article ID: IJARET_11_11_121
Available online at http://www.iaeme.com/ijaret/issues.asp?JType=IJARET&VType=11&IType=11
ISSN Print: 0976-6480 and ISSN Online: 0976-6499
DOI: 10.34218/IJARET.11.11.2020.121

KERNEL PCA BASED DIMENSIONALITY REDUCTION TECHNIQUES FOR PREPROCESSING OF TELUGU TEXT DOCUMENTS FOR CLUSTER ANALYSIS

Srinivas Mekala
Research scholar, JNTUH, Hyderabad, Telangana, India.

Dr. B. Padmaja Rani
Supervisor, Professor, JNTUH, Hyderabad, Telangana, India.

ABSTRACT

In this paper we focus on investigating the effect of Dimensionality reduction on text document clustering. Clustering is the process of finding groups of objects such that the objects in a group will be similar to one another and different from the objects in other groups. Dimensionality reduction is the transformation of high dimensional data into a meaningful representation of reduced dimensionality of the data. Indian languages are highly inflectional. The dimension of the feature vector hence is very large resulting in poor performance when K-means clustering algorithm is applied. To improve the clustering efficiency KPCA (Kernel Principal Component Analysis) technique is investigated on Indic Script documents and obtained a reduced data set. We aim to investigate Principle Component Analysis (PCA), and Kernel PCA feature reduction technique (KPCA) for dimensionality reduction on Indic script documents and then apply to K-means clustering algorithm. Telugu text documents are chosen as case study for a baseline. Various Kernel functions applied for improving efficiency is also aimed and compared the results with basic PCA technique.

Keywords: Dimensionality reduction, Clustering, K-means clustering algorithm, Principal Component Analysis (PCA),Kernel Principal Component Analysis (Kernel PCA).

Cite this Article: Srinivas Mekala and Dr. B. Padmaja Rani, Kernel PCA Based Dimensionality Reduction Techniques for Preprocessing of Telugu Text Documents for Cluster Analysis, *International Journal of Advanced Research in Engineering and Technology*, 11 (11), 2020, pp. 1337-1352.

http://www.iaeme.com/IJARET/issues.asp?JType=IJARET&VType=11&IType=11

1. INTRODUCTION

Text document clustering is a fundamental and enabling tool for efficient document organization, summarization, navigation and retrieval. Text clustering plays a significant role in Machine Learning and Information Retrieval, which facilitate Knowledge Discovery from text mining. As part of text mining, we analyse large quantities of text data, detect useful patterns and extract precious information or knowledge. The most critical problem for text document clustering is the *high dimensionality* of the natural language text, often referred to as the *"curse of dimensionality"*. High dimensionality problem is addressed under Data Reduction strategies. Data Reduction is achieved through Dimensionality reduction, Numerosity reduction and Data Compression. In this paper we focus on dimensionality reduction approaches which can be categorized according to different perspectives such as, *linear* versus *non-linear*. Wavelet Transforms and Principal Component Analysis are linear dimensionality reductions techniques in which Feature extraction method project the original high dimensional space onto a lower dimensional space. The other Attribute subset selection methods for Dimensionality reduction select a subset of "meaningful" dimensions from the original ones. This paper analyses the effect of various dimensionality reduction techniques for natural language text documents under nonlinear dimensionality reduction category for Clusterization process which is a significant step in text mining. Feature extraction methods like Principal Component Analysis (PCA) [7] and a nonlinear Kernel PCA with several Kernel functions [11] are compared with base method TFIDF which have an established reputation in text document dimensionality reduction.

2. ISSUES RELATED TO DIMENSIONALITY REDUCTION

2.1. Curse of Dimensionality Reduction

The curse of dimensionality refers to various phenomena that arise when analysing and organizing data in high dimensional space (often with hundreds or thousands of dimensions) that do not occur in low-dimensional settings such as the three-dimensional physical space of everyday experience. Most Machine learning and Data mining algorithms may not effective for high-dimensional data. Query accuracy and efficiency degrade rapidly as dimension increases. The expression was coined by Richard E. Bellman when considering problems in dynamic optimization. Telugu language documents analysis for Knowledge discovery is done through Text Mining process, which is an interdisciplinary field that draws on information retrieval, data mining, machine learning, Statistics, and Computational linguistics. Enormous information is stored as text such as news articles, books, digital libraries, email messages, blogs, and web pages. Hence, a high degree of dimensionality exists analysis of Text data.

There are multiple phenomena referred to by this name in domains such as numerical analysis, sampling, combinatory, databases. The common theme of these problems is that when the dimensionality increases, the volume of the space increases so fast that the available data become sparse. This sparsity is problematic for any method that requires statistical significance. In order to obtain a statistically sound and reliable result, the amount of data needed to support the result often grows exponentially with the dimensionality. Also, organizing and searching data often relies on detecting areas where objects form groups with similar properties; in high dimensional data, however, all objects appear to be sparse and dissimilar in many ways, which prevents common data organization strategies from being efficient. There are statistical techniques which can find the best representation of data in a lower-dimensional space than that in which it was originally provided

2.2. Enhancement in clustering through Dimensionality reduction

We need to derive high-quality information from text. To obtain optimized results for a high dimension data set it is needed to apply dimensionality reduction. Removing redundant and noisy features, discovery of hidden correlations or topics improve clusterization so that interpretation and visualization of results is easy. Intrinsic dimension may be small in some cases, for example number of genes cause for disease is small.

