Dimensionality Reduction in Natural Language Text document using PCA Techniques

Srinivas Mekala1, Dr. B. Padmaja Rani2
1Research scholar in JNTUH, Hyderabad, Telangana
2Supervisor, Professor in JNTUH, Hyderabad, Telangana

Abstract-- In this paper to focus on a systematic study of Clustering, the 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 efficiency PCA (Principal Component Analysis) technique to be investigated on Indic Script documents and obtain a reduced data set. we aim to investigate PCA feature reduction technique (PCA) 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 by other means is also aimed to investigate and compare the result with basic PCA technique

Index Terms-- Dimensionality reduction, Clustering, K-means clustering algorithm, Principal Component Analysis (PCA)**.**

I. Introduction

Document clustering is a fundamental and enabling tool for efficient document organization, summarization, navigation and retrieval. The most critical problem for document clustering is the high dimensionality of the natural language text, often referred to as the "curse of dimensionality". While various dimension reduction techniques have been proposed [1, 2], there are two major types, feature transformation and feature selection [2]. Feature transformation methods project the original high dimensional space onto a lower dimensional space, while feature selection methods select a subset of "meaningful" dimensions from the original ones.

The dimensionality reduction techniques like PCA have been popular since the early 90s in text processing tasks [7, 8]. Tsymbal et al. [9] propose two variants of PCA that use the within and between class covariance matrices to take into account the class information. They test the results on typical database data, but not to text categorization. Brutlag and Meek [10] investigate the effect of feature selection by means of common information statistic on email filtering. Xia and Wong [11] discussed the email categorization problem in the context of personal information management.

This paper analyses the effect of various dimensionality reduction techniques in text classification. Feature extraction methods like Principal Component Analysis (PCA) [7] and Latent Semantic Analysis (LSA) [12] are compared with classical feature selection techniques like Chi-Square ($\chi 2$) [13], and Information Gain (IG) [14], which have an established reputation in text classification. In order to study the effectiveness of various dimensionality reduction techniques in phishing email classification, each technique were tested with Bagging classifier [8], which has already proved by researchers, good for e-mail classification.

II. MATERIALS AND METHODS

2.1 Dimensionality Reduction Techniques

In text classification tasks, 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 [8]. The new space is easier to handle because of its size, and also to carry the most important part of the information needed to distinguish between emails, allowing for the creation of profiles that describe the data set. Two major classes of dimensionality reduction techniques are described in the following sections.

2.2 Curse of Dimensionality reduction

The curse of dimensionality refers to various phenomena that arise when analyzing and organizing data in high-dimensional spaces (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. The expression was coined by Richard E. Bellman when considering problems in dynamic optimization.

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**.**

2.3 Text Document Clustering

Text Document Clustering is useful in finding closely related observations based on their respective topics. High number of profiling elements may diminish the effectiveness of clustering process and therefore arises a need to explore use of dimensionality reduction technique as preprocessing step.

2.4 Feature Extraction & Survey

- Title: A Novel Dimensionality Reduction Method for Cancer Dataset using PCA and Feature ranking
- 2015 –ICACCI International Conferences on Advances in Computing, Communications and Informatics, Page no's 2261-2264,
- **Concept:** This paper introduces a novel method to reduce dimensionality using PCA and Feature Ranking. An empirical method is proposed that can reduce number of dimensions, for a given dataset and learning machine algorithm such that high classification accuracy is maintained with minimum number of dimensions.

- Title: Analysis of N-gram model on Telugu document classification
- IEEE Congress on Evolutionary Computation,CEC-2008, 3199 – 3203,
- **Concept:** This paper analyzes character N-gram model on Telugu documents. Tokenization of syllables is described. A combination of Bayes probabilistic classifier and character N-gram model is discussed

- Title: Feature reduction using principal component analysis for agricultural data set.
- 3rd International Conference on Electronics Computer Technology(ICECT),2011,Pagenos 209-213
- **Concept:** In this paper k-means clustering algorithm and PCA approach is used for attribute reduction. PCA is applied first to reduce uncorrelated attributes and then k-means is applied.
- Title: Robust Script Identification using wavelet and EHD Features in PCA Space
- (ICRAIE-2016)IEEE International Conference on Recent Advances and innovations in Engineering. Dec-2016,1-6
- **Concept:** This paper identifies the type of script of printed documents for multi-script OCR system. This scheme uses Wavelet features and MPEG-7 Edge Histogram Descriptor feature.

