International Journal of Scientific Research in Computer Science and Engineering
Vol.6, Issue.2, pp.15-19, April (2018)

Research Paper

E-ISSN: 2320-7639

Improving Clustering Accuracy using Feature Extraction Method

T.SenthilSelvi[1*], R.Parimala[2]

[1] Department of Computer Science, Periyar E.V.R College, Trichy-23, India
[2]Department of Computer Science, Periyar E.V.R College, Trichy-23, India

[*]*Corresponding Author: senthilselvikumar@yahoo.co.in*

Available online at: www.isroset.org

Received: 09/Mar/2018, Revised: 19/Mar/2018, Accepted: 31/Mar/2018, Online: 30/Apr/ 2018

Abstract— Clustering is the technique employed to group documents containing related information into clusters, which facilitates the allocation of relevant information. Clustering performance is mostly dependent on the text document features. The first challenge concerns difficulty with identifying significant term features to represent original content by considering the hidden knowledge. The second challenge is related to reducing data dimensionality without losing essential information. Clustering techniques were proposed to use feature extraction methods Principal Component Analysis (PCA) and Kernel Principal Component Analysis (KPCA) to improve the clustering efficiency and quality. Documents are pre-processed, converted to vector space model and then clustered using the proposed algorithm. The goal of this work is to design a suitable model for clustering text document that is capable of improving clustering performance. In this paper, the problems are discussed with empirical evidence. Experimental results show that the proposed method is effective for the text clustering task.

***Keywords*—** Clustering; Euclidean Distance; Document frequency; Dimensionality reduction; Principal components

I. INTRODUCTION

Document clustering is an active research area for exploring and grouping set of data objects into multiple groups or clusters so that objects within the cluster have high similarity, but are very dissimilar to objects in the other clusters. The task of text clustering is a group which exhibits resemblance of similar documents together. The user wants to investigate and explore the data to find some induced structures in them. Clustering is a technology for finding such structures. Although a text document consists of a sequence of sentences each sentence consists of sequence of words. A document is usually considered as a "bag" of words in document clustering. Each clustering algorithms has its own merits and demerits, popular algorithm for clustering results in losing potential clusters and/or constructing less important ones. Due to the innate nature of sparsity in high dimensional data, majority of clustering algorithm may not perform well. A dimensionality reduction or feature extraction method extracts new features by combining the existing features. Firstly, PCA and KPCA are introduced to reduce the dimensionality of the original high dimensional data and then k-means clustering algorithm is adopted to make clustering analysis on the dimension of the reduced data. The so-called dimension reduction refers to the process in which samples from the high dimensional space are mapped to the low-dimensional space for meaningful representation of the high dimensional data by linear or non-linear method.

The paper is organized as follows: Section I introduces clustering. Section II outlines the Literature Review. Section III presents the feature selection and feature extraction method. Section IV presents the Proposed Methodology. Section V details about the various Corpus used for study. Section VI outlines the used Environment and Libraries. Section VII discusses on the Experimental Results and finally Section VIII concludes with future scope.

II. RELATED WORK

Dimensionality reduction provides an extensive challenge to *k*-means clustering. The time complexity of K-means clustering depends on the number of features. Feature Selection and Feature Extraction method are employed to find relevant features. C.Boutsidis, M.W. Mahoney and P. Drineas in their research computed probabilities using probability distribution for the feature space. From these computed probabilities a small number of features were selected for clustering. On evaluation of the two datasets (NIPS and Bio), the most relevant features were selected and that the clustering obtained after feature selection is found to be accurate for constant k and different size of the dataset[1]. Diagonal dominance is a phenomenon proposed by D. Greene and P. Cunningham, an observable fact which occurs when kernel functions are applied to sparse high-dimensional