2.3. Approaches to handle Dimensionality reduction

Dimensionality reduction is a pre-process step in which size of the vector space form reduced. The reduced space is called reduced term set. Dimensionality reduction methods have been derived from information theory or linear algebra literature. Dimensionality reduction methods primarily categorized based on applicability of domain, like natural language text is undergoing nonlinear dimensionality reduction. In this paper nonlinear dimensionality reduction techniques are surveyed for dimensionality reduction by applying different type of Kernel functions, that improves clustering efficiency which in turn improves text mining.

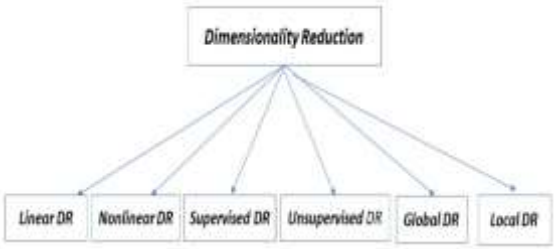

Figure 1 Primary categorization of Dimensionality Reduction.

There are three types of linear dimensionality reduction methods, in which feature extraction has applied to the following methods for dimensionality reduction. Principal Component Analysis (PCA) Independent Component Analysis (ICA)Orthogonal Component Analysis (OCA). Nonlinear dimensionality reduction methods are Kernel PCA, ISOMAP, LLE, MDS, and MVU.

In text clustering process, the documents or examples are represented by thousands of tokens, which make the classification problem very hard for many classifiers. Dimensionality reduction is a typical step in text mining, which transform the data representation into a shorter, more compact, and more predictive one. The new space is easier to handle because of its size, and space that describe the data set effectively. Limitations of linear dimensionality reduction through Principal Component Analysis (PCA) are understood and studied Nonlinear dimensionality reduction techniques by applying Kernel functions. Telegu language dataset is applied for study of the Kernel Principal Component Analysis methods.

2.3.1. Principle Component Analysis (PCA)

PCA is a well-known technique that can reduce the dimensionality of data by transforming the original attribute space into smaller space. In the other words, the purpose of principle components analysis is to derive new variables that are combinations of the original variables and are uncorrelated. This is achieved by transforming the original variables Y = [y1, y2... yp] (where p is number of original variables) to a new set of variables, T = [t1, t2... tq] (where q is number of new variables), which are combinations of the original variables.[10] Transformed attributes are framed by first, computing the mean (μ) of the dataset, then covariance matrix of the original attributes is calculated [5]. The second step is, extracting its eigenvectors and these eigenvectors (principal components) introduce as a linear transformation from the

original attribute space to a new space in which attributes are uncorrelated. Eigenvectors can be sorted according to the amount of variation in the original data. The best n eigenvectors (those one with highest eigenvalues) are selected as new features while the rest are discarded.

Procedure for PCA

1. Scale the given matrix X with respect to its mean

   i.e. X-$\overline{X}$
2. Find covariance matrix of given matrix X

$$\text{Covariance matrix } (\sum) = \frac{\mathbf{X}\mathbf{X}^{\mathbf{T}}}{\mathbf{n-1}}$$

3. Find the eigen values of the system

$$\sum \boldsymbol{\alpha_i} = \boldsymbol{\lambda_i}\boldsymbol{\alpha_i}$$

These $\boldsymbol{\alpha_i}$ are called principal components, where is an eigen vector corresponding to the eigen value $\boldsymbol{\lambda_i}$.

4. The Components corresponding to larger eigen values explain most of the variance in the data and are hence considered important whereas the eigen vectors corresponding to lower eigen values can be discarded.

Limitations of the PCA are as follows:

1. It assumes that the relationships between variables are linear.
2. Its interpretation is only sensible if all of the variables are assumed to be scaled at the numeric level.
3. It lacks a probabilistic model structure which is important in many contexts such as mixture modelling and Bayesian decision.

2.3.2. Kernel Principle Component Analysis (KPCA)

Kernel Principle Component Analysis (KPCA*)* overcomes the first limitation by using a kernel trick. The essential idea of KPCA is avoid the direct evaluation of the required dot product in a high dimensional feature space using the kernel function. Therefore, no explicit nonlinear function projecting the data from the original space to the feature space is needed. In an approach to analyse kernel principle component in a probabilistic manner has been proposed called probabilistic kernel principle component analysis (PKPCA) that naturally combines PPCA and KPCA to overcome limitations of PCA.

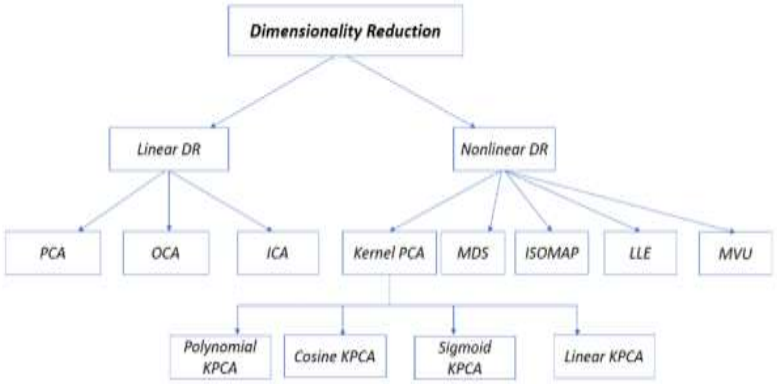

Figure 2 Taxonomy of nonlinear Dimensionality Reduction

Kernel PCA is achieved through the following given procedure.