In feature extraction [8], the original feature space is converted to a more compact new space. All the original features are transformed into the new reduced space without deleting them but replacing the original features by a smaller representative set. That is when the number of feature in input data is too large to be processed then the input data will be transformed into a reduced representation set of features.

III. PROPOSED METHOD

3.1 Principal Components 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 word, 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 variable) to a new set of variables, T = [t1, t2... tq] (where q is number of new variables), which are combinations of the original variables. Transformed attributes are framed by first, computing the mean (μ) of the dataset, then covariance matrix of the original attributes is calculated [5]. And the second step is, extracting its eigenvectors. The 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.

3.2 Latent semantic Analysis (LSA)

LSA method is a novel technique in text classification. Generally, LSA analyzes relationships between a term and concepts contained in an unstructured collection of text. It is called Latent Semantic Analysis, because of its ability to correlate semantically related terms that are latent in a text. LSA produces a set of concepts, which is smaller in size than the original set, related to documents and terms [11, 12]. It uses SVD (Singular Value Decomposing) to identify pat- tern between the terms & concepts contained in the text, and find the relationships between documents. The method commonly referred to as concept searches. It has ability to extract the conceptual content of a body of text by establishing associations between those terms that occur in similar contexts. LSA is mostly used for page retrieval systems and text clustering purposes. LSA overcomes two of the most problematic keyword queries: multiple words that have similar meanings and words that have more than one meaning.

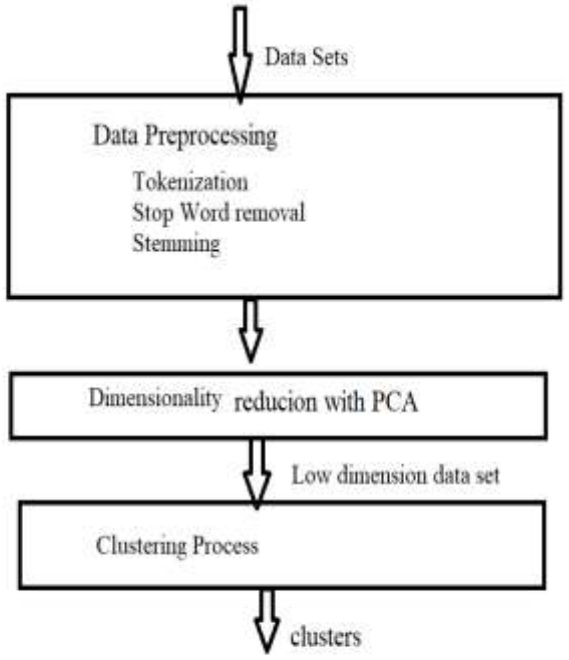

Fig.Proposed work for the problem

3.3 Procedure for the Dimensionality reduction problem

- Step1: For each document
  - Calculate Term Frequency Vector
  - Calculate Inverse Document Frequency
  - Calculate TF*IDF(Term Frequency * Inverse Document Frequency)vector
- Step2: Calculation of Covariance matrix
- Step3: Calculation of Eigen values
- Step4: sort eigen values in ascending order
- Step5: Apply Threshold for best dimensionality reduction tuning purpose select top eigen values and corresponding term vector.

IV. Implementation

4.1. Sample Problem of PCA

Document/ Term	Proliferation	Rumour	Technology	Regrettable
Hindu	1	0	1	0
Indian Express	0	1	1	0
Times	1	3	0	1

Document/Term	Proliferation	Rumour	Technology	Regrettable
Hindu	.005649	0	.005649	0
Indian Express	0	.00373134	.00373134	0
Times	.0037878	.011363636	0	.0037878

4.2 Inverse Document Frequency values

- Inverse Document Frequency of a term is defined as log-e[(Total no of documents)/(No of documents with term t in it)]
- Inverse Document Frequency of Proliferation is log(1.5)=0.176091259
- Inverse Document Frequency of rumour is log(1.5)=0.176091259
- Inverse Document Frequency of Technology is log(1.5)=0.176091259
- Inverse Document Frequency of regrettable is log(3)=0.4712125471

4.3 Inverse Document Frequency and TF*IDF

- TF*IDF Values

Document/Term	Proliferation	Rumour	Technology	Regrettable
Hindu	.005649	0	.005649	0
Indian Express	0	.00373134	.00373134	0
Times	.0037878	.011363636	0	.0037878

- Mean(technology)=.0005505953333
- Mean(proliferation)=.00055390666
- Mean(rumour)=.00088603096666
- Mean(regrettable)=.000602413