data, such as text corpora like BBC News Abstract, BBC Sports, Classic3, Classic, NG17-20, NG3 and Reviews. The diagonal shift (DS), Diagonal shift with empirical Kernel Map (DSM), Algorithm adjustment (AA) and Sub polynomial Kernel with Empirical Kernel map (SPM) represent efficient strategies for reducing diagonal dominance and are found to be the best choice for implementation in large dataset [2]. Z. Miner and L.Csat proposed feature selection for the text document categorization problem were a kernel matrix was built and kernel PCA based clustering was used to group words into clusters and these clusters were used subsequently. The best results were achieved with different kernels like linear, RBF and polynomial kernel. These kernels achieved the highest performance value [3]. R. Mall and J.A.K. Suykens proposed the kernel spectral document clustering (KSDC) model which generates homogeneous clusters of documents and K-means clustering generates heterogeneous clusters also evaluates the cluster quality by a quality metric. They compared the quality of the clusters obtained by the proposed KSDC technique with k-means and neural gas algorithm on several world textual data like NIPS, Kos, R8, Classic4, Reuters, 20NewsGroups, Enron, NY Times and found that KSDC outperformed well for small data and also prevents formation of high-dimensional data. A major advantage of this algorithm is to identify homogeneous and heterogeneous clusters [4]. R. Jenssen, T.Eltoft, M.Girolami and D. Erdogmus proposed kernel technique called kernel MaxEnt where maximum entropy with eigen vectors is calculated. This new kernel-based data transformation technique named kernel MaxEnt, is based on Renyi's quadratic entropy estimated via Parzen windowing. The data transformation is obtained using eigen vector. An enhanced spectral clustering algorithm was proposed by the authors were kernel PCA is replaced by kernel MaxEnt. This small adjustment enhanced greater performance of the algorithm [5]. T. Shi, M. Belkin and B. Yu proposed and explained spectral clustering, where eigen vectors can be used for clustering when the distribution is a mixture of multiple components. Among the several kernel functions available Gaussian kernel and Polynomial kernels are commonly used. Gaussian kernel is more flexible and comparatively effective to Polynomial kernel when a careful choice of degree is chosen [6]. L. Kaufmann discussed the application of polynomial kernels to handwritten digits recognition and checkerboard problem [7]. Researchers have recently studied the design of digital filters and regression frameworks through kernel methods. Very limited research has been conducted on an unsupervised dimensionality technique for text document clustering. This motivated our research to bring out a Three-Phase framework for unsupervised datasets.

III. FEATURE SELECTION AND FEATURE EXTRACTION METHODS

Feature Selection Methods

Document frequency is the number of documents in which a term occurs. It is the simplest criterion for term selection and easily scales to a large dataset with linear computation complexity. It is simple but is an effective feature selection method for text categorization [8].

Feature Extraction Methods

Standard PCA works on the linear space, whereas kernel PCA works on the non-linear feature space using kernel functions. A kernel matrix can be constructed by applying the kernel function for every pair of data objects. Kernel PCA computes the required number of top eigen values and eigen vectors of the kernel matrix and using this produces the low-dimensional representation of data objects.

a) Principal Component Analysis (PCA)

PCA is an unsupervised feature extraction method that selects the original data projection with the maximum feature covariance. The desired goal is to reduce the dimensions of a d-dimensional dataset by projecting it into a p-dimensional subspace where p<d in order to increase the computational efficiency while retaining most of the information.

b) Kernel Principal Component Analysis (KPCA)

The majority of the PCA approaches proposed to assume linear data dependencies. But in many applications, non-linear data dependencies naturally arise. The non-linear nature of various phenomena constituted the extension of real PCA and KPCA one of the most popular methodologies for non-linear dimensionality reduction and feature extraction. The use of kernel functions provides a powerful and principled way of detecting non-linear relations using well-understood linear algorithms in an appropriate feature space. Kernel maps the data into some other dot product space F called the feature space via a non-linear map $\phi : \mathrm{R}^{N} \to F$ and perform linear algorithm in F. Kernel based learning methods uses an implicit mapping of the input data into a high dimensional feature space defined by the kernel function. The kernel function is defined as the inner product between two points in a suitable feature space in high dimensional space. Kernels commonly used with kernel methods are listed in Table 1.

