International Journal of Engineering & Technology, 7 (3.29) (2018) 648-654

International Journal of Engineering & Technology

Website: www.sciencepubco.com/index.php/IJET

Research paper

Unified Framework of Dimensionality Reduction and Text Categorisation

K.M.M Rajashekharaiah[1], Sunil S Chikkalli[2], Prateek K Kumbar[3], Dr. P. Suryanarayana Babu[4]

[1] Associate Professor, [2,3] Final year students, [4]Research Supervisor
[1] School of Computer Science and Engineering, KLE Technological University, Hubli, Karnataka and Research Scholar,
[2,3] Computer science and Engineering, Rayalaseema University, Kurnool, AP, India.
[4]Research Supervisor, Dept of Computer Science, Rayalamaseema University, Kurnool, AP, India.
**Corresponding author E-mail: kmmr@bvb.edu*

Abstract

Text classification (categorization) is a supervised learning task that assigns text documents to pre-defined classes of documents. It is used to organize and manage the collection of text documents available in digital form. To accomplish the task, support vector machine (SVM) is regarded as the suitable classifier for any kind of applications. Though SVM's computational complexity is independent of number of dimensions, still high dimensionality poses the problem of 'curse of dimensionality' that can be solved effectively by the process of Dimension Reduction (DR). This work contemplates on developing a framework for dimensionality reduction and text classification. A comparative analysis of the classification accuracies using two approaches viz., text classification with dimensionality reduction and text classification without dimensionality reduction completes the scope of the paper. It also evaluates the efficiency of various dimensionality reduction techniques to include one of the most coherent methods in the framework.

Keywords:*Classification accuracy, Classifier, Dimension Reduction, Framework, Supervised learning, Support Vector Machine(SVM), Text Classification/Categorisation (TC) ;*

I. Introduction

During the recent years, data in digital form has been growing exponentially in size. And according to International Data Corporation (IDC), global digital universe is expected to reach 44 zettabytes by 2020. The prime reason is the increased popularity of applications that deal with the data. Thus, managing and organizing the data becomes necessary. Text categorisation is one such automated process that can achieve this goal. It is a machine learning task that categorises text documents on the basis of class labels or genres. A major characteristic of the TC problem is the high dimensionality of textual data. The solution to such a problem is the usage of some form of Dimensionality Reduction (DR). DR reduces the number of dimensions (features), noise and sparseness by preserving the significant variance of the original dataset. Thereby it betters the classification performance in terms of speed and accuracy.

This paper focuses on including the most coherent dimensionality reduction in the unified framework considering the accuracy rate, computation cost and space cost. Firstly, we compare the performances of classifiers like single decision tree (R-part), Random forest and SVM on the Email dataset. Then we show that, applying feature extraction DR before the actual classification can significantly improve accuracy compared to classification involving no dimensionality reduction. Lastly, we familiarise the generic as well as the application specific enhancing measures that can increase the accuracy of classification result. The next section deals with the need of dimensionality reduction and its scope. Section III outlines the related work in the field of text mining. In section IV we explain the methodology of text classification involving text pre-processing and dimensionality reduction. In section V we present the results obtained from the implementation. In the final section we conclude about the most coherent DR algorithm that is to be included in the unified framework.

2. Background

Dimensionality reduction is the transformation of a high dimensional data into a low dimensional data space, while retaining most of the useful structure in the original data thereby circumventing the 'Curse of Dimensionality'.[1] It is the sparseness and noise that degrade the classification accuracy and being referred to as Curse of Dimensionality. Dimensionality reduction techniques can be of two forms, Feature Selection and Feature Extraction. Feature selection, also known as variable selection or subset selection is the selection of a subset of features or elimination of less significant features that is most suitable for the task at hand. Though feature selection gives better lower dimensional

Copyright © 2018 Authors. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

representation, even for moderate value of N (features), 2N candidate subsets are possible for the dataset. Thus we focus on feature extraction, where a small set of new features is constructed by a general mapping from the high dimensional data. Benefits of using DR is that it reduces the computation time and also the resulting classifiers take less space to store. The surprising thing about DR is it enhances the accuracy of task at hand (classification here).