1. Project the data on to the higher dimensional space that makes it linearly separable. This is done by adding a new dimension which is a function of existing dimensions. The high dimensional mapping is represented as φ.

Example: - $(x_1, x_2) \rightarrow (x_1, x_2, x^2_1 + x^2_2)$

$\varphi: R^2 \rightarrow R^3$

2. The covariance matrix in the higher dimension is represented as $\sum = \varphi(x)\ \varphi(x)^T$

3. This can be easily calculated as dotproduct of feature vectors in the high dimensional space without actually calculating the high dimensional representation. Typically done via kernel method to obtain a matrix as follows.

$$K = \begin{bmatrix} K(X_1,X_1) & K(X_1,X_2) & \ldots. & K(X_1,X_d) \\ K(X_2,X_1) & K(X_2,X_2) & \ldots. & K(X_2,X_d) \\ \ldots.. & \ldots.. & \ldots. & \ldots.. \\ K(X_d,X_1) & K(X_d,X_2) & \ldots. & K(X_d,X_d) \end{bmatrix}$$

K (X_1, X_2) is called the Kernel function. Most Kernel functions used are Polynomial, Sigmoid, Cosine, and Linear functions.

3. PROPOSED METHOD

In proposed method unsupervised text mining process is applied for Telegu text documents data set. Text documents are given for pre-processing, words or terms are extracted and doing the dimensionality reduction. Kernel Pca is applied for dimensionality reduction. The data reduced documents have given for clusterization process. K-means clustering algorithm is applied for clustering dimensionality reduced documents. The following steps are adopted for proposed methodology.

3.1. Pre processing

Telugu text documents from three domains Sports, Political News, and literature are given for pre-processing. Lexicalization is done by eliminating special characters and tokenization. Terms are extracted from the documents. Vectorization is done and obtained Term Frequency matrix and Inverse Document Frequency matrix. In the whole pre-processing step, the TF_IDF vector is produced and given for dimensionality reduction by applying Kernel PCA technique.

3.2. Dimensionality Reduction by applying Kernel PCA

Data reduction is a significant step in the Knowledge discovery process. When applying text mining algorithms on massive amounts of data for extraction of knowledge, large size of data cannot give efficient results. In order to improve efficiency of results, dimensionality reduction which is a method of data reduction is applied as a prior step that effects the text mining algorithm applied. Kernel Principal Component Analysis is applied to reduce data dimensionality. The main idea of in Kernel PCA is to map from input data x via a nonlinear mapping $\Phi(x)$ to feature space F, and then execute the linear PCA in the feature space F. For the computation of eigen value in the feature space and the vector projection in the feature space, KPCA does not require the mapping $\Phi(x)$ having explicit format, but only computing the dot product can use the kernel function given as follow to compute. The nonlinear of KPCA is achieved by kernel transformation, transforming input space to Hilbert feature space, so it can be said that the PCA is computed in the input space, while Kernel PCA in the feature space.

$Kij = k(xi, xj) = (\Phi(xi).\ \Phi(xi))$

The following steps carried out in order to perform dimensionality reduction.

i. Pick a Kernel function K (Xi, Xj)
ii. Calculate the Kernel Matrix K
iii. Centre the Kernel Matrix.
iv. Solve the eigen system

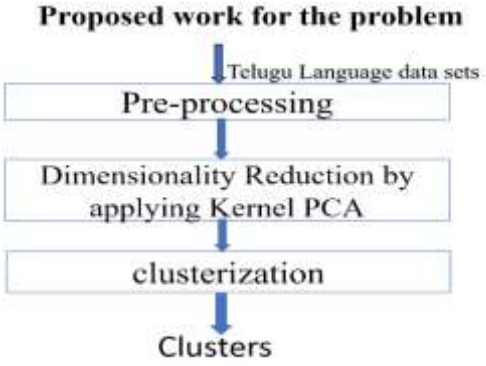

Figure 3 Procedure for DR by using Kernel PCA

3.3. Properties of Kernel PCA

If we use a kernel which satisfies the conditions for computing Dot products in Feature Space, we know that we are in fact doing a standard PCA in F. Consequently, all mathematical and statistical properties of PCA[4] carry over to kernel based PCA, with the modifications that they become statements about a set of points $\Phi(x_i)$, i=1,…M, in F rather than in RN. In F, we can thus assert that PCA is the orthogonal basis transformation with the following properties. (Assuming that the Eigenvectors are sorted in ascending order of the Eigenvalue size).

i. The first q ($q \epsilon \{1,\ldots,M\}$) principal components, i.e. projections on Eigenvectors, carry more variance than any other q orthogonal directions.
ii. The mean-squared approximation error in representing the observations by the first q principal components is minimal.
iii. The principal components are uncorrelated
iv. The representation entropy is minimized
v. The first q principal components have maximal mutual information with respect to the inputs.

3.4. Clustering Low dimensional data

After applying Kernel PCA technique we obtain a reduced set of components which can be applied for clusterization process. For better interpretation of the data the reduced data set can be clustered. We adopt k-means clustering algorithm for clusterization.

k-means clustering algorithm

We obtain the purified clusters by applying the original k-means clustering algorithm. The idea is to classify a given set of data into k number of disjoint clusters where the value of k is fixed in advance. The algorithm consists of two phases. The first phase is to define k centroids one for each cluster.[9] The next phase is to take each point belonging to the given data set and associate it to the nearest centroid. Pseudo code for the algorithm is given as follows.