4.4 The Covariation Matrix of the sample is

[1.7130855E-7 - 4.5887925E-8 2.4850832E-8 - 5.31945E-8]
[- 4.5887925E-8 6.9357288E-7 - 3.4393699E-7 6.2152978E-7]
[2.4850832E-8 - 3.4393699E-7 1.7058157E-7 - 3.0836176E-7]
[- 5.31945E-8 6.2152978E-7 - 3.0836176E-7 5.5783618E-7]

4.5 Eigen values for the sample data set

The sorted order eigen values and corresponding eigen values are as follows:

- 1.4258680450E-6 -→ proliferation
- 1.67431134792208E-7 -→rumour
- 1.4052744158060863E-15 -→Technology
- 1.5974740338749847E-15 -→regrettable

A. If we consider the threshold as 50% then first 2 terms which are maximum and next to maximum.

B. Select the first 2 bigger values which corresponds to most significant terms in the data set. They are proliferation and rumour terms.

C. Apply Threshold for best dimensionality reduction tuning purpose.

V. Conclusion

We aim to investigate PCA feature reduction technique (PCA) 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 by other means is also aimed to investigate and compare the result with basic PCA technique. The results of feature extraction methods (PCA, LSA) are not dependent on number of features chosen. It is an advantage in text classification because choosing the correct number of features in the high dimensional space is a difficult problem. Moreover, 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 efficiency PCA (Principal Component Analysis) technique to be investigated on Indic Script documents and obtain a reduced data set.

REFERENCES

[1] B.Padmaja Rani, B. Vishnu Vardhan, A. Kanaka Durga, L.Pratap Reddy, A.. Vinay Babu , “Analysis of N-gram model on Telugu Document Classification”

[2] P.Vijayapal reddy, B. sasidhar, B.Harinathreddy, B. Vishnu Vardhan “Approaches of Dimensionality Reduction for Telugu Document Classification”

[3] Joan Farjo, Rawad Abou Assi, Wes Masri, and Fadi Zaraket” Does Principal Component Analysis Improve Cluster-Based Analysis?”

[4] N Tajunisha, V Saravanan, “ An increased performance of Clustering high dimensional data using Principal Component Analysis”

[5] K. Ramakrishna , B.Padmaja Rani, D. Subrahmanyam” Information Retrieval in Telugu Language Using Synset Relationships”

[6] Subhadra Mishra, Debahuti Mishra , Satyabrata Das and Amiya Kumar Rath “ Feature Reduction using Principal Component Analysis for Agricultural Data Set”

[7] G. Suresh Reddy “Dimensionality Reduction Approach for High Dimensional Text Documents”

[8] Daniel Kondor , Istvan Csabai, Laszlo Dobos, Janos Szule, Norbert Barankai, Tamas Ganyecz, Tams Sebok, Asofia Kallus and Gabor Vattay “ Using Robust PCA to estimate regional characteristics of language use from geo-tagged Twitter messages”

[9] D.A. Meedeniva, A.S.Perera “Evaluation of Partition-Based Text Clustering Techniques to Categorize indic Language Documents”

[10] A.K.Verma, Neema Verma” Robust Script Identification using Wavelet and EHD Features in PCA space”

[11] Rajini Jindal, Shweta Taneja “ Ranking in Multi Label Classification of Text Documents Using Quantifiers”

[12] Nitika Sharma, Kriti Saroha, “ A Novel Dimensionality Reduction Method for Cancer Dataset using PCA and Feature Ranking

[13] D. Lakshmi Padmaja, Dr.B. Vishnuvardhan “ Comparative Study of Feature Subset Selection Methods for Dimensionality Reduction on Scientific Data”

[14] P.Vijayapal Reddy, Dr. B. Vishnu Vardhan “ Corpus based Extractive Document Summarization for Indic Script.

[15] Aina Musdholifah, Siti Zaiton Mohd Hashim and Razali Ngah” Hybrid PCA-ILGC Clustering Approach for High Dimensional Data

[16] Shahana A. H. ,Preeja v “ Survey on Feature Subset Selection for High Dimensional Data

[17] Yang Y. Pedersen J. O.: A Comparative Study on Feature Selection in Text Categorization. Proc. ICML (1997) 412-420

[18] Berry M.W., Dumais S.T., and O'Brien G.W.: Using linear algebra for intelligent information retrieval. SIAM Review. 37(4) (1995) 573-595

[19] Bingham E., Mannila H.: Random Projection in Dimensionality Reduction: Applications to Image and Text Data. Proc. SIGKDD (2001) 245-250