Table 1: Kernel Functions and Parameters

Kernel Name	Kernel Function	Parameters
Rbf (R)	$k(x,x') = \exp(-\sigma\left\|x-x'\right\|^2)$	$\sigma > 0$
Polynomial (P)	$k(x,x') = (scale.\langle x,x'\rangle + offset)^{\text{degree}}$	offset >0, degree ≥2
Tangent (Sigmoid) (T)	$k(x,x') = \tanh(scale.\langle x,x'\rangle + \deg ree)$	offset ≥ 0
Bessel (B)	$k(x,x') = \frac{bessel^{n}_{v+1}(\sigma\left\|x-x'\right\|)}{(\left\|x-x'\right\|)^{-n(v+1)}}$	
Laplace (L)	$k(x,x') = \exp(-\sigma\left\|x-x'\right\|)$	$\sigma > 0$

IV. METHODOLOGY

The proposed work presents a Three-Phase framework for text corpora dimensionality reduction and clustering. The Methodology is illustrated in Figure 1.

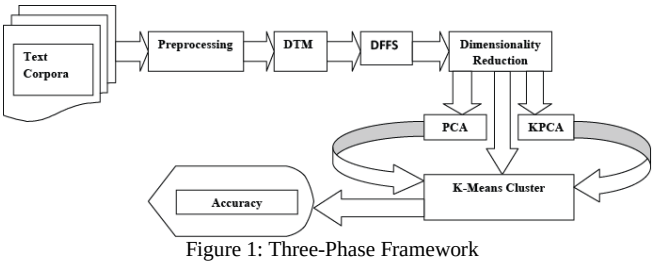


Figure 1: Three-Phase Framework

The framework consists of the following major computational states: Preprocessing DTM, Dimensionality Reduction and clustering. Text corpora is the compilation of various forms of texts usually stored digitally where the text analysis is performed. A text document comprises of several sentences and each sentence consists of a sequence of words. A text document is usually considered as a "bag" of words/terms in document clustering. Preprocessing includes tokenization converting the text to lower case, removing numbers and punctuations, removing stop words and finally perform stemming. Document term matrix (DTM) size consists of documents row (m) and column (n) terms. The Document term matrix is constructed using TF-IDF (Term Frequency-Inverse Document Frequency) weighting score. The TF-IDF weighting scheme in normalized form is given in the interval [0, 1]. In Phase-I the formal approach for dimensionality reduction is by first applying DFFS on DTM. The relevant features are selected for clustering using document Frequency Feature Selection (DFFS). DFFS selects the features whose frequency in the documents is greater than the threshold (T). The resultant size of DTM is m x d where d < n.

The feature extraction is performed on the selected feature set. PCA and KPCA are applied to compute the principal components. In Phase-II PCA is a method which involves finding the linear combination of a set of variables that has maximum variance in a low dimensional data. In Phase-III the main strengths of kernel PCA is that it can map non-linearly separable data to a high dimensional feature space where the data becomes linearly separable data by deriving low-dimensional features that incorporate the higher order statistics. PCA is good at reducing dimensionality for linearly separable data, while kernel PCA can handle non-linearly separable data. The kernel Matrix is first computed for the DTM and the principal components of DTM are passed into the k-means clustering algorithm to group the text corpora and finally the accuracy is displayed. Pseudo code for proposed framework is stated as:

Cluster Formation – Phase –I
Input : Desired Number of clusters k, threshold T
Output: Cluster Accuracy
Algorithm

1. Collect text corpus for clustering
2. Preprocess the collected Text corpus
3. Calculate tf-idf score
4. Construct Document-Term-Matrix.
5. Perform DFFS for Threshold T
6. Perform K-Means Clustering
7. Validate cluster accuracy for C_i, i=1,2, …….k

Cluster Formation – Phase – II
Input : Desired Number of clusters k, threshold T, number of Principal Components p
Output: Cluster Accuracy
Algorithm

1. Implement the steps 1 to 5 in Phase –I algorithm
2. Subtract the mean
3. Calculate the covariance matrix
4. Calculate the eigenvectors and eigen values of the covariance matrix
5. Sort eigen values in descending order
6. Choose the p eigenvectors (principal components) that corresponds to the p largest eigen values
7. Construct the projection matrix w from the selected p principal components
8. Find the reduced projected DTM
9. Go to step 6 of Phase-I algorithm