Input

X= {x1, x2,…..,xn}// set of n data items

K // Number of desired clusters

Output: A set of k clusters

Steps:

1. Arbitrarily choose k data items from X as initial centroids;

2. Repeat

Assign each item xi to the cluster which has the closest centroid;

Calculate new mean for each cluster Until convergence criteria is met

After clusterization process we would get a set of clusters and each cluster contains a set of documents that are most related.

4. IMPLEMENTATION

Telugu Language dataset collected from three categories namely Sports, Political News and Literature comprising 141 documents are considered for experimentation. Both PCA and KPCA techniques are applied. In Kernel PCA the Linear, Polynomial, Cosine, and Sigmoid Kernel methods are used to reduce dimensionality. As dimensionality reduced k-means clustering is applied on reduced documents. The clustering accuracy is measured in terms of the ratio correctly assigned documents to clusters and total documents. Dimension reduction is measured as the ratio of reduced words to the total words in a document. Cluster evaluation is done by applying several cluster efficiency index methods like, Davies Bouldin Index, Completeness, and V-measure.

The following tables depicts the number of components selected and corresponding dimensionality reduced in percentage. The resultant clustering efficiency is shown as cluster evaluation index. The experimentation is done in two phases, first we apply dimensionality reduction technique and then as a second phase applying clustering algorithm

Table 1 Dimensionality reduction achieved when PCA applied and Cluster efficiency indices

Components Chosen for PCA	DR % achieved	v-measure	completeness
141	0	0.2780651088	0.28827468335
136	3.546	0.3608815373	0.3618374002
131	7.092	0.3958254142	0.3990624222
126	10.064	0.3290786855	0.3345533676
121	14.184	0.3024355354	0.3294447764
116	17.731	0.2293089243	0.2342017637
111	21.277	0.2210406248	0.2423483445
106	24.823	0.1153220286	0.1856493850
101	28.369	0.1265987441	0.1402206158
96	31.915	0.3070825983	0.3077338797
91	35.461	0.1608175280	0.1717783444
86	39.007	0.1787879770	0.1964907852
81	42.553	0.1513246566	0.2747459050

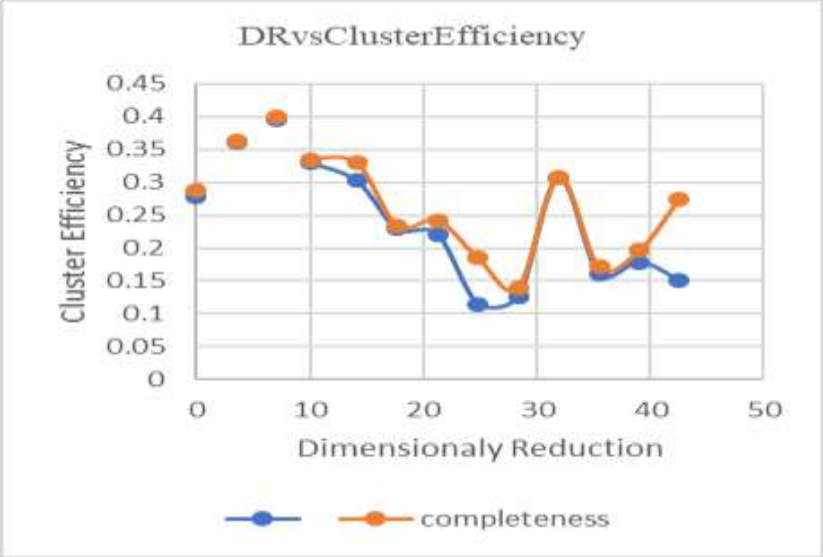

Figure 4 Dimensionality reduction with PCA.

Table 2 Dimensionality reduction achieved when PCA applied and Cluster efficiency index.

Components Chosen for PCA	DR % achieved	Davies Bouldin Index
141	0	6.8255869158
136	3.546	6.9887774883
131	7.092	6.8871047776
126	10.064	6.9561863562
121	14.184	6.6044458757
116	17.731	7.0685424846
111	21.277	6.8288590088
106	24.823	4.0856591818
101	28.369	6.8009962824
96	31.915	7.1591625185
91	35.461	7.0761557663
86	39.007	6.6671322717
81	42.553	3.8173672397

Table 3 Dimensionality reduction achieved when Polynomial KPCA applied and Cluster efficiency indices

Components Chosen for KPCA	DR % achieved	Davies Bouldin Index
141	0	6.9302001706
136	3.546	6.9887774883
131	7.092	6.8871047776
126	10.064	6.9561863562
121	14.184	6.6044458757
116	17.731	7.0685424846
111	21.277	3.3746448835
106	24.823	5.0773243294
101	28.369	5.2081407357
96	31.915	4.8140699927

91	35.461	6.4449270687
86	39.007	6.0796049994
81	42.553	6.7301045869

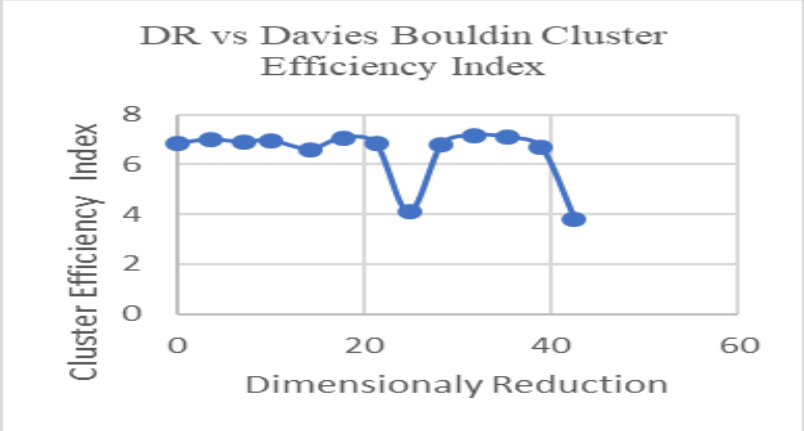

Figure 5 Dimensionality reduction with PCA.