It finds its place in wide variety of applications like Text mining, Image retrieval, Microarray data analysis, Protein classification and Intrusion detection. Text Classification is one such automated process in the field of text mining which is highly in want of the pre-processing task, DR.

3. Related Work

DR has wide range of applications i.e. it is applied to data sources such as text, images, biological data etc. Of these, some of the popular works on text are:

Yiming Yang and Jan O. Pedersen[2] presented a comparative study of five feature selection methods in statistical learning of text categorisation. Of the five feature selection methods viz. document frequency(DF), information gain(IG), chi-square text(CHI), mutual information(MI) and term strength(TS) applied on the Reuters corpus with K-nearest neighbour and Linear Least Square Fit (LLSF) as classifiers, they found strong correlations between the DF, CHI and IG values of a term. But CHI and IG measures involved expensive computations. Hence DF was inferred as the suitable feature selection method for text categorisation.

Hyunsoo Kim, Peg Howland and Haesun Park[3] described class of cluster preserving DR techniques which area able to dramatically reduce the number of features needed to perform text classification without sacrificing the classification accuracy. They compared these algorithms with LSI/SVD (Latent Semantic Indexing/Singular Value Decomposition) and found that the latter is not able to preserve the cluster structure as well as achieve the accuracy rate as the former algorithms do. However these techniques can be applied only if the terms in the data set are already clustered into logical groups.

Bingham and Mannila[4] compared statistically optimal methods of DR with Random Projection. They assert that random projection is an alternative to traditional and optimal DR methods as it is computationally simple and tend to preserve the similarity of data vectors to a high degree. They also find that it is able to achieve comparable performance. However they evaluate distortion error rather than the classification accuracy.

Underhill and McDowell [5] examined the five DR techniques on the accuracy of two supervised classifiers (Linear and KNN classifiers) on three distinct corpuses and concluded that MDS and ISOMAP outperformed all other unsupervised algorithms, Principal Component Analysis (PCA), Lafon's Diffusion Map(LDM), Local Linear Embedding(LLE). They also infer that of the five DR techniques they had taken into consideration, all were able to achieve substantial improvements compared to not performing the DR.

IV. Methodology

To effectuate the unified framework, in the first approach we start with a collection of emails, encode the emails (documents) into a suitable representation and then classify each document using classifiers. Meanwhile the other approach we follow is applying dimensionality reduction method on the encoded representation and then classifying the emails using the efficient classifier. Finally, we evaluate and compare the results of both the approaches. Below we elaborate each of these steps.

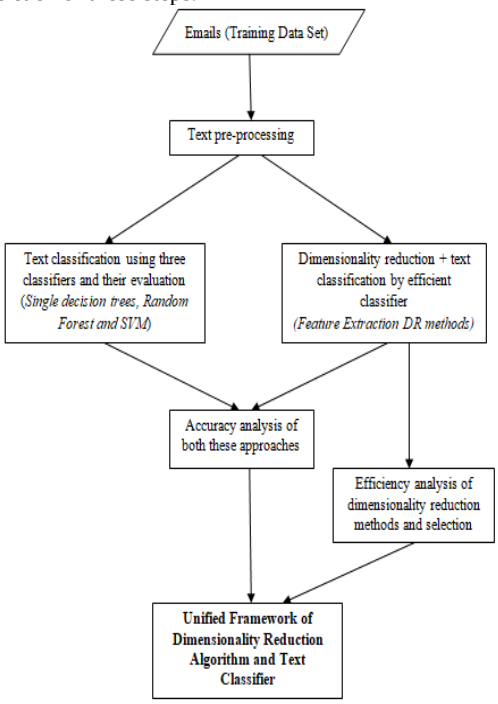


Fig. 1. Unified Framework Methodology

A. Data Set Collection and Partition

The dataset we employ for the implementation of unified framework and comparison of DR methods is the email dataset which is a collection of 5572 emails. Out of 5572 emails, it is known in advance that 4825 are hams and 747 are spams i.e. 86.59% of the emails are relevant and other 13.41% are irrelevant. Further this dataset is split into training data(70%) and test data(30%) keeping the proportion of ham and spam emails same in both the splits. So our training dataset contains 3901 Emails. Initially all 3901 emails (observations) are stored in a column matrix. Thereafter text pre-processing is accomplished which involves text cleaning and text encoding.