Cluster Formation – Phase – III
Input : Desired Number of clusters k, threshold T, Kernel function, number of Principal Components p
Output: Cluster Accuracy
Algorithm

1. Implement the steps 1 to 5 in Phase-I Algorithm
2. Compute the kernel matrix of DTM for a particular function
3. Implement the steps 2 to 9 in Phase-II Algorithm

K-Means Clustering
// Apply k-means clustering with reduced DTM

1. Choose randomly k input vectors (centroid) to initialize the clusters
2. For each input vector, find the cluster center that is closest using Euclidean distance and assign that input vector to the corresponding cluster
3. Update the cluster centers in each cluster using the mean (centroid) of the input vectors assigned to that cluster
4. Repeat steps 2 and 3 until no more change in the mean values

V. CORPUS USED

Past research work [8] reported text categorization performance varies greatly on different datasets. Different text corpora are selected to evaluate the text clustering performance for a proposed framework. The first set BBC News abstract an English dataset corresponds to stories in five topical areas. The second set, the BBC Sports website consists of five different documents such as athletics, cricket, football, rugby and tennis. The third set 20Newsgroup corpuses consists of 20 different news retrieved from the website. It is divided into four categories fromNg20-group1 to Ng20-group4. The Fourth set the Reuters 21578 collection are originally taken from Reuter's newswire of 1987. The documents are broadly divided into five categories (Exchanges, people, topics, organizations and places). The fifth dataset c50 corpus is used to determine the author's identification where the top 50 authors were selected and stored in this corpus. The sixth data set the Enron email (Enron1 - Enron6) contains approximately 500,000 emails generated by employees of the Enron Corporation and finally the seventh data set the Ohsumed Corpus a subset of the MEDLINE database were all taken for study.

VI. USED ENVIRONMENT AND LIBRARIES

R is a programming language and software environment used for statistical computing. The package "tm" [9] is used to perform all manipulations regarding text. "kernlab" is a package used for creating and manipulating kernel-based methods and algorithms [10].

VII. RESULTS AND DISCUSSION

This study implemented the three algorithms namely Phase-I, Phase-II and Phase-II using R software and evaluated the results for k=2, 5, 10, 15 and 50 for different datasets with threshold value T>=30 and principal component p=2. Results show that proposed approaches using Phase-II algorithm perform better than Phase-I approach and also Phase-III approach outperforms the approach of Phase-II also. Measurement of cluster validation is one of the steps that need to be done after the clustering analysis. The validity of the cluster is done by evaluating clustering algorithms so that it can be seen that formed cluster m matches the natural partition. The study uses an accuracy validation. Accuracy refers to the clustering of data about a known target. Table 2 shows the cluster results for Phase-I algorithm. Table 3 show the clustering result for Phase – II and Phase – III algorithm. A sample plot of clustering using DFFS, DFFS + PCA, DFFS + KPCA is depicted in Figure 2a, 2b and 2c respectively.

The results obtained from Phase-I of Table 2 shows that our accuracy is very less. To improve the accuracy we proposed and implemented the algorithm as given in Phase-II and Phase-III and found that the overall accuracy improved from that of the existing accuracy. The accuracy of the proposed work is compared with R.Mall et.al. for BBC Sports, BBC News Abstract and Newsgroup20 and the work of Dereck Greene for Reuters and Enron with our proposed work.

Table 2: Cluster Result of Phase - I

Corpora Type	Text Corpus	No.of Docs	Features	Cluster size k	DFFS	K-Means Clustering Accuracy % (Phase –I)
News wires	BBC Sports	737	13057	5	2	91.3
	BBC News Abstract	2225	29069	5	157	03.8
	Ng20-group1	5000	57467	5	573	01.2
	Ng20-group2	5000	56459	5	479	01.2
	Ng20-group3	5000	10020	5	479	01.6
	Ng20-group4	5000	75731	5	554	01.7
	Reuters	7316	23529	15	335	07.2
	c50	2500	28785	50	131	05.4
Spam	Enron1	5172	49766	2	354	03.0
	Enron2	5857	39542	2	433	55.8
	Enron3	5512	53005	2	393	00.9
	Enron4	6000	68654	2	362	00.8
	Enron5	5175	41496	2	380	00.5
	Enron6	6000	70413	2	473	00.6
	Lingspam	2894	57057	10	225	01.0
Medical Abstract	Ohsumed (5 cases)	2894	42266	5	225	00.2