Table 4 Dimensionality reduction achieved when Polynomial KPCA applied and Cluster efficiency

Components Chosen for KPCA Polynomial	**DR % achieved**	**v-measure**	**completeness**
141	0	0.4142154538	0.4181524566
136	3.546	0.3608815373	0.3618374002
131	7.092	0.3958254142	0.3990624222
126	10.064	0.3290786855	0.3345533676
121	14.184	0.3024355354	0.3294447764
116	17.731	0.2293089243	0.2342017637
111	21.277	0.0749765416	0.1503420265
106	24.823	0.0954856231	0.1305788924
101	28.369	0.1283104227	0.1857023633
96	31.915	0.1377788797	0.1940382829
91	35.461	0.1732531637	0.1931722836
86	39.007	0.1088373110	0.1310298533
81	42.553	0.1300183036	0.1421154675

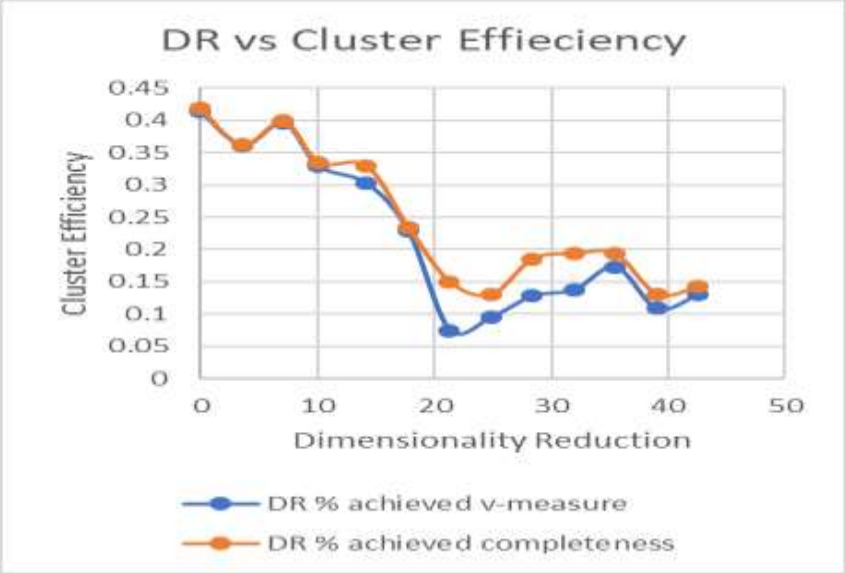

Figure 6 Dimensionality reduction with Polynomial Kernel PCA.

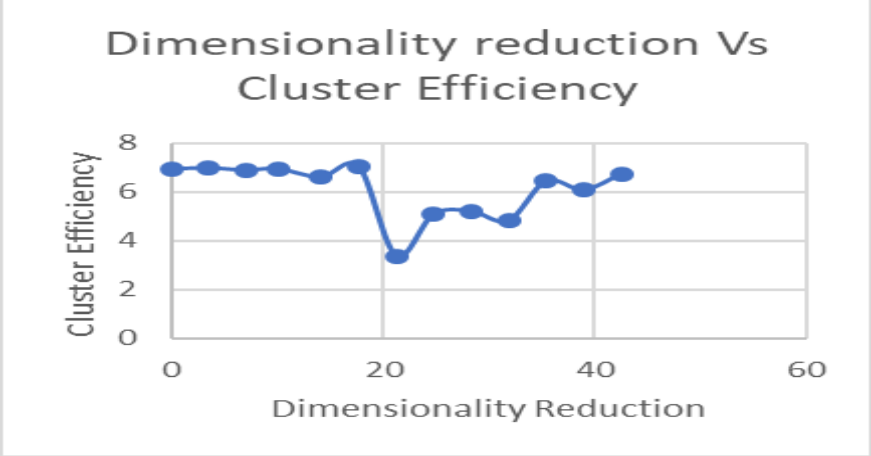

Figure 7 Dimensionality reduction with Polynomial Kernel PCA

Table 5 Dimensionality reduction achieved when Sigmoid KPCA applied and Cluster efficiency

Components Chosen for Sigmoid Kernel PCA	DR % achieved	v-measure	completeness
141	0	0.3181525669	0.3196680331
136	3.546	0.3618374002	0.3608815373
131	7.092	0.3958254142	0.3990624222
126	10.064	0.3290786855	0.3345533676
121	14.184	0.3024355354	0.3294447764
116	17.731	0.2293089243	0.2342017637
111	21.277	0.0749765416	0.1503420265
106	24.823	0.0954856231	0.1305788924
101	28.369	0.1283104227	0.1857023633
96	31.915	0.1377788797	0.1940382829
91	35.461	0.1732531637	0.1931722836
86	39.007	0.1088373110	0.1310298533
81	42.553	0.1300183036	0.1421154675