B. Text pre-processing

The matrix containing a single column and 3901 rows is first tokenised to avail each words (terms) separately in each column. Further the data matrix is cleaned by

i. Removal of numbers, punctuations and symbols.
ii. Lowercasing of each character to resolve case ambiguity.
iii. Removal of stopwords.
iv. Performing stemming

The stopwords like pronouns, prepositions, conjunctions carry no information and hence have to be eliminated. Word stemming is the

process of suffix removal to generate word stems. This is done to group words that share the same origin form like run, ran, runs, running all are from the origin word run. Then the cleaned data matrix is represented in the vector space model. In the vector space model, documents are represented by vectors of words. The document set is represented by a word-by-document matrix A i.e.

$$\mathbf{A=f_{id}}$$

where $\mathbf{f_{id}}$ is the weight of word **i** in document **d**. Matrix is normally sparse as every word does not appear in every document. The sparseness creates a lot of issues that can only be prevented by the usage of DR. Further the term-document matrix is rationalised by determining the weight $\mathbf{a_{id}}$ of word **i** in document **d**. A well known approach for computing word weights is the tf-idf weighting which assigns the weight $\mathbf{a_{id}}$ to word **i** in document **d** in proportion to the number of occurrences of the word in the document and in inverse proportion to the number of documents in the collection for which the word occurs at least once.

$$a_{id} = f_{id} * \log \frac{N}{n_i} \quad (1)$$

Here $\mathbf{f_{id}}$ is the term frequency(tf) discussed earlier and $\mathbf{\log(N/n_i)}$ is the inverse document frequency (idf). **N** is the total number of documents in the corpus and $\mathbf{n_i}$ is the number of documents that contain the term **i**. In other words, tf-idf weighting scheme gives higher importance to rare/unique terms compared to common ones.

C. Dimensionality Reduction

Data represented in a high dimensional space is very difficult to train or mine for the purpose of knowledge discovery. So it is necessary to reduce the dimensions dramatically. As feature selection is NP-hard, we only choose extraction algorithms. Depending upon the prior knowledge of class labels of samples we classify feature extraction methods into supervised and unsupervised extraction methods. Supervised methods not only maintain the variances across features, but also seek to preserve the relationships between features and labels. Over it supervised algorithms classify more precisely than unsupervised algorithms (LSA being comparable to Supervised FE methods). If there's no information of class labels then unsupervised algorithms like Principal Component Analysis (PCA), Latent Semantic Analysis (LSA), Independent Component Analysis(ICA), Multidimensional Scaling (MDS) and ISOMAP have to be employed for reduction.

Further supervised algorithms can be divided into two classes, cluster preserving algorithms and the rest (cluster non-preserving algorithms). Cluster preserving algorithms like Centroid Algorithm, Orthogonal Centroid Algorithm and Linear Discriminant Analysis/Generalised SVD perform tremendously as inferred by Kim[3]. But any of the three algorithms can be applied only if the dataset is clustered. If the dataset is non-clustered, then the whole dimension reduction task involves two steps:
i. Clustering data using K-means, KNN or K-mediods
ii. Reducing by any of the cluster preserving algorithms

Though for any data it gives the best lower dimensional representation than any other class of algorithms it involves two processes (clustering and reducing). Thus when combined the computation becomes more intensive than the algorithms we discuss here. Therefore we assess only the class of unsupervised algorithms and cluster non preserving supervised algorithms. Following are the Algorithms:

Unsupervised Algorithms: PCA, LSA, MDS, ISOMAP. Cluster non preserving supervised algorithms: LDA, Supervised ISOMAP