Table 3: Cluster result of Phase II and Phase III

Text Corpus	K-Means Clustering Accuracy % (Phase–II)	K-Means Clustering Accuracy % (Phase –III) Kernel Functions					Existing Accuracy %
		R	P	T	B	L	
BBC Sports	94.4	93.0	84.3	92.8	90.0	96.8	92.0[4]
BBC News Abstract	90.5	88.4	79.0	85.8	79.4	93.4	80.0[4]
Ng20-group1	90.9	85.4	88.0	77.5	85.5	82.4	63.0[4]
Ng20-group2	91.8	83.6	84.3	79.2	93.8	84.2	
Ng20-group3	90.8	90.7	94.9	84.6	86.3	81.0	99.0[4]
Ng20-group4	86.9	94.3	93.0	82.7	86.1	84.3	
Reuters	98.1	99.8	99.9	97.8	98.3	98.8	98.7[2]
c50	98.1	99.0	92.6	99.2	97.6	98.6	
Enron1	46.1	79.4	42.8	47.6	31.2	24.3	89.8[2]
Enron2	43.0	04.0	46.1	42.5	13.8	70.0	
Enron3	43.9	90.3	45.5	43.3	15.6	09.8	
Enron4	46.7	84.1	46.2	44.7	20.0	16.5	
Enron5	48.7	18.1	49.4	47.9	28.6	68.6	
Enron6	44.0	91.5	46.3	41.8	18.8	18.6	
Lingspam	98.5	97.9	99.2	94.5	97.7	95.6	
Ohsumed (5 cases)	86.9	91.2	85.5	84.2	85.5	86.8	

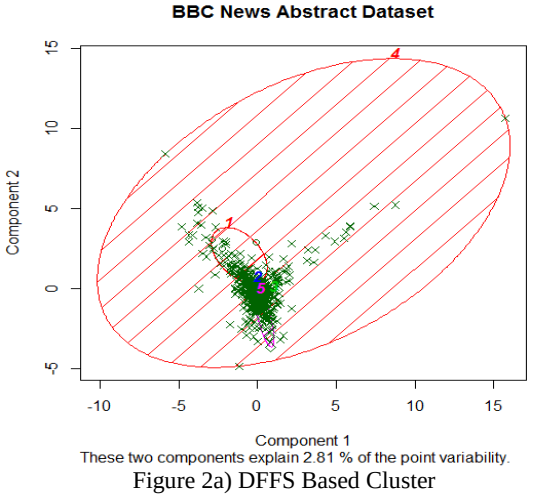
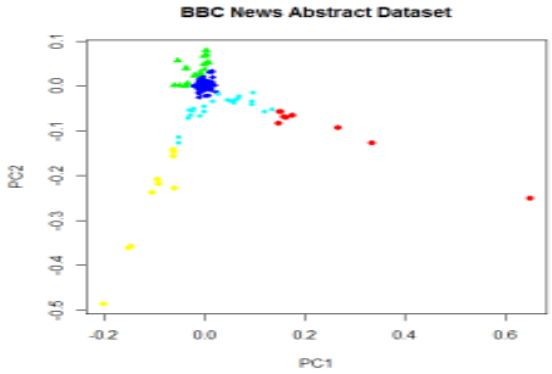


Figure 2a) DFFS Based Cluster



.

Figure: 2b) PCA based cluster

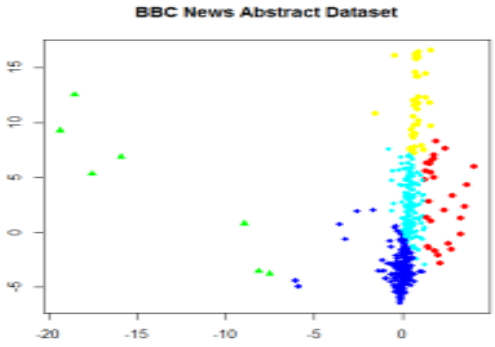


Figure 2c) KPCA based cluster

VIII. CONCLUSION AND FUTURE SCOPE

The algorithms are implemented using R and experiments are conducted and the results are evaluated based on performance parameters. The paper introduced a new algorithm DFFS, DFFS + PCA and DFFS + KPCA to reduce dimensionality and enhance accuracy for various text datasets. The clustering accuracy was improved using PCA and KPCA. A future work in this direction would be the application of proposed algorithm on the unlabeled text corpus.