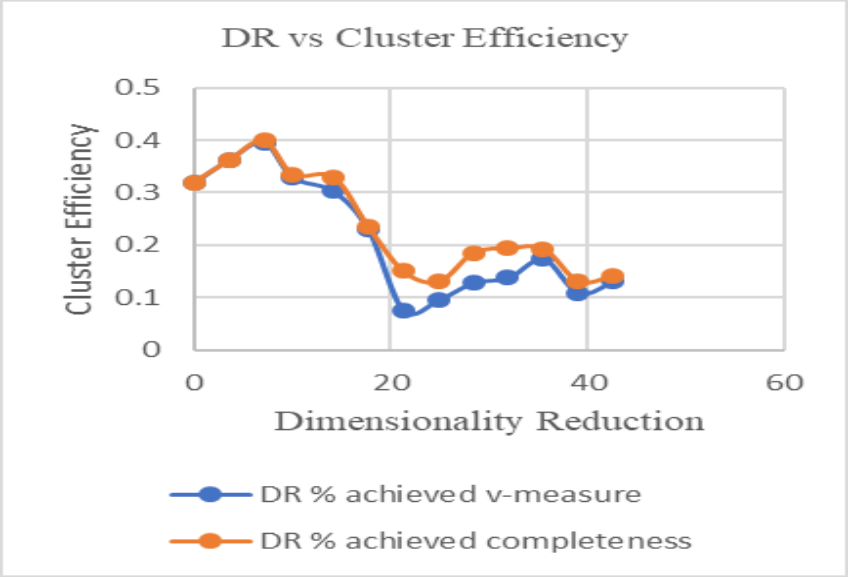

Figure 8 Dimensionality reduction with Sigmoid Kernel PCA

Table 6 Dimensionality reduction achieved when Sigmoid KPCA applied and Cluster efficiency

Components Chosen for Sigmoid Kernel PCA	DR % achieved	Davies Bouldin Index
141	0	6.9887774883
136	3.546	6.9887774883
131	7.092	6.8871047776
126	10.064	6.9561863562
121	14.184	6.6044458757
116	17.731	7.0685424846
111	21.277	3.3746448835
106	24.823	5.0773243294
101	28.369	5.2081407357
96	31.915	4.8140699927
91	35.461	6.4449270687
86	39.007	6.0796049994
81	42.553	6.7301045869

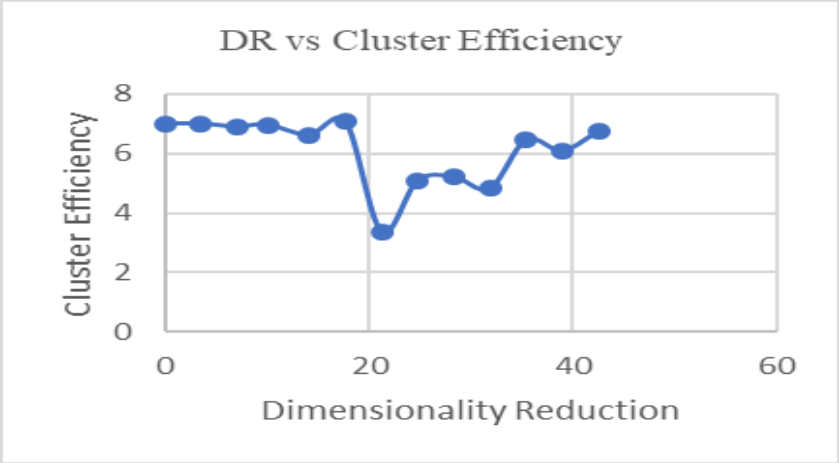

Figure 9 Dimensionality reduction with Sigmoid Kernel PCA

Table 7 Dimensionality reduction achieved when Cosine KPCA applied and Cluster efficiency

Components Chosen for Cosine Kernel PCA	DR % achieved	v-measure	completeness
136	3.546	0.3608815373	0.3618374002
131	7.092	0.3958254142	0.3990624222
126	10.064	0.3290786855	0.3345533676
121	14.184	0.3024355354	0.3294447764
116	17.731	0.2293089243	0.2342017637
111	21.277	0.0749765416	0.1503420265
106	24.823	0.0954856231	0.1305788924
101	28.369	0.1283104227	0.1857023633
96	31.915	0.1377788797	0.1940382829
91	35.461	0.1732531637	0.1931722836
86	39.007	0.1088373110	0.1310298533
81	42.553	0.1300183036	0.1421154675

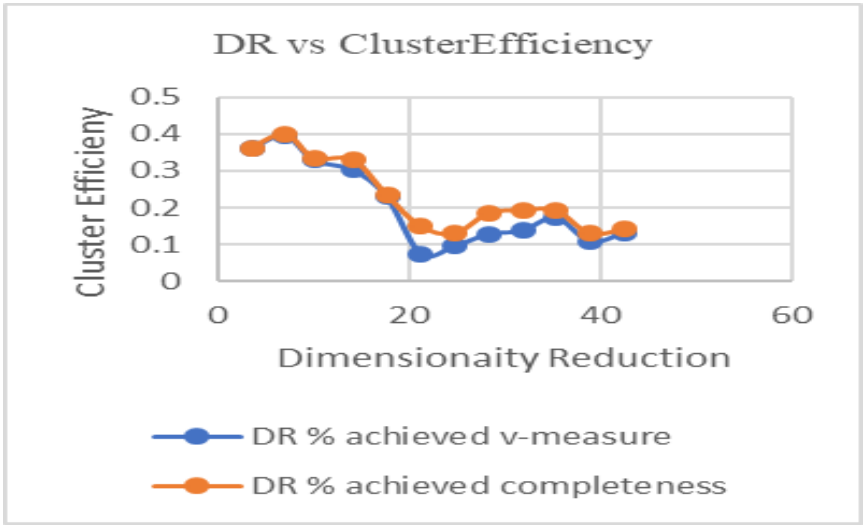

Figure 10 Dimensionality reduction with Cosine Kernel PCA

Table 8 Dimensionality reduction achieved when Cosine KPCA applied and Cluster efficiency