Among these above algorithms, only ISOMAP and its supervised version S-ISOMAP are non linear algorithms i.e. they depend on geodesic distance rather than the Euclidean distance. The simple idea behind geodesic distance is that it tends to preserve the local structure of the data points. Here we describe some of the coherent DR techniques belonging to the two classes (Class of unsupervised algorithms and cluster non preserving supervised algorithms):

i. Principal Components Analysis (PCA) – PCA results in new dimensions that capture the degree of variation in the original data. These smaller numbers of dimensions are called principal components. The computation involved in PCA is quite simple. Firstly singular value decomposition (SVD) is performed on the document covariance matrix and thereafter the resulting eigenvectors are multiplied with their corresponding eigen values. And the principal components must be chosen in such a way that they contribute to 85% - 95% variance of data. However the issue with PCA is its inability to scale up for very high dimensional data i.e. it can be effective only when reducing tens or a few hundreds of dimensions.

ii. Independent Components Analysis (ICA) - It is based on the idea that each sample of data is a combination of components that are statistically independent of each other. The independence of features is acquired by minimising mutual information and maximising non-Gaussianity. In other words it tries to de-correlate the data by separating each source(independent component) of signal.

iii. Latent Semantic Analysis(LSA) – Here the high dimensional space is reduced into lower-dimensional data space where the original features are combined together to form higher concepts. The idea behind the combination is the co-occurrence of the terms. These co-occurrence patterns help in combining the similar terms & context oriented terms. Singular Value Decomposition (SVD) is used to map the original vectors into new vectors. However a truncated form of SVD where reduction is done rigidly is enough to preserve the information in original data. For a few thousands of dimensions in original data space, the truncated SVD results in approximately 300 extracted dimensions which are not interpretable.

iv. Linear Discriminant Analysis (LDA): It is a supervised DR algorithm that works on the basis of class label information. It tries to generate features that are distinct from each other i.e. it computes the within-scatter matrix (Sw) of the data, a between-scatter matrix(SB) and solve the eigen value problem, (Sw)-1 SB. There after those dimensions are extracted that define the maximum variance of the data.

v. Multidimensional Scaling (MDS): MDS focuses on preserving distances between pairs of points [9]. It is a linear and unsupervised algorithm like PCA and LSA. The eigenvalue decomposition is performed on the matrix of pairwise distances between the original points. Even in this algorithm only those dimensions are extract that correspond to eigen vectors of high eigen values.

vi. Isomap (Non-Linear) - Isomap works as same as Multidimensional Scaling (MDS) except it preserves the non-linear information. That is its matrix is based on the geodesic distance, which is computed by connecting points in a nearest-neighbours graph. On the other hand the linear DR methods are based on the Euclidean distance. The geodesic distance though a costlier

computation can capture intrinsic geometry of the data more precisely for advanced applications. The remaining computation task is done the same way as in MDS.

D. Classifiers and Evaluation

In data mining and machine learning, classification is a supervised learning task in which the algorithm learns from the input data and class labels given to it and later uses this learning to classify new samples or observations. The input data set may be binary class (like relevant or irrelevant, ham or spam) or it may be multi-class. Document categorisation, image recognition, speech recognition, bio metric identification, hyperspectral image classification are some of the trending applications of this supervised learning process. Below we describe the three classifiers that we consider for comparison:

i. Decision Tree (R-part): It is a predictive modelling approach that builds classifier in the form of a tree structure. The leaves (leaf nodes) in the tree are the class labels and branches (internal nodes) are the features that determine the class label of a sample. The Information Gain (IG) decides the significant features upon which the class label of a sample has to be predicted. Though the classifier is close to human decision making, it isn't as powerful as other classifiers. Over it, any change in the data affects the tree considerably.

ii. Random Forest Classifiers: Random forests or random decision forests are known as ensemble learning method as it follows a divide and conquer approach to categorise the data. Multiple decision trees are built for the samples each resulting in the decision of the category (class label). The final prediction is made on the basis of highest number of votes. Higher the number of trees in the forest higher is the classification accuracy as it betters off decision trees due to the capability of solving over fitting problem. Meanwhile it takes more time for the prediction process.

iii. Support Vector Machine Classifier: The primary objective of SVM is to find the optimal boundaries between the different classes of the dataset. It is both a linear and non linear classifier. That is kernels are used when handling the non-linearly separable data. It performs very well in high dimensional space as the number of support vectors determines the complexity of the dataset not the number of dimensions. And the number of support vectors is independent of the dimensions. Though the time it takes in training and computation costs are high, it proves to be an all-round performing classifier for any type of data. Other classifiers like kNN and Neural Networks cannot cope with high dimensionality as SVM does. Additionally, SVM can work even when the data consist of outliers or noise.