ACKNOWLEDGEMENT

We would thank all souls who had helped us to make this paper a successful one and giving several openings in research. We also thank the R Core Team for providing this open source software towards successful implementation of this research work.

REFERENCES

1] C. Boutsidis, M.W. Mahoney, P. Drineas,"Unsupervised Feature Selection for the k-Means clustering problem", In the NIPS'09 Proceedings of the 22nd International Conference on Neural Information Processing Systems, Canada, pp.153-161, 2009.

[2] D. Greene, P. Cunningham, "Practical Solutions to the Problem of Diagonal Dominance in Kernel Document Clustering", In the 23rd International Conference on Machine Learning, Pittsburgh, PA, pp.377-384, 2006.

[3] Z. Miner, L. Csat, "Kernel PCA Based Clustering for Inducing Features in text Categorization", In the ESANN'2007 Proceedings-European Symposium on Artificial Neural Networks, Bruges,Belguim., pp.349-354, 2007.

[4] R. Mall, J.A.K.Suykens,"Kernel Spectral Document Clustering Using Unsupervised Precision-Recall Metrics.", 2015 International Joint Conference on Neural Network,Killarney, Ireland, pp. 1-7, 2015.

[5] R. Jenssen, T.Eltoft, M.Girolami and D. Erdogmus ,"Kernel Maximum Entropy Data Transformation and an Enhanced Spectral Clustering Algorithm.", In the NIPS'06 Proceedings of the 19th International Conference on Neural Information Processing Systems, Canada, pp.633-640, 2006.

[6] T. Shi, M. Belkin, B. Yu, "Data spectroscopy: eigenspaces of convolution operators and clustering", The Annals of Statistics, Vol. 37, No.6B, pp.3960-3984, 2009.

[7] L.Kaufmann, "Advances in Kernel Methods — Support Vector Learning -Solving the quadratic programming problem arising in support vector classification, MIT Press, Cambridge, MA, pp.147–168, 1999.

[8] Y.Yang, J.O. Pedersen, "A Comparative study of feature selection in Text Categorization", In the Proceedings of the Fourteenth International Conference on Machine Learning (ICML'97), USA, pp.412-420, 1997.

[9] I. Feinerer, K.Hornik, D. Meyer, "Text Mining Infrastructure in R", Journal of Statistical Software, Vol.25, Issue 5, pp.1-54, 2008..

[10] A. Karatzoglou, A. Smola, K. Hornik, A. Zeileis, "kernlab - An S4 Package for Kernel Methods in R", Journal of Statistical Software Vol.11, Issue 9, pp.1-20, 2004.

Authors Profile

Mrs. T. SenthilSelvi graduated Master of Science and M.Phil in Computer Science from Srimathi Indira Gandhi College, Trichy affiliated to Bharathidasan University and now is a Reseach scholar and currently working as Assistant Professor in Periyar E.V.R. College, Tiruchirappalli. She has a teaching experience of about 21 years. Her research interest is in the field of Web Mining, Artificial Intelligence and Information Retrieval.

Dr. R. Parimala graduated with M.Sc., Applied Science at the National Institute of Technology, (formerly Regional Engineering College) Tiruchirappalli in 1990. She received her M.Phil., Computer Science at Mother Teresa University, Kodaikanal in 1999. She started teaching in 1999 at National Institute of Technology and is currently working as Assistant Professor in Department of Computer Science, Periyar E.V.R. College (Autonomous), Tiruchirappalli. She completed her Ph.D. at National Institute of Technology, Tiruchirappalli. Her area of research interests include Neural Networks, Data mining and Optimization Techniques