Components Chosen for Sigmoid Kernel PCA	DR % achieved	Davies Bouldin Index
136	3.546	6.9887774883
131	7.092	6.8871047776
126	10.064	6.9561863562
121	14.184	6.6044458757
116	17.731	7.0685424846
111	21.277	3.3746448835
106	24.823	5.07732432942
101	28.369	5.20814073577
96	31.915	4.81406999275
91	35.461	6.44492706879
86	39.007	6.07960499941
81	42.553	6.73010458699

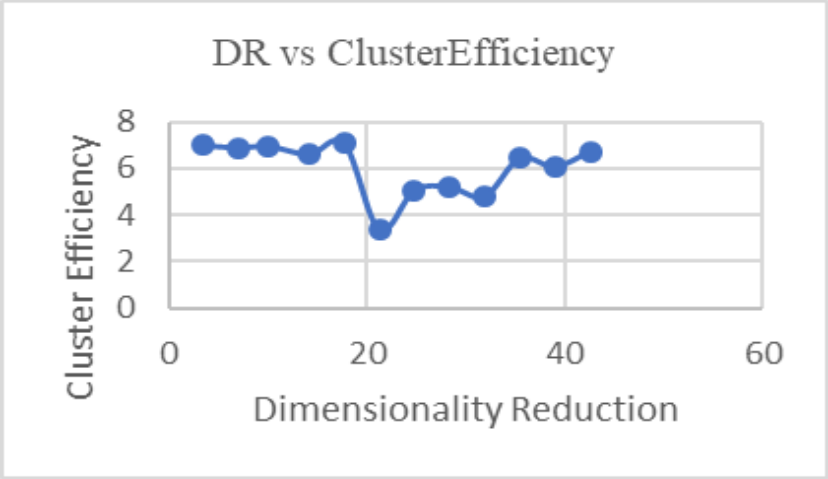

Figure 11 Dimensionality reduction with Cosine Kernel PCA

Table 9 Dimensionality reduction achieved when linear KPCA applied and Cluster efficiency

Components Chosen for KPCA linear n	DR % achieved	Completeness	V-Measure
136	3.546	0.3618374002	0.3608815373
131	7.092	0.3990624222	0.3958254142
126	10.064	0.3345533676	0.3290786855
121	14.184	0.3294447764	0.3024355354
116	17.731	0.2342017637	0.2293089243
111	21.277	0.1503420265	0.0749765416
106	24.823	0.1305788924	0.0954856231
101	28.369	0.1857023633	0.1283104227
96	31.915	0.1940382829	0.1377788797
91	35.461	0.1931722836	0.1732531637
86	39.007	0.1310298533	0.1088373110
81	42.553	0.1421154645	0.1300183036

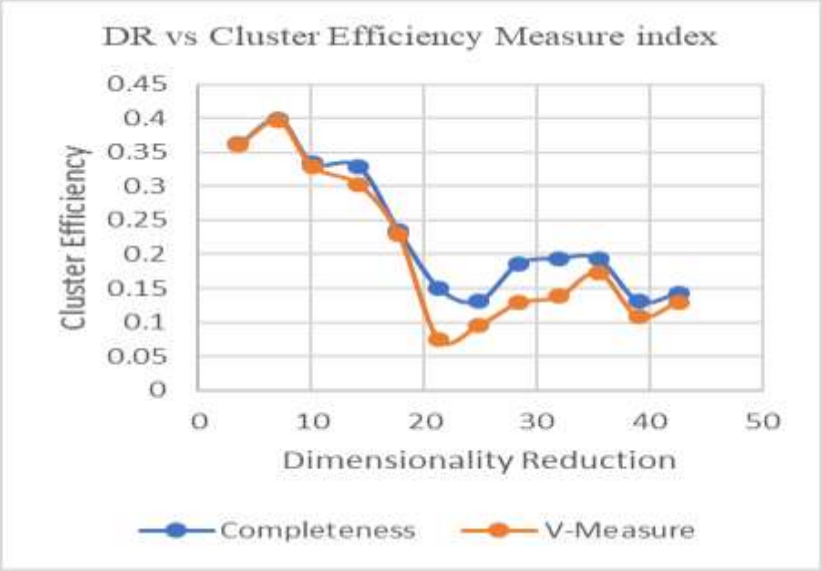

Figure 12 Dimensionality reduction with Cosine Kernel PCA

Table 10 Dimensionality reduction achieved when linear KPCA applied and Cluster efficiency

Components Chosen for KPCA linear function	DR % achieved	Davies Bouldin Index
136	3.546	6.9887774883
131	7.092	6.8871047776
126	10.064	6.9561863562
121	14.184	6.6044458757
116	17.731	7.0685424846
111	21.277	3.3746448835
106	24.823	5.0773243294
101	28.369	5.2081407357
96	31.915	4.8140699927
91	35.461	6.4449270687
86	39.007	6.0796049994
81	42.553	6.7301045869

5. CONCLUSION

We aim to investigate Kernel PCA feature reduction technique (KPCA) for dimensionality reduction on Indic script documents and then apply to K-means clustering algorithm. Telugu text documents are chosen as case study. Various ways of improving efficiency is also aimed to investigate and compare the result with basic PCA technique. Telugu documents suffer from more inflectionality initial reduction resulted in words with its variants mostly efficiency is increased. Whereas further reduction eliminated some of root words and resulting in efficiency decrease.