The table below details the classification decisions predicted with reference to the actual decisions. It is called as the confusion matrix.

Table I. Confusion Matrix(Binary Classification)

Predicted	Actual		
		True	False
	True	TP	FP
	False	FN	TN

TP - Number of documents correctly assigned to the category.
FP - Number of documents incorrectly assigned to the category.
FN - Number of documents incorrectly rejected from the category.
TN - Number of documents correctly rejected from the category.

The accuracy of a classification is measured using these details of the above confusion matrix. It is the total number of correctly predicted decisions out of the total decisions. Error rate is complement of Accuracy rate.

$$A = \frac{TP + TN}{TP + TN + FP + FN} \quad (2)$$

$$E = \frac{FP + FN}{TP + TN + FP + FN} = 1 - A \quad (3)$$

Other measures of evaluation are precise and recall. Precision (πi) is the number of true positive predictions made out of total number of positive predictions i.e.

$$\pi_i = \frac{TP_i}{TP_i + FP_i} \quad (4)$$

Recall (ρi) is the number of true positive predictions made out of total number of actual true values i.e.

$$\rho_i = \frac{TP_i}{TP_i + FN_i} \quad (5)$$

5. Implementation

The email ham-spam classification is done in two ways, first without applying any dimensionality reduction method and second by applying DR methods on the best classifier.

A. Without inclusion of DR

The classification was done using three separate classifiers that are described in the previous section. When single decision tree R-parts (the simplest classifier) was used the accuracy scaled to 88.7%. Out of 3901 emails, only 3460 emails were correctly classified (3107 as hams and 353 as spams). Its confusion matrix is as follows:

TABLE II. Confusion Matrix of Classification through R-Parts

Predicted	Reference	
	Ham	Spam
Ham	3107	60
Spam	381	353

Then the mighty Random Forest was used in place of R-parts and the accuracy increased by 2 percent i.e. 90.81%. The confusion matrix is as follows:

Table III. Confusion matrix of classification through random forest

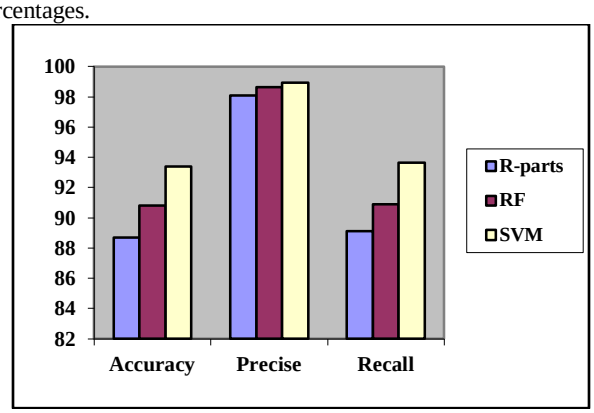
Predicted	Reference	
	Ham	Spam
Ham	3172	43
Spam	316	370

Finally classification was done by Support Vector Machine which gave the better results as expected. The accuracy achieved through SVM is 93.41%. Its confusion matrix is as follows:

Table IV. Confusion Matrix of Classification through Svm

Predicted	Reference	
	Ham	Spam
Ham	3266	35
Spam	222	378

The following graphs, expresses the performance of all three classifiers in terms of classification accuracy, precise and recall percentages.



Graph 1: Performance measures of classifiers

B. Applying DR prior to classification

From the previous three results it is clear that SVM is the better classifier and it is appropriate to compare the performances of different DR methods classified by SVM.

In the first we applied the simplest extraction method Principal Component Analysis(PCA) and reduced dimensions (features) from 5742 to 840 preserving 90% variance of data. Later SVM was used to classify the emails. The accuracy obtained is 93.89%. Even with ICA (successor of PCA), the accuracy is only 94.11%. Compared to ICA, PCA is quick and less complex.