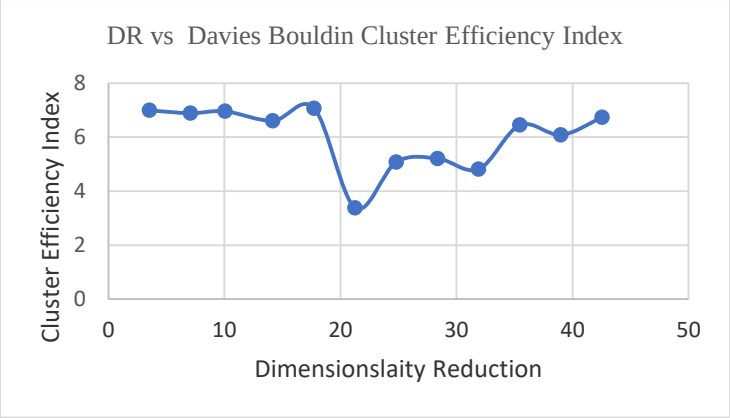

Figure 13 Dimensionality reduction with Cosine Kernel PCA

It is observed that for both PCA and Kernel PCA are consistent as the dimensionality reduction is increased. Then clustering efficiency is increased to a level and then dropped. It is found that for dimensionality reduction at 7% right efficiency of clustering is 0.39 is obtained as computed even dimensionality reduction is increased Davies Bouldin Index is almost constant with the range 6 to7.

REFERENCES

[1] Bernhard Scholkopf, Alexander Smola, Klaus-Robert Muller, "Kernel Principal Component Analysis", Vol.1327, pp.583-588, Artificial Neural Networks- ICANN1997.

[2] B. Padmaja Rani, B. Vishnu Vardhan, A. Kanaka Durga, L. Pratap Reddy, A. Vinay Babu, "Analysis of N-gram model on Telugu Document Classification", IEEE-2008.

[3] Sam T Roweis and Lawrence K. Saul, "Nonlinear Dimensionality Reduction by Locally Linear Embedding" Vol.290, pp.2323-2326,2000.

[4] Bernhard Scholkopf, Alexander Smola, and Klaus-Robert Muller "Nonlinear Component Analysis as a Kernel Eigenvalue Problem", IEEE journal on Neural Computation, Vol:10 and Issue:5, pp.1299 – 1319, IEEE-1998.

[5] Vasileios megalooikonomou Guo Li &Qiang Wang "A Dimensionality Reduction Technique for Efficient Similarity Analysis of Time Series Databases", pp.160-161, ACM-2004.

[6] Mykola Penchenizkiy, Seppo Puuronen "Comparing Dimension Reduction Techniques for Document Clustering", pp.553-558, ACM-2006.

[7] Mahdi Shafiei Singer Wang "Document Representation and Dimension Reduction for Text Clustering", pp.770-779, IEEE-2007.

[8] Hans-peter Kriegel peer Kroger Arthur Zimek "Clustering High-Dimensional Data: A Survey on Subspace Clustering, Pattern-Based Clustering, and Correlation Clustering",58 pages, ACM-2009.

[9] N. Tajunisha V Saravanan "An increased performance of Clustering high dimensional data using Principal Component Analysis", PP.17-21, IEEE-2010.

[10] Jaffali Soufiene, Jamoussi Salma "Text document dimension reduction using Principal Component Analysis",2012

[11] Maysa I Abdulhussain John Q Gan, "An Experimental Investigation on PCA Based on Cosine Similarity and Correlation for Text Feature Dimensionality Reduction", IEEE-2015.

ABOUT AUTHORS

Dr. B. Padmaja Rani received her B.Tech(ECE) from Osmania University, M.Tech Computer Science from JNTU, Hyderabad and Ph.D. from JNTU, Hyderabad. She is currently working as a Professor in the department of Computer Science and Engineering JNTUH College of Engineering, JNTUH, Hyderabad. She is having 25 years of experience in Industry and Academia. Her area of Research includes Information Retrieval, Data Mining, Machine Translation, Cloud Computing, Software Engineering, Computer Networks etc. She is guiding 8 Research Scholars in the area of Information Retrieval and Computer Networks. To her credit she is having more than 70 publications in international Journals and Conferences. She is a member of various advisory committees and Technical Bodies. She is also a member of various Technical Associations including ISTE, CSI, IEEE etc.

Srinivas Mekala received B.Tech degree in Computer Science and Systems Engineering, from Sir CRR College of Engineering, Andhra University, M.Tech Software Engineering from AIET, JNTUK. Presently he is a research scholar from JNTUH, Hyderabad and also he is working as an Assistant Professor in the department of Computer Science and Engineering, Keshav Memorial Institute of Technology, JNTUH, Hyderabad. His area of Research includes Data Mining, Information Retrieval and Software Engineering