Then LDA was applied under supervised conditions and accuracy after classification was 94.36%. Next the non linear algorithm ISOMAP was applied and the accuracy achieved after SVM classification was 95.27%.

Table V. Performance of DR methods

DR method	Extracted Dimensions	Accuracy of classification
PCA	840	93.89%
ICA	1064	94.11%
LSA/SVD	300	96.75%
LDA	751	94.36%,
ISOMAP	688	95.27%
MDS	688	94.54%

Finally, Latent Semantic Analysis was applied on the tf-idf matrix by employing truncated SVD where we obtained 300 'higher concept' dimensions. The accuracy we got after classifying this reduced data by SVM was 96.75%. Thus we could achieve an increase of 3% percent accuracy with the best performing DR.

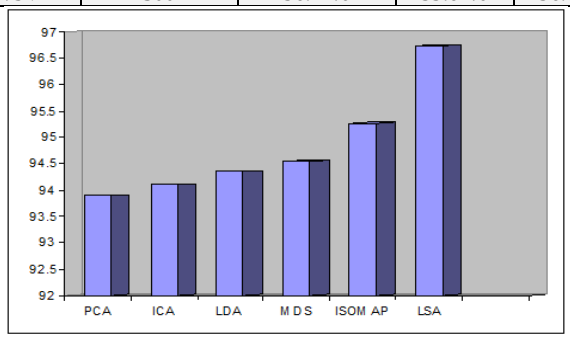
Table VI. Confusion Matrix of Classification through LSA/SVD and SVM

Predicted	Reference	
	Ham	Spam
Ham	3373	5
Spam	108	415

The below table lists out the performance measures of classification carried out using different DR methods followed by the SVM procedure. Measures considered are classification accuracy, precision and recall percentages. A graph following the table represents the comparison of classification accuracies obtained from the six methods.

Table VII. Performance of DR Methods

DR method	Extracted Dimensions	Accuracy of classification	Precision Value	Recall Value
PCA	840	93.89%	98.76%	94.35%
ICA	1064	94.10%	98.83%	94.52%
LDA	751	94.36%,	98.95%	94.69%
MDS	688	94.54%	98.87%	94.98%
ISOMAP	688	95.26%	99.04%	95.61%
LSA/SVD	300	96.74%	99.67%	96.67%



Graph 2: Classification accuracies of DR methods

Meanwhile the cluster preserving DR algorithms would even more enhance the accuracy of classification as shown by Kim[3]. But it would appeal for intensive computation time as it involves two processes, clustering and reducing. A higher accuracy of 97.1% was achieved by applying an enhancing measure called the Cosine Similarity Measure. This measure was applied on the data space reduced by LSA and then Support Vector Machine was used to classify the Emails.

Table VIII: Confusion Matrix of Classification by

Predicted	Reference	
	Ham	Spam
Ham	3380	5
Spam	108	408

It has also been pointed out that the length of email is specifically a significant measure to classify an email as ham or spam. All the spams are lengthy emails whereas the hams are generally short. It is based on the idea that spam emails or messages contain too many words that are irrelevant whereas hams are short & brief. The length of text as an enhancing measure resulted in 97.04% classification accuracy. However it is an application specific enhancing measure that can only work on the basis of principle discussed above.

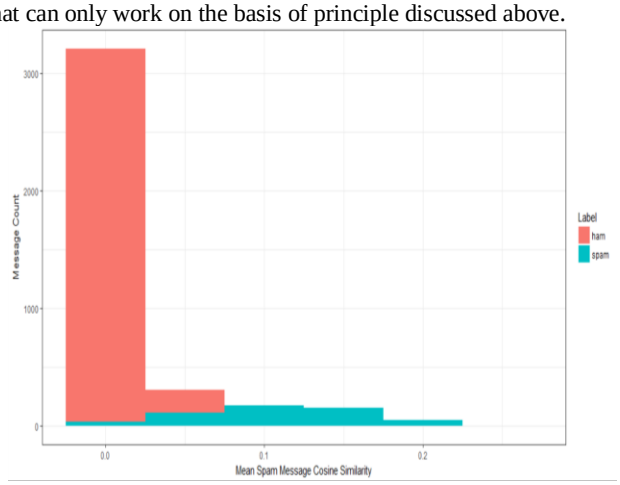


Fig 2: Distribution of 'ham' vs 'spam' using Cosine Similarity Measure

VI. Conclusion

This paper has examined how pre-processing with dimensionality reduction can improve text analysis using classification accuracy to measure performance. Without the intervention of DR, the accuracy is only 93.41%. Of the six DR techniques that we considered, all were able to achieve substantial improvements compared to the approach where DR was not applied. Even the simplest DR technique PCA showed a slight increase in the accuracy from 93.41% to 93.89%. Of the all six algorithms, LSA gave the highest accuracy of 96.75% which is 3.34% higher than not applying DR.

Meanwhile, the data points of text based applications lie on the linear sub space requiring no other information than the Euclidean distance. However non linear DR methods like ISOMAP are based on geodesic distance which preserve the local structure that are tailor made for complex applications like image retrieval, handwriting recognition but not for text categorisation. And by applying ISOMAP we could achieve an accuracy of 95.27% which is 1.48% less than applying LSA. Apart from ISOMAP all other algorithms used are of same complexity. Thus Latent Semantic Analysis is regarded as the most suitable and coherent Dimensional Reduction method for Text Categorization which can be included in the unified framework of DR and TC of any text based application. It is regarded as the most coherent DR method of all other methods as it is a best overall performer in terms of classification accuracy, computation and space cost and complexity of algorithm.

References

[1] Richard Ernest Bellman, “Adaptive control processes: a guided tour”, Princeton University Press, 1967.

[2] Y. Yang and J. O. Pedersen, “A comparative study on feature selection in text categorization”, International Conference on Machine Learning, 1997.

[3] Hyunsoo Kim, Peg Howland and Haesun Park, “Dimension Reduction in Text Classification with Support Vector Machines”.

[4] E. Bingham and H. Mannila, “Random projection in dimensionality reduction: applications to image and text data”. In ACM Special Interest Group on Management of Data. ACM Press, 2001.

[5] Underhill, D.G., McDowell, L., Marchette, D.J., & Solka, J.L. (2007). Enhancing Text Analysis via Dimensionality Reduction. 2007 IEEE International Conference on Information Reuse and Integration, 348-353.

[6] Aas, K. and Eikvil, L. 1999. Text Categorization: A survey. Tech. rep. 941. Norwegian Computing Center, Oslo, Norway.

[7] Yang, Y., Slattery, S., and Ghani, R., 2002, “A study of approaches to hypertext categorization”, J. Intell. Inform. Syst. 18, 2/3 (March-May), 219–241.

[8] Fabrizio Sebastiani, “Machine learning in automated text categorization”, ACM Computing Surveys, 34(1):1-47, 2002.

[9] F. Wickelmaier. An introduction to MDS. Technical report,Aalborg University (Denmark), May 2003.

[10] Manning, Christopher; Raghavan, Prabhakar; Schütze, Hinrich, "Vector space classification: Introduction to Information Retrieval”, Cambridge University Press, 2008

[11] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman, “Indexing by latent semantic analysis”. Journal of the Society for Information Science, 41:391-407, 1990.

[12] Chelsea Boling and Kumar Das, “Reducing Dimensionality of Text Documents using Latent Semantic Analysis”

[13] P. Howland, M. Jeon, and H. Park, “Structure Preserving Dimension Reduction for Clustered Text Data based on the Generalized Singular Value Decomposition”, SIAM Journal of Matrix Analysis and Applications, 25(1):165–179, 2003.

[14] Yogesh Jain, Amit kumar Nandanwar, A Theoretical Study of Text Document Clustering, “International Journal of Computer Science and Information Technologies”, Vol. 5 (2), 2014, 2246-2251

[15] Pratiksha Y Pawar and S H Gawande, “A Comparitive Study on Different Types of Approaches to Text Categorization”, International Journal of Machine Learning and Computing, Vol 2, No 4, August 2012